Meten en maximaliseren van basale schrijfvaardigheid bij eerstejaarsstudenten in het hoger beroepsonderwijs

Anouk van Eerden

Mik van Es

Meten en maximaliseren van basale schrijfvaardigheid
bij eerstejaarsstudenten in het hoger beroepsonderwijs
Proefschrift

Anouk van Eerden, Mik van Es
Rijksuniversiteit Groningen

Groningen, 2014
ISBN: 978-90-367-6710-1 (pdf)
ISBN: 978-90-367-6709-5 (boek)

Lay-out: Mik en Anouk
Omslag: Ontwerpburo Peter van der Weele - PLW'12
Druk: Off Page, Amsterdam

Website bij boek: http://basaleschrijfvaardigheid.blogspot.nl/

rijksuniversiteit groningen

Meten en maximaliseren van basale schrijfvaardigheid bij eerstejaarsstudenten in het hoger beroepsonderwijs

Proefschrift

ter verkrijging van de graad van doctor aan de
Rijksuniversiteit Groningen
op gezag van de
rector magnificus, prof. dr. E. Sterken
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

donderdag 8 mei 2014

om 14.30 uur door

Anneke van Eerden
geboren op 17 juni 1953
te Groningen

en om 16.15 uur door

Marinus van Es
geboren op 26 juni 1949
te Joure

Promotor:

Prof. dr. C.L.J. de Bot

Beoordelingscommissie:

Prof. dr. W.F. Admiraal

Prof. dr. ir. J. Nerbonne

Prof. dr. Th. Wubbels

Verkorte inhoudsopgave

Inhoudsopgave

An ounce of practice is worth more than tons of preaching

Mahatma Gandhi

[]

1

Inleiding

1.1 Tekortschietende schrijfvaardigheid

"Mijn leven gaat voorbij met het corrigeren van teksten, althans daar lijkt het soms op. Als de meester van groep 7 zit ik dag in, dag uit elementaire spel- en taalfouten te corrigeren. De teksten zijn alleen niet geproduceerd door kindertjes, maar door hoog opgeleide academici, studenten", aldus onderzoeker en emeritus hoogleraar klinische biochemie Piet Borst (2009) in een column. Hoewel zijn opmerking zich niet beperkt tot studenten, snijdt hij hier een probleem aan dat in het Nederlandse hoger onderwijs voortdurend terugkomt. Aan de ene kant gaat men er in dat onderwijs vanuit dat studenten een goede schrijfvaardigheid bezitten; aan de andere kant blijkt in de praktijk dat daar geen sprake van is.

In kranten verschenen artikelen over studenten die niet kunnen schrijven. Teksten van studenten zonder taalfouten zouden uitzonderlijk zijn. De Paauw, docent inleiding tot de rechtswetenschap aan de Erasmus Universiteit Rotterdam, beoordeelde taaltoetsen van eerstejaars rechtenstudenten. Volgens hem was het ongelofelijk wat hij soms las. "Ik heb me vaak afgevraagd tijdens het nakijken hoe sommige van deze studenten hun vwo gehaald hebben" (geciteerd in Bouma, 2007a). In opdracht van de overheid zijn nota's geschreven over de mogelijke oorzaken van de slechte taalbeheersing en welke vaardigheden in het basisonderwijs en voortgezet onderwijs verworven zouden moeten worden (Tijd voor onderwijs, 2008; Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2008). Vanuit het hoger beroepsonderwijs (hbo) zijn verschillende pogingen ondernomen om de taalbeheersing van studenten te verbeteren (Beijer, Gangaram Panday & Hajer, 2010; Gangaram Panday, Droop & Rutten, 2008; Van der Westen, 2006; Van der Westen, 2011b).

Studenten werden vanaf 2006 grootschalig getoetst op taalfouten. De resultaten op de taaltoetsen waren slecht tot zeer slecht. De eerste uniforme taaltoets is ontwikkeld door het Cito in samenwerking met het Expertisecentrum Nederlands en in 2006 voorgelegd aan 5748 eerstejaars pabo-studenten. Van de havisten zakte 70% voor de toets en van de mbo'ers 85% (Bal, Berger, De Jonge, Oudmaijer & Tan, 2007). In 2007 liet de hbo-raad een onderzoek doen onder 2000 studenten van een aantal andere hbo-opleidingen, waaronder bouwkunde en verpleegkunde. Ruim 30% van de studenten was niet in staat een voldoende te halen op een taaltoets van niveau havo-4 (Gerrits, 2008). Onderzoek onder universitaire eerstejaarsstudenten leverde een soortgelijk beeld op. In 2006 zakte bij de Erasmus Universiteit 46% van de eerstejaars rechtenstudenten voor de taaltoets (Bouma, 2007a). De Vrije Universiteit Amsterdam heeft in 2007 een taaltoets afgenomen onder 1100 eerstejaarsstudenten van de

faculteiten Rechten, Letteren en Exacte Wetenschappen en circa 30% haalde een onvoldoende (Reijn, 2008).

In de bovenstaande voorbeelden wezen de resultaten op de taaltoetsen weliswaar allemaal in dezelfde richting, maar verder was het niet mogelijk om de uitkomsten goed met elkaar te vergelijken. Er was namelijk niet sprake van een uniforme taaltoets. De onderdelen die getoetst werden, waren niet exact hetzelfde, de manier waarop getoetst werd, verschilde en het niveau van de toetsen varieerde, evenals de norm. Bovendien is het de vraag in hoeverre het resultaat op een niet-gevalideerde taaltoets iets voorspelt over schrijfvaardigheid (Deygers & Kanobana, 2010; Peters, Van Houtven & El Morabit, 2010).

Als de schrijfvaardigheid van eerstejaarsstudenten te wensen overlaat, hoe zit het dan met de schrijfvaardigheid van leerlingen in het voortgezet onderwijs? Bonset en Braaksma (2008) merkten naar aanleiding van onderzoek dat is verricht in de periode 1969-1997, hierover op: "De conclusies van de onderzoekers over de gemiddelde schrijfprestaties van de leerlingen zijn somber: deze liggen onder de norm van wat onderzoekers en bevraagde panels wenselijk achten" (p. 129). Voor deze periode zijn dus gegevens beschikbaar, die echter moeilijk interpreteerbaar zijn. Hetzelfde probleem doet zich hier voor als bij de scores die studenten in het hoger onderwijs halen op taaltoetsen. Om dit probleem op te lossen zouden gestandaardiseerde toetsen met vaste normen ontwikkeld moeten worden.

In de periode daarna, 1997-2007, is er geen gericht onderzoek verricht naar de schrijfvaardigheid van de doorsnee leerling in het voortgezet onderwijs, zodat recente gegevens niet beschikbaar zijn. Bonset en Braaksma (2008) stelden: "Opvallend is het volledig ontbreken van instrumentatieonderzoek na 1997, omdat dit in de periode daarvoor in zeer ruime mate is verricht. . . . Een oorzaak kan zijn dat schrijfvaardigheid sinds 1998 geen deel meer uitmaakt van het centraal examen in havo en vwo, zodat de noodzaak van het oplossen van de beoordelingsproblematiek in de onderwijs- en examenpraktijk minder groot is geworden" (p. 134-135).

Wat vinden leerlingen in het voortgezet onderwijs zelf van het tegenwoordige schrijfonderwijs? Bonset en Braaksma (2008) citeerden Van de Ven (2003) en Van de Ven, Martens en Imants (2005) die leerlingen in de bovenbouw van havo/vwo interviewden. Schrijven wordt een typisch schoolse activiteit gevonden. Het kost veel tijd. Het is tekstproductie voor

school, vooral voor Nederlands. Het is het invullen van een formulier: de door de methode aangegeven tekststructuur. Een verband met teksten en schrijven buiten school zien de leerlingen niet. Leren ze er iets van? De leerlingen rapporteerden nauwelijks leerervaringen.

Tekortschietende schrijfvaardigheid is overigens geen nieuw probleem en het is ook geen specifiek Nederlands probleem. In het kader van dit onderzoek worden de twee laatste punten alleen aangestipt. Klachten over moedertaalonderwijs hebben een lange historie. In het overzicht dat Van de Ven (1986) gaf, was volgens hem een bepaald patroon zichtbaar. Het onderwijs deugt niet en studenten kunnen niet fatsoenlijk schrijven. Iedere generatie leek dit opnieuw te vast te stellen. In 1893 zijn bij Kalff negatieve uitspraken van professoren te vinden over het taalgebruik van hun studenten. Volgens Van Ginneken (1917) kan de jeugd niet meer dragelijk schrijven. Leest vermeldde in 1932: dat "een jongmens . . . niet in staat blijkt een simpel briefje of verslagje zonder fouten te schrijven" (geciteerd in Van de Ven, 1986, p. 4). Van den Ent vermeldde in 1941 dat universiteiten klagen over studenten die niet in staat bleken om te schrijven zonder ernstige taal- en stijlfouten. Stuiveling werd in 1952 buitengewoon hoogleraar taalbeheersing. Hij was verontrust over de "slordigheid van stijl, de onduidelijkheid van formulering, de beperktheid van woordkeus, de aanwijsbare fouten zelfs in spelling en zinsbouw, die scripties en proefschriften, examens en promoties te zien en te horen geven" (Van de Ven, 1986, p. 6).

De situatie in de tachtiger jaren lijkt eveneens op die van nu. 'Er wort steets meer fout gesgrefen', heette een verkennende ministeriële nota met klachten over het moedertaalonderwijs uit 1981 (Baltzer 1988, p. 1). In de tachtiger jaren verscheen over dit onderwerp een reeks publicaties van onder andere Baltzer (1986), Baltzer, De Glopper & Van Schooten (1988), Bochardt (1985), Van Dijck (1985). In de lijst van problematische aspecten die genoemd werden op het gebied van schriftelijke taalvaardigheid, staat ook het schrijven zonder spel- en taalfouten. Onvoldoende beheersing van taalvaardigheid werd als een van de belangrijkste oorzaken gezien van de aansluitingsproblemen bij de overgang naar het hoger onderwijs. Verschillende universiteiten gingen, net als nu, cursussen schrijfvaardigheid aanbieden. In 1985 werd een schriftelijke enquête over taaltaken gehouden onder circa tweeduizend eerstejaarsstudenten. Een van de belangrijkste conclusies was, dat studenten moeite zeggen te hebben met het schrijven van verslagen, scripties en samenvattingen (Bochardt, 1986). Bij deze taken gaven de studenten aan, dat hierin onvoldoende aandacht was besteed in de vooropleiding.

Deze klachten klinken eigentijds. Dergelijke uitspraken van soms meer dan honderd jaar geleden relativeren de huidige commotie over het schriftelijk taalgebruik van studenten. Wat ook op de tegenwoordige situatie lijkt, is dat systematisch onderzoek naar de omvang en de aard van de toenmalige klachten over de schriftelijke taalvaardigheid ontbrak.

Het is evenmin een typisch Nederlands probleem. Ook elders bestaat zorg over het schrijfniveau van studenten. Vlaamse docenten aan hogescholen in Gent, Brussel en Antwerpen zijn bezorgd over de Nederlandse schrijfvaardigheid van hun studenten (Peters & Van Houtven, 2010). Engelse en Amerikaanse studenten bleken moeite te hebben met basale taalvaardigheden en veel fouten te maken als ze zakelijk en academisch schrijven (Anson, 2000; Bacon & Scott Anderson, 2004; Binder & Watkins, 1990; Connor, 1990; Connors & Lunsford, 1998; Gilbert, 2004; Graham & Perin, 2007a; Lunsford & Lunsford, 2008; Miller & McCardle, 2011). Volgens Cyr (2011) scoorden Amerikaanse studenten nummer één op het gebied van zelfvertrouwen, terwijl zij niet in staat waren een simpel schriftelijk vragenlijstje zonder opvallende taalfouten te maken. Een schrikbarend aantal Amerikaanse studenten zou overigens in het algemeen weinig elementaire kennis paraat hebben (Foer, 2011). Engelse ondernemers zouden bij het werven van nieuw personeel geschokt zijn door de slechte kwaliteit van het geschreven Engels (Coughlan, 2011).

In geen enkel ander land waren de beoordelaars echter zo negatief over het schrijfniveau als in Nederland, bleek uit het IEA-onderzoek (International Association for the Evaluation of Educational Achievement) van 1992. Van de twaalf vermelde landen waren de Duitse beoordelaars (Hamburg) het positiefst over het schrijfniveau van de leerlingen met slechts 2.2% op het allerlaagste niveau. Nederland scoorde het slechtst met meer dan een van de vijf leerlingen (22.4%) op het allerlaagste niveau. Zweden en Nigeria scoorden daarna het laagst met respectievelijk 18.4% en 18.0%. Het gemiddelde van de twaalf landen bedroeg 12.3% leerlingen op het laagste niveau (Purves, 1992, p. 118).

Het probleem van een tekortschietende schrijfvaardigheid beperkt zich al met al niet tot Nederland en het is niet nieuw, maar het is ook nog steeds niet opgelost. Dit probleem kan wel verstrekkende gevolgen hebben. Opleidingen kunnen niet het gewenste niveau handhaven als basisvaardigheden, waaronder die van Nederlands, bij de student ontbreken. Omgekeerd is het te verwachten dat studenten met gebrekkige basisvaardigheden eerder een opleiding

zonder diploma zullen verlaten en langer over hun studie zullen doen. Zowel het schrijven van een zakelijke tekst als het schrijven van een effectieve tekst lijkt onmogelijk als studenten problemen hebben met het schrijven van een correcte, heldere zin. Ook na het afronden van de opleiding blijft een tekort aan basale schrijfvaardigheid een levenslange handicap. Verbetering van basale schrijfvaardigheid bij eerstejaars hbo-studenten stond daarom centraal in dit onderzoek.

Hoewel er redenen zijn om aan te nemen dat het slecht gesteld is met de schrijfvaardigheid van studenten in het Nederlandse hoger onderwijs, zijn er weinig kwantitatieve gegevens waaruit dit ondubbelzinnig blijkt. Een belangrijk punt is dat gestandaardiseerde instrumenten ontbreken. Schrijfvaardigheid is lastig of niet kwantificeerbaar en rapportages over het tekortschietende niveau vallen daardoor terug op geselecteerde waarnemingen, anekdotes en niet-gestandaardiseerde toetsresultaten. Een paar van zulke geselecteerde waarnemingen staan in Tabel 1.1. De fragmenten komen uit teksten van hbo-studenten. Ze zijn slechts bedoeld als illustratiemateriaal.

1.2 Onderzoeksvragen

De aanleiding voor dit onderzoek was het grote aantal fouten in het werk van eerstejaars hbo-studenten dat een van beide onderzoekers dacht waar te nemen. Bevatte het Nederlands van deze studenten werkelijk zoveel fouten of was zij als docent zo gericht op fouten dat ze overal fouten zag? Wanneer studenten inderdaad zoveel fouten maken, zou het dan niet mogelijk zijn daar iets aan te doen met een gerichte training? Het doel van het onderzoek was het beantwoorden van deze twee vragen. De beantwoording van die vragen riep vervolgens nieuwe vragen op.

In dit onderzoek gaat het om twee problemen: het meten en het maximaliseren. Het meten van basale schrijfvaardigheid veronderstelt dat men vrij precies kan aangeven wat verstaan wordt onder 'basale schrijfvaardigheid' en wel op zo'n manier dat dit bij studenten waargenomen en gekwantificeerd kan worden. Het aantal objectief aantoonbare fouten in het werk van studenten leek hiervoor een goed uitgangspunt.

Tabel 1.1 Geselecteerde fragmenten uit teksten van hbo-studenten

---

** Vaak worden consumenten verward met vetten. Dit komt door dat er verzadigde en onverzadigde vetten zijn. de verzadigde vetten heb je dagelijks nodig om te verbranden de onverzadigde vetten zijn vetten die blijvent zijn.*

** Binnen de bedrijf zijn verschillende afdelingen. Functies op de afdelingen met focus op de Nederlandse markt wordt mbo als hbo diploma gevraagd.*

** Supermarktondernemers krijgen vooral in het bijzonder veel van deze producten in de winkels.*

** Maar toch zal de belangstelling van de product en door ontevredenheid terug lopen.*

** Bovenaan staat de vestingmanager van de organisatiestructuur.*

** Deze kant en klare diepvriestaart is in 3 smaken op de markt gekomen , heeft geen bereidingstijd nodig, kan bevroren en ontdooid genuttigd woren en is voor een langere tijd houdbaar.*

** Dit rapport is geschreven voor de net gestarte onderneming onder leiding van vier studenten. En ze een optimale interne situatie te schetsen om zo effectief mogelijk te opereren.*

** Er mag geconcludeerd worden dat de ambitie van dit bedrijf een potentiële misfit is met een grote verandering.*

** Daarentegen zijn er ook producten en consumenten die waarde hechten en tevreden zijn over bepaalde producten zoals de frisdranken waar de suikergehalte bijna met de helft gehalveerd is.*

** Ondanks dit alles worden de smaak en de groei van de omzet van light frisdranken gewardeerd.*

Het idee dat eerstejaarsstudenten slecht schrijven, is zolang het niet goed onderbouwd wordt, in de eerste plaats een idee. Iemand anders kan op dit punt een andere mening verkondigen. Deze overweging leidde tot de eerste onderzoeksvraag:

- wat is het niveau van basale schrijfvaardigheid voor Nederlands bij eerstejaarsstudenten in het hoger onderwijs?

Deze onderzoeksvraag leidde tot een aantal vragen. De eerste vraag was of fouten objectief vaststelbaar zijn of dat ze alleen een subjectieve projectie van een beoordelaar op de tekst vormen. De tweede vraag was hoeveel fouten, objectief gedefinieerd, in teksten van eerstejaarsstudenten voorkomen. De daaropvolgende vraag was welke soorten fouten voorkomen.

Het maximaliseringsprobleem is vervolgens uitgewerkt in drie onderzoeksvragen. De eenvoudigste oplossing om basale schrijfvaardigheid te verbeteren is een bestaande onderwijsmethode te gebruiken om studenten op dit punt bij te spijkeren. De tweede onderzoeksvraag was daarmee:

- wat is de waarde van bestaande methodes om basale schrijfvaardigheid te verbeteren bij eerstejaars hbo-studenten?

De veronderstelling dat deze methodes misschien niet ideaal zouden zijn, leidde tot de derde onderzoeksvraag:

- hoe moet een nieuw onderwijsprogramma eruit zien om basale schrijfvaardigheid bij hbo-studenten te verbeteren?

Het construeren van een nieuw onderwijsprogramma waarvan men aanneemt dat het betere resultaten zal opleveren, is vaak veel werk. Zonder de effectiviteit echter daadwerkelijk te onderzoeken en te toetsen, blijft de veronderstelde grotere effectiviteit niets meer dan een veronderstelling. Een nieuw onderwijsprogramma zal daarom empirisch op effectiviteit onderzocht moeten worden. Dit leidde tot de vierde onderzoeksvraag:

- hoeveel effect heeft dit nieuwe onderwijsprogramma op de basale schrijfvaardigheid van eerstejaars hbo-studenten?

Het vaststellen van taalfouten in teksten van studenten en het verbeteren van de schrijfvaardigheid op dit punt stonden centraal in dit onderzoek. Bij het verbeteren van de schrijfvaardigheid van studenten rijst echter de vraag of fouten in teksten wel zo belangrijk zijn en in welke mate fouten in de tekst invloed hebben op de lezer. Met andere woorden: gaat het

erom te schrijven zonder fouten, omdat dit de norm is of hebben fouten in de tekst inderdaad een aantoonbaar negatief effect op de lezer? De vijfde onderzoeksvraag luidde daarom:

- in hoeverre hebben taalfouten in een tekst effect op de waardering van die tekst door de lezer?

Het fouteneffect-onderzoek maakte het vervolgens mogelijk een relatie te leggen tussen het aantal fouten per honderd woorden en de door docenten meest gebruikte methode voor het vaststellen van schrijfvaardigheid: het holistische oordeel. Bij holistische beoordeling beoordeelt een 'expert-beoordelaar' in korte tijd een tekst door die globaal te lezen en te bekijken. Op basis daarvan kwam een globale vraag naar voren:

- welke manieren zijn er nog meer om basale schrijfvaardigheid vast te stellen?

De laatste vraag die naar voren kwam, was:

- kan basale schrijfvaardigheid betrouwbaar vastgesteld worden op basis van een tekst ter lengte van een halve A4 (250 woorden)?

Basale schrijfvaardigheid

Het begrip 'schrijfvaardigheid' wordt in dit onderzoek als volgt opgevat: een kort stukje tekst kunnen schrijven zonder duidelijke fouten. Deze enge opvatting van schrijfvaardigheid wordt hier aangeduid als 'basale schrijfvaardigheid' en verder uitgewerkt en gespecificeerd in termen van het aantal (objectief aantoonbare) fouten per honderd woorden.

Schrijfvaardigheid wordt in het onderwijs vaak zeer breed opgevat. Bij schrijfvaardigheid lijkt men te denken aan succesvolle schrijvers die jarenlang bronnenonderzoek gepleegd hebben, hun informatie zorgvuldig geordend en geëvalueerd hebben en vervolgens alles zo hebben opgeschreven dat een vlot lezend boek voor een groot publiek het resultaat is. Zo vermeldde het rapport *Over de drempels met taal* van de Expertgroep Doorlopende Leerlijnen Taal en Rekenen (2008) voor schrijfniveau 4F (eind vwo): "Kan schrijven voor zowel publiek uit de eigen omgeving als voor een algemeen lezerspubliek (bv. instanties, media)" (p. 63). Met andere woorden: de leerling kan aan het einde van het vwo schrijven op het niveau van een journalist. Voor het veel lagere niveau 2F (eind vmbo) wordt zelfs expliciet vermeld: "kan een . . . krant . . . maken" (p. 63). In het rapport gaat men ervan uit dat 75%

van de leerlingen aan het einde van het desbetreffende onderwijstype de omschreven vaardigheid 'waarschijnlijk' beheerst (p. 6).

'Schrijven' wordt in het rapport van de Expertgroep gedefinieerd als: "het lezen en analyseren van bronnen, het genereren van ideeën, het structureren van ideeën, het stellen van retorische doelen, het formuleren van tekst, het lezen van tekst die al geschreven is, het evalueren en verbeteren van die tekst, en . . . het coördineren van die activiteiten in de tijd" (p. 61).

Deze brede opvatting van schrijfvaardigheid was niet het uitgangspunt voor dit onderzoek. Basale schrijfvaardigheid is in het onderwijs vaak niet populair, dit in tegenstelling tot de brede opvatting van schrijfvaardigheid. Zo stelde de Expertgroep dat 'correct schrijven' niet behoorde tot het domein 'schrijven', maar tot het domein 'taalbeschouwing en taalverzorging'. Een restcategorie waarin men grammatica en correct schrijven liet vallen. De impliciete boodschap was: echt schrijven en correct schrijven zijn totaal verschillende vaardigheden. De Expertgroep motiveerde deze splitsing als volgt: "Uit veel onderzoek blijkt dat de cognitieve belasting bij het schrijven van een tekst zo groot is dat het niet mogelijk is tegelijkertijd aandacht te besteden aan inhoudelijke en vormelijke aspecten van formulering. Daarom is het in het onderwijs van belang deze twee aspecten van het schrijven afzonderlijk aandacht te geven" (Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2008, p.70). Het is echter de vraag of een zin met één duidelijke fout per zes woorden (een waarde die in het foutenonderzoek bij benadering gevonden werd) nog eenvoudig te begrijpen valt.

Onder 'taalverzorging' bleek volgens de door de Expertgroep vermelde kerndoelen te vallen: interpunctie, spelling, woordenschat, lay-out, grammaticale fouten, niet lopende zinnen en het reviseren van de tekst (p. 81). De Expertgroep vond de doelstellingen voor de verschillende schooltypen dermate vaag dat men geen kans zag op dit punt een doorlopende leerlijn te formuleren (p. 69). Afgaande op de vermelde doelstellingen werd foutloos schrijven gezien als iets dat vooral in het basisonderwijs moet worden geleerd (p. 81).

Uit de eerste deelstudie, het foutenonderzoek, zal echter blijken dat de schrijfproducten van eerstejaarsstudenten zeer grote aantallen fouten bevatten. Dit resultaat wordt bevestigd in de derde deelstudie, de constructie en evaluatie van het TAVAN-programma. De conclusie lijkt dan ook gerechtvaardigd dat foutloos schrijven niet (meer) geleerd wordt in het basisonderwijs en ook niet in het vervolgonderwijs. Een vaardigheid die zo basaal leek, dat de

Expertgroep dacht dat hij vooral in het basisonderwijs geleerd zou worden, blijkt helemaal niet meer geleerd te worden.

Doordat schrijfvaardigheid in het onderwijs zo breed wordt opgevat, is het in de praktijk ook moeilijk meetbaar (zie 2.1). Beoordelaars stemmen doorgaans weinig overeen bij het beoordelen van teksten. Basale schrijfvaardigheid (het aantal fouten per honderd woorden) blijkt echter reeds betrouwbaar vastgesteld te kunnen worden door een enkele expert-beoordelaar (hoofdstuk 4).

Basale schrijfvaardigheid blijkt absoluut gemeten te kunnen worden, dat wil zeggen op een schaal met een echt nulpunt, zodat 0 ook echt 0 fouten is (hoofdstuk 4). Dit maakt het gemakkelijker om de aantallen fouten per honderd woorden die bij studenten gemeten worden, te interpreteren.

Basale schrijfvaardigheid kan niet alleen betrouwbaar gemeten worden, maar blijkt ook goed te remediëren. In de derde deelstudie (hoofdstuk 6) wordt aangetoond dat de basale schrijfvaardigheid van eerstejaars hbo-studenten in korte tijd (20 uur) belangrijk verbeterd wordt door een gerichte training.

Basale schrijfvaardigheid blijkt ook betrouwbaar vastgesteld te kunnen worden door het bij de training gebruikte online-programma. Dit programma werkt niet met meerkeuze vragen, maar de student moet daadwerkelijk (her)schrijven (hoofdstuk 6).

De basale schrijfvaardigheid van een student en het holistische oordeel van een expert-beoordelaar over een door de student geschreven tekst blijken op vrijwel hetzelfde neer te komen. Beide maten correleren zeer hoog (hoofdstuk 7).

Is basale schrijfvaardigheid wel belangrijk? Het gaat er toch om dat de betekenis van de tekst overkomt? Uit een experiment (hoofdstuk 7) blijkt dat fouten in de tekst inderdaad een zeer groot effect hebben op de waardering door lezers. Een goede basale schrijfvaardigheid is dus bij schrijven belangrijk.

Verder blijken studenten met een goede basale schrijfvaardigheid sneller te schrijven, een hogere vooropleiding te hebben en belangrijk minder kans te hebben de studie te staken (hoofdstuk 6) en beter in staat te zijn de kwaliteit van teksten van medestudenten in te schatten (hoofdstuk 7).

Hoeveel tekst is er nodig om basale schrijfvaardigheid betrouwbaar vast te stellen? In de zesde deelstudie (hoofdstuk 9) blijkt dat een korte tekst (250 woorden) al voldoende is.

1.3 Drie benaderingen van het meetprobleem

Dit proefschrift gaat uit van de veronderstelling dat goed meten een voorwaarde is om met succes te kunnen maximaliseren (trainen, effectief onderwijs verzorgen). Effectief schrijfvaardigheidsonderwijs is vanuit dat uitgangspunt in eerste instantie vooral een meetprobleem en pas in tweede instantie een didactisch probleem.

Op welke manier kan een bepaalde eigenschap gemeten worden? De gangbare manier is het juiste meetinstrument te gebruiken (Gertsbakh, 2003). We meten ons lichaamsgewicht met een personenweegschaal. We meten tijd door op een klok te kijken. De prijs van een huis wordt bepaald via een beëdigd taxateur. De rijvaardigheid van een kandidaat wordt bepaald door een examinator. Om te meten gaan we uit van het juiste meetinstrument, waarbij dat meetinstrument ook een deskundige kan zijn. De waarde van de meting, de 'validiteit', veronderstellen we op basis van het gebruikte meetinstrument. We nemen aan dat we het gewicht bepaald hebben, omdat we een weegschaal hebben gebruikt.

Een tweede manier van meten is op het laatste punt belangrijk anders. Er is nog steeds een meetinstrument, maar men is niet langer volledig overtuigd van de waarde. Klopt de meting wel? Geeft de weegschaal wel het juiste gewicht aan? Het openmaken van een weegschaal om te kijken of binnenin alles in orde is, lijkt een wat onzekere onderneming. Om bevestiging te krijgen zal men daarom proberen het aangegeven gewicht te checken met een andere weegschaal. Bij deze tweede manier van meten staat men argwanend tegenover het gebruikte meetinstrument en probeert men dat te controleren door de meetresultaten te vergelijken met de resultaten van andere meetinstrumenten.

Hoe moet schrijfvaardigheid gemeten worden? Ook deze vraag kan in beginsel op dezelfde twee manieren beantwoord worden. We kunnen voor de validiteit van de meting focussen op het meetinstrument of we kunnen voor de validiteit focussen op het meetresultaat. We zullen laten zien dat beide benaderingen wat eenzijdig zijn en beter gecombineerd kunnen worden.

In de traditionele benadering van het meten van schrijfvaardigheid wordt uitgegaan van een 'expert-beoordelaar': een beoordelaar waarvan verondersteld wordt dat die zelf goed kan schrijven en goed in staat is teksten te beoordelen (Garinger, 2002). Een goede expert-beoordelaar zou in korte tijd (enkele minuten) de kwaliteit van een door een student geschreven tekst vrijwel feilloos kunnen vaststellen, de zogenaamde holistische beoordeling. Bij deze benadering is de deskundigheid van de beoordelaar de enige garantie dat het uitgebrachte oordeel juist en waardevol is. De eigenschappen van het meetinstrument, in dit geval de beoordelaar, bepalen de validiteit van het oordeel.

De aanhangers van de tweede manier van meten zijn echter door deze benadering niet bij voorbaat overtuigd. Klopt de meting wel? Volgens de tweede manier van meten moeten we kijken naar het meetresultaat en argwanend staan tegenover het meetinstrument. De waarde van de meting berust in dit geval volledig op de veronderstelde deskundigheid van de expert-beoordelaar. Hoe kunnen we zeker zijn van deze deskundigheid? Wanneer de expert-beoordelaar inderdaad deskundig is, zouden we verwachten dat hij zelf een aantoonbaar goede schrijfvaardigheid heeft. Nu wordt een probleem zichtbaar. De deskundigheid van de expert-beoordelaar kunnen we aantonen op het moment dat we een meetinstrument hebben om schrijfvaardigheid te meten, maar het enige instrument dat we hebben, wordt gevormd door onze deskundige. We zijn terug bij het beginprobleem.

Een volgend probleem met de traditionele benadering dat uit onderzoek blijkt, is dat beoordelaars die verondersteld worden expert te zijn, in de praktijk vaak nauwelijks overeenstemmen (zie 2.1). Een tweede beoordeling door een andere expert-beoordelaar kan tot belangrijk andere uitkomsten leiden. De beoordeling blijkt zeer onbetrouwbaar.

Godshalk, Swineford & Coffman toonden in 1966 aan met behulp van psychometrie en statistiek dat het mogelijk is schrijfvaardigheid via holistische beoordeling betrouwbaar te meten door een groot aantal expert-beoordelaars te gebruiken en de studenten meerdere teksten te laten schrijven. Met 'betrouwbaar' wordt hier bedoeld dat de meting bij herhaling op dezelfde personen soortgelijke uitkomsten oplevert. Deze psychometrische benadering gaat dus niet uit van een enkele expert-beoordelaar, maar van herhaald meten via meerdere beoordelaars en meerdere teksten. Een praktisch bezwaar dat blijft, zijn de vele beoordelaars waardoor de methode voor de onderwijspraktijk vaak te arbeidsintensief is om toe te passen.

In dit proefschrift speelt deze psychometrische benadering van meten (gebaseerd op de klassieke testtheorie) die sterk focust op de meetresultaten, een belangrijke rol, daarom enkele als verduidelijking bedoelde opmerkingen. Het belangrijkste voordeel van de psychometrische benadering is dat het op basis van beoordelaars die amper overeenstemmen, toch mogelijk blijkt betrouwbaar te meten. Dit kan doordat een groot aantal soortgelijke metingen, gecombineerd wordt tot een enkele totaalscore. De afzonderlijke metingen zijn onbetrouwbaar, maar door het resultaat van veel van deze laag betrouwbare metingen te combineren (door optellen of middelen) ontstaat een betrouwbare score. De ruis in de afzonderlijke metingen middelt uit, terwijl de systematische component in de metingen overblijft en optelt (Nunnally, 1967).

Een tweede voordeel van de psychometrische benadering is, dat het niet nodig is dat de beoordelaars echte experts zijn. Het model veronderstelt alleen dat de beoordelaars als groep enigszins expert zijn, dat wil zeggen, het onderling gemiddeld enigszins eens zijn. Of dat inderdaad het geval is, wordt bepaald via de gemiddelde onderlinge correlatie tussen de beoordelaars die positief moet zijn, dus minimaal iets boven nul moet liggen. In hoofdstuk 7 laten we zien dat ook studenten met gemiddeld een matige schrijfvaardigheid met succes als beoordelaars van teksten kunnen worden ingezet.

Een derde voordeel van de psychometrische benadering is dat de kwaliteit van de beoordelaars gemeten en gekwantificeerd wordt. De kwaliteit van de totale groep beoordelaars wordt gekwantificeerd via de gemiddelde onderlinge correlatie. De kwaliteit van een enkele beoordelaar kan op dezelfde manier gekwantificeerd worden via de gemiddelde correlatie met de andere beoordelaars. In de praktijk wordt in plaats hiervan vaak de gecorrigeerde beoordelaar-totaal correlatie gebruikt (overeenkomend met de gecorrigeerde item-totaal correlatie bij een studietoets), wat op ongeveer hetzelfde neerkomt.

Een vierde voordeel van de psychometrische benadering is dat de betrouwbaarheid van de resulterende totaalscore kan worden geschat uit de gemiddelde onderlinge correlatie tussen de beoordelaars (of de items). Dit laatste punt, de betrouwbaarheid van de meting, is belangrijk, omdat een correlatie van 0.40 tussen twee variabelen die beide gemeten zijn met een betrouwbaarheid van 0.45 iets totaal anders betekent dan een correlatie van 0.40 tussen twee variabelen die beide gemeten zijn met een betrouwbaarheid van 0.90 (zie bijlage 1).

De basis van de psychometrische benadering wordt gevormd door de onderlinge correlatie tussen twee metingen (beoordelaars of teksten). Dat twee beoordelaars perfect correleren

(+1.0) wil nog niet zeggen dat beide beoordelaars exact dezelfde oordelen leveren. Bij gebruik van correlaties is de veronderstelling dat iedere beoordelaar zijn eigen gemiddelde en eigen standaarddeviatie (SD) heeft, waarvoor desgewenst eerst gecorrigeerd wordt. Pas nadat de waarnemingen van beide beoordelaars gestandaardiseerd zijn op hetzelfde gemiddelde en dezelfde SD, stemmen ze ook in absolute zin overeen. Dat de ene beoordelaar dus misschien veel lager beoordeelt dan de andere, heeft in de psychometrische benadering geen betekenis, omdat er voor dit verschil wordt gecorrigeerd door beide beoordelaars te standaardiseren op hetzelfde gemiddelde. Ook is niet relevant dat de ene beoordelaar veel gematigder is dan de andere (dichter bij zijn gemiddelde blijft), omdat de SD's van de beoordelaars kunnen worden gestandaardiseerd. Hierdoor wordt voorkomen dat de ene beoordelaar (met de grootste SD) meer invloed uitoefent op het eindoordeel dan de andere beoordelaar.

De validiteit van de psychometrische benadering wordt aangetast door maatregelen als 'overleggen over gevallen waarin beoordelaars het oneens zijn'. Daarna is het immers begrijpelijk dat de beoordelaars hoog correleren, maar bewijst dat niets meer over de waarde van hun oordeel. Het is daarom belangrijk dat de beoordelaars volstrekt onafhankelijk van elkaar (zonder overleg en zonder communicatie over specifieke gevallen) werken.

De psychometrische benadering lijkt daarmee duidelijke voordelen boven de traditionele benadering te hebben: de schrijfvaardigheid van de studenten kan betrouwbaar gemeten worden, de deskundigheid van de beoordelaars wordt gekwantificeerd, de nauwkeurigheid van de meting (de betrouwbaarheid) kan berekend worden. Het is daardoor verleidelijk deze benadering als de enig juiste te zien. De aanhangers van de traditionele benadering zijn echter geneigd hetzelfde te denken van die benadering.

Een voorbeeld van een deze tegenstelling rond het meten van schrijfvaardigheid wordt beschreven in hoofdstuk 8: de discussie tussen voor- en tegenstanders van objectieve tests bestaande uit meerkeuze vragen om schrijfvaardigheid te meten. De aanhangers van de psychometrische benadering stellen dat objectieve tests sneller, goedkoper en betrouwbaarder meten en minstens zo valide zijn. De aanhangers van de traditionele benadering stellen dat een objectieve test niet schrijfvaardigheid meet, maar leesvaardigheid, omdat de test scherp lezen vereist en niet laat schrijven. Een objectieve test voor het meten van schrijfvaardigheid is daarom niet valide. Beide partijen focussen daarmee op tegenoverliggende kanten van het meetproces. De ene partij kijkt voor de validiteit vooral naar het meetresultaat, de andere partij kijkt vooral naar hoe de meting tot stand kwam (de meetmethode).

Een complete beschrijving van een meetmethode moet echter zowel de meetmethode omvatten als de meetresultaten. Wanneer twee belangrijk verschillende meetmethodes dezelfde resultaten leveren, is dat niet een gegeven dat eenvoudigweg geaccepteerd moet worden als feit, zoals de aanhangers van de psychometrische benadering soms geneigd lijken te veronderstellen, maar een fenomeen dat verklaring verdient. Er lijken twee redenen te bestaan voor aanhangers van de psychometrische benadering om het verschil tussen twee meetmethodes die hetzelfde meetresultaat opleveren, als niet-relevant te zien. Allereerst is de focus op de meetresultaten en niet op hoe die precies verkregen zijn. De meetmethode wordt gezien als een black box. Ten tweede valt een meetmethode, bijvoorbeeld de vragen in een studietoets, niet eenvoudig te kwantificeren.

In hoofdstuk 8 blijken de twee bestaande methodes om schrijfvaardigheid te meten (holistisch oordeel en objectieve tests) en drie nieuwe methodes (aantal fouten per honderd woorden, TAVAN-score, kwaliteit holistisch oordeel) alle vijf soortgelijke resultaten te leveren. Uitgaande van een (psychometrische) resultaat-benadering is hier slechts sprake van een enkele factor, uitgaande van een instrument-benadering is hier sprake van vijf belangrijk verschillende meetinstrumenten. Beide kanten van het meetproces combinerend is er echter sprake van vijf belangrijk verschillende instrumenten die dezelfde factor meten: een resultaat dat vraagt om verklaring. In hoofdstuk 8 wordt een verklaringsmodel geschetst.

In bijlage 1 worden enkele psychometrische en statistische begrippen kort toegelicht die in de psychometrische benadering vaak gebruikt worden. Het gaat om de termen: percentage verklaarde variantie, correlatie als t-test, gemiddelde onderlinge correlatie, standaardiseren, betrouwbaarheid en maximale correlatie, Spearman-Brown formule voor testverlenging, coëfficiënt alfa en de correctie voor onbetrouwbaarheid.

1.4 Opbouw

In hoofdstuk 2 wordt ingegaan op factoren waardoor het hedendaagse schrijfonderwijs mogelijk faalt en worden een aantal uitgangspunten geformuleerd voor een meer effectieve aanpak. Vervolgens wordt in hoofdstuk 3 ingegaan op de problemen rond onderwijsevaluatie en -constructie. De informatie in deze twee hoofdstukken vormt het kader voor de beoordeling van de bestaande onderwijsmethodes, het nieuw geconstrueerde programma en de evaluatie daarvan op leerwinst.

De zeven in paragraaf 1.2 geformuleerde onderzoeksvragen vormden het onderwerp van zes deelstudies die in de hoofdstukken 4 tot en met 9 worden behandeld.

De eerste onderzoeksvraag, wat het niveau is van basale schrijfvaardigheid voor Nederlands bij eerstejaarsstudenten in het hoger onderwijs, is beantwoord in de eerste deelstudie: het foutenonderzoek (hoofdstuk 4). In het foutenonderzoek werd aan de hand van een steekproef uit teksten van eerstejaarsstudenten onderzocht of fouten objectief vaststelbaar zijn en hoeveel fouten voorkomen (per honderd woorden). Ook is nagegaan welke soorten fouten gemaakt worden. Het aantal objectief constateerbare fouten werd vervolgens als uitgangspunt gebruikt om 'basale schrijfvaardigheid' te kwantificeren.

De tweede onderzoeksvraag, naar de waarde van bestaande methodes om basale schrijfvaardigheid te verbeteren bij eerstejaars hbo-studenten, is beantwoord in de tweede deelstudie: de beoordeling van taalmethodes (hoofdstuk 5). Bij de beantwoording van deze onderzoeksvraag werd getracht via een beoordeling aan de hand van een beoordelingsschema een overzicht te krijgen van de didactisch sterke en zwakke punten van bestaande methodes om basale schrijfvaardigheid te vergroten.

De derde onderzoeksvraag, hoe een nieuw onderwijsprogramma eruit moet zien om basale schrijfvaardigheid bij hbo-studenten te verbeteren, is beantwoord in de derde deelstudie: de constructie en het effect TAVAN-programma (hoofdstuk 6). Uitgangspunten voor de beantwoording van deze vraag waren de uitkomsten van het foutenonderzoek en de beoordeling van de bestaande taalmethodes. De opzet gaat uit van herschrijfopdrachten en een online-programma dat zorgt voor onmiddellijke feedback.

Ook de vierde onderzoeksvraag naar het effect van dit nieuwe onderwijsprogramma op de basale schrijfvaardigheid van eerstejaars hbo-studenten, wordt in deelstudie drie (hoofdstuk 6) behandeld. Voor de beantwoording van deze onderzoeksvraag werd de effectiviteit van het bestaande onderwijsprogramma en het nieuwe onderwijsprogramma bepaald via een voor- en nameting en was het vervolgens mogelijk beide programma's met elkaar te vergelijken.

De vijfde onderzoeksvraag, in hoeverre taalfouten in een tekst effect hebben op de waardering van die tekst door de lezer, is behandeld in de vierde deelstudie (hoofdstuk 7). Deze vraag is beantwoord door vast te stellen wat het verband was tussen het aantal fouten per

honderd woorden en het holistische oordeel in door studenten geschreven teksten. Vervolgens is via een experimentele opzet nagegaan of correctie van fouten in teksten leidde tot een positiever oordeel bij lezers.

De zesde onderzoeksvraag naar manieren voor het meten van schrijfvaardigheid wordt behandeld in deelstudie 5 (hoofdstuk 8). In totaal werden drie nieuwe manieren aangetoond om schrijfvaardigheid vast te stellen die criteriumvalide bleken te zijn.

De zevende onderzoeksvraag naar de benodigde tekstlengte voor het betrouwbaar beoordelen van schrijfvaardigheid wordt behandeld in deelstudie 6. Bij beoordeling op aantal fouten per honderd woorden bleek een halve A4 (250 woorden) een betrouwbaar oordeel over de basale schrijfvaardigheid van de student op te leveren. Dit lijkt in afwijking van het aantal teksten dat bij holistische beoordeling benodigd is om dezelfde betrouwbaarheid te bereiken. De mogelijke verklaring van deze discrepantie komt daarna aan de orde.

2

Falend en succesvol schrijfonderwijs

In dit hoofdstuk wordt ingegaan op de factoren waardoor schrijfonderwijs vermoedelijk vaak faalt. Een eerste punt is de lastige meetbaarheid. Beoordelaars zijn het onderling niet eens over wat een goede tekst is. Een tweede probleem is dat schrijven ingeoefend moet worden, terwijl het nakijken veel tijd kost en de vooruitgang niet zichtbaar is. Een derde probleem is wegkijken en ontkennen ('dat hebben ze toch al geleerd?'). Het nieuwe schrijfonderwijs vormt wellicht een vierde factor: de geproduceerde tekst is niet langer belangrijk, maar de beschrijving van het schrijfproces. Verder kunnen stellige beweringen van onderzoekers soms een negatieve rol gespeeld hebben. Enkele onderzoeken worden besproken in 2.2. Voorbeelden waaruit blijkt dat schrijfonderwijs een belangrijk effect kan hebben, zijn er ook en komen aan de orde in 2.3. In 2.4 wordt geschetst hoe effectief schrijfonderwijs er mogelijk uit zou moeten zien.

2.1 Factoren falend schrijfonderwijs

Moeilijk direct meetbaar

Als we aannemen dat de basale schrijfvaardigheid van studenten in het hoger onderwijs vaak te wensen overlaat, is de volgende vraag waarom het niet lukt studenten goed te leren schrijven.

Het eerste punt werd al vermeld in paragraaf 1.1. Schrijfvaardigheid is lastig of niet kwantificeerbaar. Doelen worden bijna altijd gesteld in algemene bewoordingen en daardoor is het moeilijk dwingende eisen te stellen aan het niveau van de schrijfvaardigheid. Purves (1992) merkte naar aanleiding van het mislukken van het tien jaar durende internationale IEA-onderzoek naar schrijfvaardigheid op: "School writing is an ill-defined domain" (p. 109).

Dat schrijfvaardigheid lastig kwantificeerbaar is, geldt overigens alleen voor directe metingen en niet voor indirecte metingen, zoals objectieve taaltests bestaande uit meerkeuzevragen. Dit wordt verder besproken in 4.1. Werken met objectieve taaltests is niet echt ingeburgerd. Het is moeilijk voorstelbaar dat een objectieve test iets totaal anders dan het schrijven van een tekst, goed kan voorspellen. Multiple-choice toetsen worden daarom in de praktijk vaak nog niet gezien als een aanvaardbaar alternatief.

Al in de jaren twintig van de vorige eeuw werd uit onderzoek duidelijk dat verschillende beoordelaars door studenten geschreven teksten zeer uiteenlopend beoordelen en onderling amper overeenstemmen. Cooper (1984, p. 1) stelde in een overzicht van onderzoek naar het meten van schrijfvaardigheid: "At first it was simply assumed that one must test writing ability bij having examinees write. But during the 1920s and 1930s, educational psychologists began experimenting with indirect measures because essay scorers (also called 'readers' or 'raters') were shown to be generally inconsistent, or unreliable, in their ratings."

Palmer beschreef in 1966 (p. 316) de situatie als volgt.

> *Sixty years of Board English testing have amply proved that essay tests are neither reliable nor valid, and that, whatever there faults, objective English tests do constitute a reliable and valid methode of ascertaining student compositional ability. Such a conclusion was very painfully and reluctantly arrived at. It struck at the heart of beliefs cherished by the teaching profession, and especially the English teaching branch of the profession. It made a villain and a fraud of one of pedagogy's oldest servants of all work. And it put in his place a monstrous testing device that asked the student to do virtually nothing except draw tiny marks on a sheet of paper, marks that would later be counted and tabulated by a soulless machine. Could there be virtues to such a testing monster?*

De holistische methode die taaldocenten gebruikten om essays te beoordelen blijkt niet betrouwbaar te zijn en (daardoor) niet valide. De docenten hebben altijd geclaimd een soort absoluut oordeel te hebben op hun gebied en nu blijkt uit onderzoek het tegendeel. Maar dat is nog niet alles. Objectieve tests blijken het beoordelen betrouwbaarder en beter te kunnen dan de docenten Engels met hun jarenlange opleiding en ervaring. Kortom, docenten zagen de objectieve tests niet als een ondersteuning, maar eerder als een bedreiging.

De conclusie dat holistische beoordeling niet betrouwbaar en niet valide was, bleek echter voorbarig. In hetzelfde jaar toonden Godshalk, Swineford en Coffman aan dat schrijfvaardigheid via het 'holistische' oordeel van beoordelaars wel betrouwbaar (0.841) gemeten kon worden door per student te werken met vijf essay-opdrachten en per opdracht vijf beoordelaars in te zetten (Godshalk et al., 1966). Voor toepassing in de onderwijspraktijk is deze methode door het grote aantal beoordelaars en essays echter niet bruikbaar.

Tegelijkertijd bevestigde deze studie de eerdere resultaten. Twee beoordelaars van dezelfde essay-opdracht bleken gemiddeld slechts 0.386 met elkaar te correleren; twee beoordelaars die verschillende essay-opdrachten beoordeelden waren het nog veel minder met elkaar eens met een gemiddelde correlatie van 0.263 (Coffman, 1966, p. 154). Het oordeel over een student hing daardoor ook nog sterk af van de specifieke essay-opdracht als er slechts één opdracht werd gebruikt.

Doordat per essay-opdracht vijf beoordelaars werden gebruikt, was de beoordelaarsbetrouwbaarheid van de totaalscore op de essay-opdracht 0.76. De gemiddelde onderlinge correlatie tussen de essay-scores, de scorebetrouwbaarheid, bedroeg echter slechts 0.52. Na correctie voor onbetrouwbaarheid van de beoordeling werd dit 0.68 (Coffman, 1966, p. 154). Wanneer de essays perfect betrouwbaar beoordeeld werden (oneindig veel beoordelaars) bedroeg de correlatie tussen de essays toch niet meer dan 0.68. Dit betekent dat er per student liefst vijf of meer essay-opdrachten benodigd waren die allemaal door vijf beoordelaars werd beoordeeld, om tot een betrouwbare totaalscore te kunnen komen.

Meuffels (2002) zag als het kernprobleem bij het bepalen van schrijfvaardigheid eveneens dat menselijke beoordelaars het in de praktijk vaak niet met elkaar eens blijken te zijn. Ze zouden het niet eens zijn met andere beoordelaars, maar ook niet met zichzelf. Bij een herbeoordeling van teksten door dezelfde beoordelaar verschilden de beide oordelen vaak. Meuffels haalde in dit verband een onderzoek aan van Wesdorp (1983), waaruit bleek dat bij ervaren docenten een verschil van drie punten kon zitten tussen hun eerste en tweede beoordeling. Als een opstel eerst beoordeeld werd met een 4, kon het bij de tweede beoordeling een 7 krijgen.

Deze kennis over het beoordelen van schrijfproducten zal voor docenten niet motiverend gewerkt hebben om schrijfopdrachten te geven en die daarna te beoordelen. Docenten konden terecht claimen dat de tijdrovende beoordeling vrijwel volstrekt subjectief was. Mogelijk heeft onderzoek vaker negatief ingewerkt op het schrijfonderwijs. Op deze mogelijkheid wordt verder ingegaan in paragraaf 2.2.

De lastige kwantificeerbaarheid van schrijfvaardigheid heeft vermoedelijk veel effecten. Aan leerlingen en studenten worden geen duidelijke eisen gesteld en aan scholen en opleidingen evenmin. Leerlingen zien geen duidelijke vooruitgang na schrijfonderwijs en raken daardoor hun motivatie kwijt en voor docenten geldt hetzelfde. Welk schrijfonderwijs wel

werkt en welk niet, wordt niet automatisch duidelijk, waardoor de meest effectieve vormen niet boven komen drijven.

Basisscholen zijn wettelijk verplicht de kwaliteit van hun onderwijs te bewaken. Dit geldt ook voor het stelonderwijs (het schrijven van teksten). Uit het rapport van de onderwijsinspectie voor het basisonderwijs (Inspectie van het Onderwijs, 2010) blijkt echter dat dit vertaald wordt in procesvariabelen. Hoe werd het onderwijs gegeven? Er wordt dus niet gekeken of leerlingen daadwerkelijk iets geleerd hebben van het gegeven onderwijs (p. 14). Dit laatste punt, de lastige evaluatie van schrijfonderwijs op het gerealiseerde effect, houdt verband met het algemene onderwijsevaluatie-probleem. Onderwijs wordt doorgaans niet geëvalueerd op het bereikte leerresultaat. Hierop wordt in 3.1 uitgebreider ingegaan.

Een punt dat hiermee samenhangt, is dat onderwijsinstellingen niet afgerekend worden op leerresultaten. Wanneer studenten in de vier jaar van hun studie tweemaal zo veel kennis en vaardigheden opdoen als voorheen, leidt dit niet tot extra financiering. Bij de samenstelling van het studieprogramma spelen wel de kosten voor de instelling een belangrijke rol, maar resulteert een groter leereffect niet in extra financiering. Effectievere onderwijsmethodes zoals Personalized System of Instruction (PSI) worden daardoor vaak slechts toegepast zo lang ze gezien worden als 'nieuw' (Fox, 2004, p. 206). Een effectievere methode wordt vermoedelijk pas blijvend ingevoerd, wanneer de kosten-batenverhouding voor de docenten en voor de instelling gunstiger is dan voorheen.

Moeizaam oefenen

Schrijven is ploeteren, wordt wel gesteld. Schrijfonderwijs waarbij de docent daadwerkelijk laat schrijven, is dat echter niet minder. Oefenen en laten oefenen met schrijven is arbeidsintensief en frustrerend. Schrijfopdrachten voor de leerling zijn vaak vaag en omvangrijk. De docent zit eerst met de lastige taak de onwillige leerlingen te motiveren en vervolgens met de arbeidsintensieve taak de talloze sterk op elkaar lijkende schrijfproducten te moeten nakijken en becommentariëren. Voor de docent die dat in een week tijd moet doen voor dertig werkstukken en soms voor nog veel grotere aantallen, valt dit naast zijn gewone werk bijna niet te realiseren. Docenten dienen op meerdere aspecten te letten en doorgaans ontbreekt de gelegenheid voor het leveren van feedback op correct formuleren (Zwiers, 2010). De Expertgroep Doorlopende Leerlijnen Taal en Rekenen (2008) stelde dat een docent in de

tweede fase al snel 250 leerlingen heeft (p. 13). Bij dat aantal leerlingen valt af te leiden dat een schrijfopdracht waar de docent gemiddeld 10 minuten naar kijkt al meer dan 40 uur nakijken vergt. Volgens Weerman, voorzitter van de Vereniging van Neerlandici, hebben docenten in het voortgezet onderwijs te weinig tijd om te corrigeren, omdat ze een beperkt aantal uren per klas beschikbaar hebben (Castagna, 2008). Gilbert (2004, p. 361) beweerde overigens: "Most students rarely write in school."

Als desondanks feedback gegeven wordt, is het trage feedback (Gilbert, 2004). De student krijgt normaal het commentaar van de docent op zijn schrijfproduct pas na een week of langer, zodat hij ondertussen al vergeten is waar het commentaar precies betrekking op heeft. De docent die feedback geeft op een tekst, zit verder met het probleem dat zijn commentaar te vaag is of juist te expliciet. In het eerste geval begrijpt de student het commentaar niet en in het tweede geval geeft de docent de oplossing weg. In beide gevallen leert de student uiteindelijk weinig of niets van het commentaar. Bovendien zijn studenten vaak niet echt gemotiveerd het commentaar van de docent nog te bestuderen en te verwerken. Zonder verplicht opnieuw inleveren van de tekst of een hoger cijfer na revisie, is de kans groot dat het commentaar van de docent geen leereffect sorteert. Dit maakt de feedback bij het nakijken van teksten in de onderwijspraktijk vaak vrijblijvend, terwijl het geven van dergelijke feedback zeer arbeidsintensief is. Het kost met andere woorden veel tijd en levert weinig op.

De Expertgroep Doorlopende Leerlijnen Taal en Rekenen (2008) merkte in dit verband op: "Maar de consequenties van het maken van fouten in spelling, formulering, stijl en genreconventies zijn zeer gering: een leerling ziet kringetjes, ziet het cijfer voor het werkstuk of de repetitie, en is allang blij met het resultaat. Zo lang het schoolbeleid is dat onverzorgde teksten qua vorm en taalgebruik geaccepteerd worden, dan gaat daarvan het signaal uit dat vorm en taalgebruik er niet toe doet. Zelfs voor een docent Nederlands is het bijna ondoenlijk om aandacht te besteden aan de kwaliteit van de inhoud, genreconventies, formulering, stijl en taalgebruik. Teveel leerlingen moeten teveel teksten produceren zodat er geen tijd is om precieze feedback te geven op verschillende kwaliteiten van de tekst" (p. 13-14).

Rijlaarsdam et al. (2008) stelden: "Once upon a time, writing education used to be simple. There was a writing task, students wrote a text and handed it in. . . . The teacher delivered feedback on the text, usually a grade and maybe some marginal comments, and that was that. In spite of the feedback, most teachers perceived little progress in the writing ability of their students" (p. 54).

Bacon en Scott Anderson (2004) vonden dat een schrijfcursus waarbij de studenten wel feedback kregen, maar de schrijfprestaties geen gevolgen hadden voor de beoordeling, geen effect sorteerde. Dezelfde cursus bleek echter in de groep waarin de schrijfprestaties voor 5% meetelden bij de beoordeling van een toets, een groot effect te sorteren. Vrijblijvende feedback werkt kennelijk niet; de student moet eerst gemotiveerd worden iets met die feedback te doen.

Moeizaam oefenen past vermoedelijk ook niet bij de tijdgeest. Onderwijs is niet meer een voorrecht, maar een recht. Onderwijs is niet langer ploeteren, maar moet in de eerste plaats leuk zijn. Schrijven en schrijfonderwijs is echter vaak vooral moeizaam en niet altijd even leuk en inspirerend. Het vereist daarmee een bepaalde instelling die in een bepaalde tijd en cultuur niet altijd even sterk aanwezig is.

Illustratief in dit verband zijn uitkomsten die De Groot (1993) ontleende aan Stevenson (p. 72-75). Kinderen uit de VS (Chicago) en China (Beying) kregen een open vraag voorgelegd wat ze zouden wensen als een tovenaar hun wens zou vervullen. De antwoorden werden ingedeeld in: geld, dingen, fantastische wensen en schoolse aspiraties. Van de Amerikaanse kinderen wilde 10% het liefst beter worden op school, van de Chinese kinderen 73%. Op een soortgelijke vraag over de school wilden Amerikaanse kinderen vooral minder school (30%). Van de Chinese kinderen gaf 1% een dergelijk antwoord. De meeste Chinese kinderen (57%) wilden een beter lokaal en betere lesboeken. Dat de Chinese kinderen belangrijk gemotiveerder waren voor school, lijkt daarmee duidelijk. Of dat verschil volledig mag worden toegeschreven aan de cultuur is echter de vraag. Uit het onderzoek bleek op de Chinese scholen per lesuur viermaal zoveel tijd besteed te worden aan de lesstof als op Amerikaanse scholen. Qua resultaten bleken de Chinese kinderen op de vier gebruikte tests vrijwel allemaal (90% tot 98%) boven het Amerikaanse gemiddelde te scoren. Cultuur en onderwijs grijpen in elkaar en beïnvloeden elkaar. De Groot (1993) verwoordde het als volgt: "Ik denk dat onze huidige maatschappij . . . het ontstaan van steeds meer 'domoren' bevordert. We worden het straks allemaal, voor zover we het al niet zijn" (p. 141).

Een aan de de cultuur gerelateerd punt is de invloed van de tv en de ontlezing. Jongeren kijken gemiddeld veel uren per week tv, film of video, ook via internet. Al die tijd lezen ze niet en schrijven evenmin. Otter en Schoonen (1996) rapporteerden dat leerlingen in groep 8 gemiddeld nog geen vier minuten per dag voor hun plezier een boek lezen en dat het lezen van kranten geen gewoonte is. Per dag wordt gemiddeld anderhalf uur tv gekeken.

'Geen probleem' en 'Niet ons probleem'

Een derde factor die niet bevorderlijk lijkt om studenten in het hoger onderwijs goed te leren schrijven, is de veronderstelling dat zij dat al geleerd hebben of in ieder geval geleerd behoorden te hebben, in het voorgaande onderwijs. In het hbo en op de universiteit gaat men er vanuit dat de student dat geleerd heeft in het voortgezet onderwijs. In het voortgezet onderwijs gaat men er vanuit dat de leerling dat grotendeels geleerd heeft in het basisonderwijs.

Al in het basisonderwijs gaat men er vermoedelijk vaak gemakshalve vanuit dat leerlingen veel beter kunnen schrijven dan ze daadwerkelijk kunnen. Leerlingen worden op een gegeven moment geacht volledige werkstukken te kunnen schrijven, maar omdat dit tijd vergt en tegenwoordig met een computer beter gaat, mag of moet de leerling dat vaak thuis doen. Via internet wordt dan met knippen en plakken, waarbij soms het hele gezin betrokken is, een fraai werkstuk van meerdere vellen A4 geconstrueerd. De docent beoordeelt dit globaal, geeft eventueel nog wat subjectieve kritiek en concludeert dat de leerling in staat is mooi werk te leveren dat zeker op het gewenste niveau ligt. Met die conclusie is de leerling blij en de docent vermoedelijk niet minder, omdat daarmee een lastig probleem is opgelost.

Dat de leerling veel van de gebruikte zinnen onmogelijk zelf geformuleerd kan hebben, is een detail dat de docent niet opvalt of negeert. Als leerlingen op school al zelfstandig iets schrijven, wordt dat meestal globaal beoordeeld. Fouten op zinsniveau worden daardoor gemakkelijk over het hoofd gezien, zodat docenten soms maar een vage indruk hebben van het feitelijke schrijfniveau van hun leerlingen.

Als de docent zelf maar een beperkte schrijfvaardigheid bezit, wordt het vervolgens extra aantrekkelijk om met de groep niet onnodig lang stil te staan bij dit pijnlijke punt. Uit onderzoek van Scheerens en Bosker (1997) en Hanushek en Rivkin (2010) bleek dat voor de resultaten van onderwijs de kwaliteit van de docenten van grote invloed is. In de Nederlandse situatie hebben veel docenten in het basisonderwijs evenwel een mbo-achtergrond (Van der Steeg, Vermeer & Lanser, 2011) en kunnen soms zelf slechts beperkt schrijven en zeker niet foutloos.

Geconfronteerd met een concreet schrijfproduct van een student in het hoger onderwijs, merkt een docent vaak wel dat er iets niet klopt. Doorgaans is die docent echter niet aan-

gesteld om de student te leren schrijven. Als de docent wel een onderdeel op het gebied van schrijfvaardigheid geeft, zijn zijn mogelijkheden in de praktijk uiterst beperkt. Hij kan niet afwijken van het programma en geen individuele lessen geven of voorschrijven. Verder kan hij of de instelling terecht opmerken dat leren schrijven op een basaal niveau niet een taak is, die in het hoger onderwijs thuishoort en dat het dus zijn verantwoordelijkheid niet is. Wanneer de docent de studenten wel laat schrijven en de producten becommentarieert, maakt hij zich daarmee doorgaans bij de studenten niet populair en loopt kans op een negatieve beoordeling via de door de studenten in te vullen evaluatievragenlijst.

Testen op een bepaald minimumniveau van schrijfvaardigheid wordt bovendien vaak als een oneigenlijke activiteit gezien. De opleiding is er om de student iets te leren, niet om het niveau van de student te controleren. Van der Westen (2011a, p. 4) stelde bijvoorbeeld: "Helaas, in plaats van in te zetten op leren en verwerven van de benodigde taal . . . zet het onderwijs nu juist in op *controleren.*"

Lage prioriteit

Bij de New SAT (Scholastic Aptitude Test) die in de Verenigde Staten gebruikt wordt voor toelating tot instellingen van hoger onderwijs, bleek de schrijfvaardigheidstest de beste voorspeller te zijn van studiesucces (Atkinson & Geiser, 2009). Atkinson (2009) beweerde: "Learning to write is of critical importance." In het Nederlandse onderwijs heeft schrijven echter een lage prioriteit. Op het gebied van Nederlands schrijfonderwijs wordt onvoldoende gepresteerd en die tendens begint al in het basisonderwijs (Tijd voor onderwijs, 2008; Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2008). Volgens Van den Akker, directeur van de Stichting Leerplanontwikkeling en hoogleraar Onderwijskunde aan de Universiteit Twente, is het oefenen met spelling, zinsbouw en grammatica op de achtergrond geraakt (Bouma, 2007b).

In het voortgezet onderwijs is de situatie niet anders. Van de Gein (2010) beweerde dat met name in het voortgezet onderwijs te lage eisen gesteld worden bij spelling. Spelling en formuleren krijgen weinig expliciete aandacht. Volgens Rijlaarsdam (2011) gaat de meeste tijd in de bovenbouw van het voortgezet onderwijs naar andere dingen dan schrijfonderwijs. Bonset en Braaksma (2008) merkten op: "De respondenten in de onderzoeken voor 1997 vonden schrijven als doelstellingengebied belangrijk, maar zetten het op de derde plaats (na

mondelinge taalvaardigheid en lezen) . . . . Tegelijkertijd toonden de tekortscores in de verschillende doelstellingenonderzoeken vooral schrijftaken" (p. 133). Onderzoek na 1997 op dit punt is niet verricht.

Sinds 1998 maakt het onderdeel Schrijfvaardigheid geen deel meer uit van het centraal examen havo en vwo (Bonset & Braaksma, 2008). Bij het onderdeel Leesvaardigheid (vragen over een tekst beantwoorden en een samenvatting maken) van het centraal schriftelijk vwo-examen Nederlands mocht in 2011 maximaal twintig procent van het cijfer worden afgetrokken voor taalfouten. Een leerling die in beginsel een 6,9 wist te halen, haalde met honderd taalfouten nog steeds een voldoende. Bij het centraal schriftelijk vwo-examen in 2013 werden spelling en grammatica helemaal niet beoordeeld bij de antwoorden op de open vragen van het onderdeel Leesvaardigheid. In het middelbaar beroepsonderwijs (mbo) werd het vak Nederlands vaak afgeschaft en geïntegreerd in andere vakken.

Het nieuwe schrijfonderwijs

Een nieuwe aanpak van het schrijfonderwijs is de vijfde factor die geleid kan hebben tot een lagere schrijfvaardigheid. Werken met schrijfopdrachten is door het vele werk en de onduidelijke opbrengst niet populair bij docenten en studenten. Ook vanuit wetenschappelijke hoek kwam er kritiek op de 'leren schrijven door te schrijven' methode (in 2.2 is dit verder uitgewerkt). Hayes en Flower (1980) veronderstelden dat schrijven bestond uit drie subprocessen: plannen, schrijven en reviseren. Kinderen bleken niet volgens deze drie subprocessen te werken en moesten dat dus leren. Dit leidde tot de opkomst van het procesgerichte schrijfonderwijs.

De Inspectie voor het Onderwijs (2010) beschreef deze overgang naar het 'nieuwe' schrijfonderwijs als volgt. "Van oudsher had het schrijfonderwijs een sterk productgerichte aanpak: de tekst stond centraal. De leerlingen kregen in schrijflessen gewoonlijk een opdracht om een tekst te schrijven, die dan achteraf van opmerkingen werd voorzien. De schrijfopdracht was meestal erg open en er werd nauwelijks instructie gegeven of geoefend" (p. 12). Dat veranderde toen wetenschappelijk onderzoek verricht werd naar het verloop van schrijfprocessen. De onderwijsinspectie noemde onder andere het onderzoek van Hayes en Flower (1980, 1981, 1986) en Bereiter en Scardamalia (1987). "Vooral onder invloed van het werk van Hayes en Flower heeft zich een verschuiving voltrokken van een productge-

richte aanpak van het schrijfonderwijs naar procesgericht schrijfonderwijs. . . . Hieronder verstaat men dat leerlingen moeten leren hoe ze een schrijftaak kunnen aanpakken. Een dergelijke benadering wordt ook strategisch schrijfonderwijs genoemd" (p. 12).

Er wordt aandacht besteed aan schrijfstrategieën met behulp van instructie en hardopdenkgesprekken met leerlingen. Praten over schrijven is een wezenlijk aspect van de didactiek: de docent praat of de leerlingen praten onderling over de geschreven teksten om tot reflectie te komen. Leerlingen leren hierdoor te reflecteren op hun teksten en hun aanpak (Inspectie van het Onderwijs, 2010, p. 12).

De onderwijsinspectie verwerkte de nieuwe inzichten in een beoordelingslijst om basisscholen op schrijfonderwijs te beoordelen. Voor een voldoende beoordeling moest bij schrijfopdrachten de tekstsoort aangegeven worden, het tekstdoel met eventueel het lezerspubliek en instructie over het schrijfproces gegeven worden. Leerlingen moesten de schrijftaak zo uitvoeren dat duidelijk was dat ze planden. De docent moest zich ten slotte in zijn commentaar beperken tot de organisatie, opbouw en inhoud van de tekst (Inspectie van het Onderwijs, 2010, p. 21). Het laatste betekende dat de docent niet langer commentaar mocht geven op fouten in de tekst en niet goed lopende zinnen.

De in de praktijk uitgetrokken tijd voor het schrijfonderwijs hield tegelijkertijd niet over. Meer dan de helft (60%) van de basisscholen besteedde maximaal 45 minuten, vaak belangrijk minder, per week aan het schrijven van teksten. Dat is minder dan 10% van de totale tijd die beschikbaar is voor taal. De onderwijsinspectie vond dit 'aan de krappe kant' (Inspectie van het Onderwijs, 2010, p. 5).

In die drie kwartier of minder per week moest niet alleen geschreven worden, maar vooral worden geleerd hoe het schrijven aangepakt moest worden. Het was de bedoeling dat leerlingen leerden te reflecteren op hun aanpak en te praten over het doel van de tekst. Het doel was ook dat ze leerden rekening te houden met hun lezerspubliek en leerden informatie te verzamelen, te selecteren en te ordenen. Ondanks al deze doelstellingen lukte het de overgrote meerderheid van de leerlingen (80%) niet, zelfs niet op het einde van het basisonderwijs, steeds grammaticaal correcte zinnen te schrijven. Ook het nadenken over de structuur en het plannen van de opzet leek niet altijd even effectief te zijn. Zo ontbrak bij bijna de helft van de 10% best schrijvende leerlingen de inleiding en het slot van de tekst (Inspectie van het Onderwijs, 2010, p. 36).

Dat de procesgerichte benadering in de praktijk niet altijd goed uitwerkte, viel ook af te leiden uit het rapport. De inspectie constateerde dat de hedendaagse taalmethodes meer aandacht besteden aan het procesgerichte schrijfonderwijs dan methodes in het verleden. In verhouding tot de voorgaande peiling gebruikten nu drie maal zoveel scholen een procesgerichte methode (Inspectie van het Onderwijs, 2010, p. 25). De verschuiving naar meer procesgerichte methodes resulteerde echter niet in een grotere schrijfvaardigheid (p. 6).

De inspectie stelde over de basisscholen die een procesgerichte benadering gebruikten, dat zij er niet in slaagden om de mogelijkheden van deze benadering voldoende in de praktijk te brengen (Inspectie van het Onderwijs, 2010, p. 14). "De vraag is of de lesuitwerkingen van procesgericht schrijven die doorgaans in de methodehandleidingen staan, de leraren in de praktijk voldoende houvast bieden. Verder is de vraag of leraren op vakdidactisch gebied voldoende zijn toegerust" (p. 6). Met andere woorden: het procesgerichte schrijfonderwijs leidde niet tot betere resultaten en bleek in de praktijk niet goed te werken, wat echter niet aan de nieuwe methode lag, maar aan de tekortschietende leraren.

Bij 179 willekeurig getrokken scholen beoordeelde de inspectie via schoolbezoeken het schrijfonderwijs in de groepen 5 tot en met 8 aan de hand van een zelf opgesteld beoordelingsschema. Om voldoende te kunnen scoren op schrijfdidactiek volgens dit schema moest een school procesgericht schrijfonderwijs geven (Inspectie van het Onderwijs, 2010, p. 21). Verder beschikte de inspectie over de objectieve scores op de Cito-eindtoets voor onder andere het schrijven van teksten. Op basis hiervan kon ze de scholen indelen als schrijfzwak, schrijfneutraal en schrijfsterk. Daarna werd gekeken of er significante verschillen waren in de kwaliteit van het schrijfonderwijs tussen schrijfzwakke en schrijfsterke scholen die deel uitmaakten van de steekproef (p. 34-35).

Het rapport van de inspectie vermeldde de uitkomst van deze significantietoetsing niet. Wel wordt gesteld: "Omdat de omvang van de verdeling schrijfsterk versus schrijfzwak van de scholen in dit onderzoek te beperkt is, bleek het niet mogelijk om op dit punt conclusies te trekken" (Inspectie van het Onderwijs, 2010, p. 35). Men had gegevens van 179 scholen en daarvan moet naar verwachting 25% schrijfzwak geweest zijn en 21% schrijfsterk (p. 35). Dat betekent dat er in de steekproef ongeveer 45 schrijfzwakke scholen zaten en 38 schrijfsterke. Voor een significantietoetsing is dat meer dan voldoende. Mogelijk was de reden dat

de uitkomst van de toetsing een andere conclusie leverde over het procesgerichte schrijfonderwijs dan volgens de ideeën van de inspectie zou moeten.

In een eerder verschenen rapport van de inspectie (Inspectie van het Onderwijs, 2009) werd wel expliciet een vergelijking gemaakt tussen taalsterke en taalzwakke scholen. Deze vergelijking was gebaseerd op de scores van de leerlingen op de taalvaardigheidsonderdelen van de Cito-eindtoets. Schrijfvaardigheid in de Cito-eindtoets correleert echter hoog met andere taalvaardigheden, zoals begrijpend lezen (.86) en woordenschat (.69), (Inspectie van het Onderwijs, 2010, p. 34) en daarom zullen taalsterke scholen vrijwel altijd ook schrijfsterk zijn en taalzwakke scholen, schrijfzwak. Doordat in dit onderzoek een totaal ander beoordelingsschema werd gehanteerd waarin de procesgerichte schrijfaanpak amper een rol speelde, leverde ook dit onderzoek geen duidelijk oordeel op over de waarde van het procesgerichte schrijfonderwijs.

Het 'nieuwe' schrijven had intussen wel gevolgen voor de gerichtheid op het schrijven van heldere en correcte teksten. Bij de productgerichte benadering moest de leerling een goede tekst produceren. Bij de procesgerichte benadering stond niet langer de tekst centraal, maar het proces. Als de tekst fouten vertoonde, was dat niet zo erg, als het proces waarmee de tekst geconstrueerd was, maar goed was. De doelstelling van het schrijfonderwijs was gewijzigd van het schrijven van een goede tekst naar het werken volgens een voorgeschreven aanpak en het kunnen praten over het proces.

2.2 Vaker negatieve rol onderzoek?

Door onderzoek dat aantoonde dat beoordelaars van schrijfproducten het vrijwel volledig niet eens zijn, werd het voor docenten niet gemakkelijker om teksten van studenten gemotiveerd te beoordelen. Aan hun oordeel kon immers weinig waarde worden gehecht. Hierop is in 2.1 ingegaan, evenals op de opkomst en de mogelijk negatieve gevolgen van het procesgerichte schrijfonderwijs. In deze paragraaf is dit verder uitgewerkt. Mogelijk heeft onderzoek vaker negatief ingewerkt op de inzet en activiteiten van docenten. Een aantal toonaangevende onderzoeken worden hierna in dit verband besproken.

Meer procesgericht schrijfonderwijs?

De basis voor de opkomst van het procesgerichte schrijfonderwijs lag in de 'ontdekking' van Hayes en Flower in 1980 dat schrijven opgesplitst kon worden in drie afzonderlijke deelprocessen. In de onderwijspraktijk werd dit opgevat als wetenschappelijke evidentie voor de gedachte dat schrijven bestaat uit drie hoofdprocessen: plannen, schrijven en reviseren (Inspectie van het Onderwijs, 2010, p. 12). Goed schrijven begon dus met goed plannen. Als leerlingen problemen hadden met schrijven, betekende dat, dat ze eerst beter moesten leren plannen.

De theorie van Hayes en Flower bleek aantrekkelijk voor onderzoekers van onderwijs. Bonset en Hoogeveen (2007) stelden weliswaar in een overzicht van 46 onderzoekspublicaties op het gebied van schrijven in het basisonderwijs: "Er is slechts één construerend onderzoek naar schrijfonderwijs verricht. Wij weten dus nagenoeg niets over de praktische bruikbaarheid en effectiviteit van aanpakken voor het schrijfonderwijs. Het is duidelijk dat hier een leemte ligt die dringend opgevuld moet worden" (p. 58). Met 'construerend onderzoek' werd onderzoek bedoeld dat betrekking had op de constructie van een onderwijsprogramma en het onderzoeken van de effectiviteit daarvan. De overige 45 publicaties hadden betrekking op diverse deelaspecten van het onderwijs (doelstelling, beginsituatie, onderwijsleeractiviteiten, onderwijsleersituatie, toetsconstructie, peilingsonderzoek), maar niet op de daadwerkelijk constructie van onderwijs met het bijbehorende effectonderzoek.

Hoewel dus amper onderzoek verricht is naar welke onderwijsmethodes wel en niet werken, werden de uitkomsten van 14 'descriptieve' onderzoeken door Bonset en Hoogeveen (2007) als volgt samengevat: "De resultaten van het descriptieve onderzoek naar de praktijk van het schrijfonderwijs stemmen niet optimistisch over het onderwijsaanbod. Bij het schrijven ontbreekt aandacht voor plannen, reviseren, reflecteren, interactie en samenwerking tussen leerlingen, voorlezen en presenteren van teksten aan elkaar" (p. 54).

Op basis van opvattingen over hoe schrijfonderwijs idealiter zou moeten zijn, werd door Bonset en Hoogeveen (2007) een negatief oordeel geveld over het onderzochte onderwijs. De noodzaak om die opvattingen eerst empirisch te toetsen, door een verbeterd onderwijsprogramma te construeren en dat te onderzoeken op effectiviteit, werd niet gezien. Het uitgaan van eigen opvattingen en de projecties daarvan vervolgens zien als werkelijkheid werd door Jaynes (1989, p. 7) aangeduid als 'mind projection fallacy'.

De aanbeveling die Bonset en Hoogeveen (2007) zelf ten slotte aan leerkrachten gaven, was: "om het productgerichte schrijfonderwijs waarin veel aandacht wordt besteed aan vorm- en verzorgingsaspecten om te buigen in de richting van procesgericht schrijfonderwijs" (p. 61). Belangrijke kenmerken van dit schrijfonderwijs zijn volgens Bonset en Hoogeveen nadruk op planning en schrijfstrategieën, de behandeling van verschillende tekstsoorten, de toepassing van genre- en tekstkenmerken en reflectie op de schrijfcontext, het schrijfproces en schrijfproduct. Teksten zouden becommentarieerd en besproken moeten worden, niet alleen met de docent, maar ook met de medeleerlingen.

De kritiek van Bonset en Hoogeveen (2007) op de gang van zaken in het schrijfonderwijs en de daaruit voortvloeiende aanbevelingen werd evenwel niet ontleend aan constructie- en evaluatie-onderzoek. Ook andere bronnen die wellicht als een soort onderbouwing voor de kritiek hadden kunnen fungeren, werden in dit geval niet vermeld. Meer voorbeelden waarin onderwijsonderzoekers stellige uitspraken menen te moeten doen over de 'slechte' didactische praktijken in het onderwijs, zijn gemakkelijk te vinden. Het punt waar het ons hier om gaat, is dat onderwijsonderzoekers aan de ene kant kennelijk nalaten het benodigde onderzoek te verrichten of daar althans niet in slagen, terwijl ze aan de andere kant wel stellige en kritische uitspraken doen alsof het onderzoek al verricht is en men al weet wat eruit komt. Men lijkt te gemakkelijk uit te gaan van de juistheid van de eigen opvattingen.

Alle onderwijsvormen effectief, behalve schrijfopdrachten?

In Engelstalig onderzoek werd iets soortgelijks gevonden. Graham & Perin (2007a) voerden een meta-analyse uit op onderzoeken naar 'adolescent writing instruction'. Het rapport was ook bedoeld voor onderwijsgevenden en schooldirecties. Zij rapporteerden dat de elf geselecteerde onderwijsvormen die in het rapport aan de orde kwamen allemaal effectief waren. "This report identifies 11 elements of current writing instruction found to be effective for helping adolescent students learn to write well" (p. 4).

De Research Brief van het Center for Comprehensive School Reform and Improvement (2007, p. 3) nam deze conclusie over en vermeldde in een reactie op dit rapport: "However, all of the Writing Next instructional elements have shown clear results for improving student writing." Als vrijwel iedere onderwijsmethode effectief blijkt te zijn om leerlingen schrijven te leren, hoe is het dan mogelijk dat hedendaagse studenten daar volgens veel

berichten problemen mee hebben? Deze vraag was de reden om het rapport nauwkeuriger door te nemen.

In bijlage B van het rapport (p. 43-66) werd een overzicht gegeven van de gebruikte onderzoeken met hun 'effect size' (het verschil in gemiddelde tussen de experimentele en controle groep uitgedrukt in standaarddeviaties) en welk onderwijsprogramma per groep gebruikt werd. De auteurs stelden in totaal '176 effect sizes' verzameld te hebben (Graham & Perin, 2007a, p. 25). In een tijdschriftartikel gebaseerd op deze meta-analyse (Graham & Perin, 2007b) was dit aantal teruggebracht tot 154. Bij het natellen van de vermelde effect sizes werden echter slechts 133 effect sizes geteld. Van deze 133 effect sizes bleken negen dubbel voor te komen, met zowel een min- als een pluswaarde (Troia & Graham, 2002; Anderson, 1997; Gamelin, 1996; Yeh, 1998; Saddler & Graham, 2005; Howie, 1979; Kanellas, Carifio & Dagostino, 1998; Pantier, 1999; Hayes, 1984). In het onderzoek van Troia & Graham (2002) werd bijvoorbeeld schrijfstrategie-onderwijs vergeleken met schrijfprocesonderwijs. De eerste conditie scoorde iets beter en werd daarom genoteerd als +.14 in de categorie 'schrijfstrategieën'. In de categorie 'schrijfprocesbenadering' werd hetzelfde onderzoek genoteerd als -.14.

Vijf onderzoeken bleken voor te komen met twee effect sizes. Dit was bijvoorbeeld het geval bij Hillocks (1982). In de eerste experimentele conditie namen de leerlingen deel aan enkele activiteiten waarna ze een schrijfopdracht kregen. In de tweede experimentele conditie namen ze ook deel aan de activiteiten, maar moesten ze na de schrijfopdracht hun teksten ten slotte nog een keer reviseren. In de controleconditie vond er eerst een discussie plaats en moesten ze daarna schrijven. Hoewel de tweede conditie door die revisiefase belangrijk verschilde van de eerste conditie, zijn beide condities opgenomen als 'inquiry activities'. Deze categorie bestond in totaal maar uit vijf effect sizes, zodat dit onderzoek daarmee twee vijfde uitmaakte van de totale categorie.

Ongeveer zestig effect sizes waren afkomstig uit een 'unpublished doctoral dissertation', zodat de kwaliteit onduidelijk was en moeilijk controleerbaar. Verder bleek een groot aantal (ongeveer zesentwintig) van de vermelde onderzoeken gebaseerd te zijn op kinderen met leerproblemen of op andere groepen kinderen die duidelijk afweken van gewone leerlingen. In de uitkomsten werden deze onderzoeken samengevoegd met onderzoek dat wel gebaseerd was op doorsneeleerlingen. Graham en Perin (2007a) stelden echter zelf: "the purpose of this review was to draw recommendations for writing instruction within regular school settings" (p. 35).

Bij onderzoek van onderwijs op schrijfgebied is een lastig probleem dat de beoordeling van de schrijfproducten onbetrouwbaar en subjectief is. Beoordelaars kunnen dus gemakkelijk onbewust een bepaalde conditie bevoordelen, zodra ze weten uit welke groep een schrijfproduct afkomstig is. De onderzoeker moet in dit verband speciale maatregelen nemen om te zorgen dat de beoordelaars volledig blind beoordelen. Dat kan door de schrijfproducten qua volgorde goed te mengen, ze daarna te nummeren en de namen te verwijderen en ervoor te zorgen dat iedere beoordelaar alle schrijfproducten onafhankelijk van de andere beoordelaars beoordeelt. Idealiter worden ook nog, om onjuiste correlaties door vermoeidheid te voorkomen, de schrijfproducten per beoordelaar in een andere volgorde beoordeeld.

Op dit belangrijke punt merkten de auteurs echter alleen op: "studies in which reliability for the writing quality measure was questionable were excluded. For example, studies that provided no information on interrater reliability or that provided no training to raters were eliminated. Studies in which low levels of interrater reliability (i.e., below 0.60) were reported were also eliminated" (Graham & Perin, 2007, p. 35). Een lage beoordelaarsbetrouwbaarheid maakt echter niet uit op het moment dat men een significant verschil vindt tussen twee condities, mits de beoordeling maar blind en onafhankelijk was. Voor de significantietest heeft de betrouwbaarheid geen toegevoegde waarde. De auteurs stelden hier een overbodige eis, terwijl ze nalieten duidelijkheid te eisen op het punt dat wel belangrijk was, namelijk dat de beoordeling volledig blind was uitgevoerd. Ook de eis dat de raters getraind moesten zijn, had weinig nut. Cooper (1984, p. 8-9) citeerde Breland en Jones (1982) die opmerkten: "When scoring the same set of papers -- even after careful instruction in which criteria are clearly defined and agreed upon -- teachers assign a range of grades to any given paper."

Bij veel onderzoek ontbrak een duidelijke controleconditie en werden in plaats daarvan twee verschillende onderwijsvormen vergeleken. Het resultaat was dat de vermelde effect sizes niet vergelijkingen waren ten opzichte van een nulconditie, maar ten opzichte van een andere onderwijsvorm. Zo leverde de categorie Grammatica-onderwijs een negatief effect (de grootte werd niet vermeld). Op basis van deze waarde concludeerden de auteurs: "This negative effect was small, but it was statistically significant, indicating that traditional grammar instruction is unlikely to help improve the quality of students' writing" (Graham & Perin, 2007a, p. 21).

Het onderzoek dat echter de grootste bijdrage (-1.40) aan dit negatieve effect leverde, was dat van Anderson (1997) waarin het effect van schrijfstrategie-instructie bij onder andere kinderen met leerproblemen op tekstkwaliteit werd vergeleken met grammatica-onderwijs als controleconditie. De schrijfstrategie-instructie leidde hier tot een beter schrijfproduct. Hieruit kan echter niet worden afgeleid dat leerlingen door grammatica-onderwijs slechter gaan schrijven. Op dit punt ontbreekt immers alle informatie.

Doordat de controlecondities waarmee vergeleken werd per categorie sterk konden verschillen, ontbrak een duidelijke vergelijkingsmaatstaf. De categorieën leverden een positief effect, maar ten opzichte waarvan was vaak niet duidelijk. De auteurs zelf concludeerden echter iets totaal anders. Alle elf onderwijsvormen waren volgens hen effectief om goed te leren schrijven. "It is important to note that all of the elements are supported by rigorous research" (Graham & Perin, 2007a, p. 4). Om dit aan te tonen had men echter per onderwijsmethode moeten werken met een pre- en posttest of met een echte nulconditie in plaats van te vergelijken met een totaal andere onderwijsmethode.

Voor de categorieën Samenvatten en Schrijfstrategieën werd een hoge gemiddelde effect size vermeld van beide keren .82. Wat opviel was de onevenwichtigheid in behandeling. Aan Samenvatten werden slechts zes regels tekst besteed en aan Schrijfstrategieën maar liefst vierenveertig. De categorie Samenvatten bestond uit slechts vier effect sizes waarvan maar twee betrekking hadden op doorsneeleerlingen.

Een van deze onderzoeken was het onderzoek van Chang, Sung en Chen (2002) dat betrekking had op leerlingen van 'grade 5'. Dit onderzoek leverde een effect size van 0.81: de samenvatgroep deed het belangrijk beter deed dan de controlegroep. Chang et al. (2002) zelf vermeldden echter drie 'concept mapping' groepen versus een controlegroep. Op basis van hun artikel konden drie effect sizes berekend worden: 1.17, 0.81 en 0.21. De middelste waarde is opgenomen, de overige twee niet. Als afhankelijke variabele gebruikten Chang et al. niet zoals Graham en Perin stelden een maat voor de 'the completeness and accuracy of the written summary' (Graham & Perin, 2007a, p. 35), maar een maat voor 'summarization efficiency' (Chang et al., 2002, p. 9). Dat wil zeggen, het aantal idee-eenheden gedeeld door het totale aantal woorden. Een leerling die het artikel samenvatte in één terecht woord, scoorde maximaal, namelijk 1. Een leerling die hetzelfde punt verwoordde in vijf woorden, scoorde slechts 0.2. In feite was dit dus vooral een maat voor beknoptheid in plaats van volledigheid en nauwkeurigheid.

Betekende de positieve uitkomst voor de categorie Samenvatten dat leerlingen door samenvattingen te leren maken, beter gingen schrijven? Het rapport gaf de volgende conclusie: "Overall, teaching adolescents to summarize text had a strong, positive effect on their ability to write good summaries" (Graham & Perin, 2007a, p. 16). Anders geformuleerd: als je leerlingen leert samenvatten, gaan ze beter samenvatten. Het was niet duidelijk of ze ook beter gingen schrijven.

Een soortgelijk probleem deed zich voor bij veel van de andere onderwijsvormen die volgens Graham en Perin effectief bleken. Bij 'collaborative writing' werkten leerlingen samen aan een tekst. Dit resulteerde in doorsnee in betere teksten. Iets wat valt te verwachten, omdat de best schrijvende leerling het niveau van de tekst verhoogt. Maar dat betekent niet dat leerlingen na 'collaborative writing' daadwerkelijk beter kunnen schrijven. Specifieke schrijfopdrachten versus globale schrijfopdrachten hebben invloed op de geproduceerde teksten, maar ook dat wil niet zeggen dat studenten daarna echt beter kunnen schrijven. Het geven van voorbeeldteksten heeft invloed op de tekstkwaliteit bij de daarop volgende schrijfopdracht. Maar schrijft de leerling vervolgens zonder voorbeeldtekst nog steeds beter?

Bij het op de eerste plaats genoemde en hoog scorende 'writing strategies' merkten de auteurs op: "Teaching adolescents strategies for planning, revising, and editing their compositions has shown a dramatic effect on the quality of students' writing. . . . The ultimate goal is to teach students to use these strategies independently" (Graham & Perin, 2007a, p. 15). De eerste zin geeft aan dat de daarna geschreven producten verbeterden; de laatste zin geeft aan dat het niet altijd duidelijk was of leerlingen de schrijfstrategie zelfstandig (bleven) toepassen zonder de instructies van de docent.

Het meest opmerkelijke aan het rapport, was dat een voor de hand liggende onderwijsvorm als een schrijfopdracht geven en het schrijfproduct beoordelen en bespreken, volledig ontbrak. Hierover merkten Graham en Perin (2007a) op: "increasing the frequency and amount of writing is generally proposed as an important practice because of widespread concerns about how little actual writing students do in schools (Applebee, 2000; National Commission on Writing, 2003), but evidence of a consistent effect is ambiguous. There were too few effect sizes, too much variability in effect sizes, too much diversity in the procedures used to promote extra writing time, and too many different comparison conditions to allow any reliable conclusions to be drawn about the impact of this approach" (p. 26).

Bij de overige onderzochte onderwijsvormen is het argument van verschillende vergelijkingscondities, hoewel vaak zeer van toepassing, nooit gebruikt. Ook het argument dat er te veel verschil zou zitten in de procedures 'to promote extra writing time' lijkt nogal gezocht. De auteurs vermeldden ook niet welke onderzoeken er dan wel waren en wat die voor effect opleverden. Er werd overigens niet gesproken van een negatief of verwaarloosbaar effect. Mogelijk lieten de beschikbare studies wel duidelijke effecten zien. De auteurs stelden nog: "As was claimed many years ago (Braddock & Jones,1969), it may well be that . . . providing more opportunities to write without effective instruction and motivation is not enough to improve writing quality" (Graham & Perin, 2007a, p. 26).

De boodschap aan de onderwijsgevende leek in ieder geval te zijn: doe in de klas wat je wilt, als het maar niet grammatica-onderwijs of een schrijfopdracht is.

Leren schrijven door te observeren?

Onderzoek dat studenten leren schrijven door te observeren is te vinden bij Raedts (2011), Rijlaarsdam et al. (2008) en Zimmerman en Kitsantas (2002). Rijlaarsdam et al. veronderstelden dat vroeger in het onderwijs meer daadwerkelijk geoefend werd met schrijfopdrachten, maar dat er voor de docent weinig vooruitgang zichtbaar was. In de jaren zestig veranderde het onderwijsmodel, volgens Rijlaarsdam et al., van leren om kennis te verwerven naar leren als een participatieproces: "in the 1960s, the paradigm of language education (L1) shifted towards language as a means of communication, moving from studying language as a system towards enhancing the language in communicative situations" (p. 54).

Moffett (1968) presenteerde een complete taalleertheorie. In zijn opvatting ontstond leren door taal te ervaren. Onderwijzers moesten zorgen voor leertaken met een echt publiek, zodat de lezers betrokken werden bij het schrijven. Volgens Elbow (1974) hing het succes van een tekst af van de subjectieve lezer. Schrijvers hadden daarom gelegenheid nodig om naar hun lezers te luisteren. Luisteren hoe verschillende lezers de tekst lazen, zou voldoende feedback leveren. Bruffee (1984) zag schrijfvaardigheid nog breder: "The way they [the students] talk with each other determines the way they will think and the way they will write" (p. 642). Hoe studenten praten en denken bepaalt uiteindelijk hoe ze schrijven. Schrijfonderwijs zou daarom kunnen volstaan met studenten beter te leren denken. Dit zou bereikt

kunnen worden door ze meer met elkaar te laten praten onder leiding van de docent. Hoewel de praktische voordelen van deze methode voor de docent evident zijn, lijkt de effectiviteit van deze methode voor het verbeteren van de schrijfvaardigheid nog even de vraag.

In de opvatting van Rijlaarsdam et al. (2008) leert een student niet door te schrijven, maar door te observeren. Wanneer twee medestudenten met elkaar praten of schriftelijk met elkaar communiceren kan een derde leerling die toekijkt, die observeert, leren. "To stimulate students' learning-to-write capacities, instruction should stimulate them to observe and evaluate relevant processes . . . . Designing writing lessons could be guided by the design rule that at least the Learner [Observer] role must be realized in the lessons" (p. 58).

De opvatting dat studenten schrijven moeten leren door te observeren, lijkt duidelijk. Men kan er vraagtekens bij zetten. Zijn er inderdaad overtuigende voorbeelden van schrijvers die alleen door observeren goed schrijven hebben geleerd? Het punt waar het hier echter om gaat, is dat dit soort opvattingen docenten niet motiveert om leerlingen en studenten te laten schrijven.

Meer focussen op plannen, minder op het product?

Mayer (2008) besteedde in zijn boek *Learning and Instruction* een apart hoofdstuk aan schrijfonderwijs en de problemen daarvan. Hij baseerde zich op een cognitief model dat hij ontleende aan Hayes en Flower (1980). Dit vereenvoudigde model gaat ervan uit dat schrijven bestaat uit drie deelprocessen: plannen (informatie genereren en organiseren), verwoorden en reviseren.

De evidentie die hij presenteerde om dit model te onderbouwen, zijn ontleend aan één hardop-denk-protocol van één proefpersoon die een schrijfopdracht hardop denkend probeerde uit te voeren. In de gepresenteerde grafiek komen wel het genereren, organiseren en verwoorden als zelfstandige fases terug, maar niet het reviseren. Mogelijk komt dit doordat met een vrij eenvoudige schrijfopdracht gewerkt werd, zodat de tekst niet eerst in concept geschreven werd en daarna nog eens bewerkt werd.

Echt vooraf plannen leek overigens niet te gebeuren. De pauzes tussen het verwoorden werden nu opgevat als teken van 'plannen'. Onderzoek van Gould (1978a, 1978b, 1980) en

Matsuhashi (1982, 1987) bevestigde dat (geciteerd in Mayer, 2008). "Interestingly, Gould's studies revealed that planning pauses occured during the writing process (as local planning) rather than before it (as global planning), suggesting that writers rarely plan before they start writing" (Mayer, 2008, p. 129). De empirische basis voor de driedeling van het schrijfproces lijkt daarmee nogal wankel geworden. Ook Lowyck en Vanmaele (1992) kwamen tot de conclusie dat leerlingen bij schrijfopdrachten nauwelijks procesgericht te werk gaan. Het (expert) model van Flower en Hayes kon volgens hen niet dienen als basis voor het schrijfgedrag van jonge leerlingen (geciteerd in Bonset & Braaksma, 2008).

Het ligt overigens voor de hand dat ervaren schrijvers eerst tijd besteden aan het verzamelen van informatie, nadenken over de indeling, een concept maken en dat vervolgens, vaak eindeloos, bewerken. Als model van hoe men zou moeten schrijven, voldoet het misschien prima. Vervolgens werden echter uit dit empirisch gezien, matig onderbouwde model, vergaande conclusies getrokken: "the foregoing analysis suggests that most of the time and effort in writing is devoted to planning rather than to actually producing acceptable text" (Mayer, 2008, p. 127). Dat is niet de enige conclusie die uit dit model werd afgeleid. "Much of the instruction in writing involves teaching procedures for producing sentences properly" (Mayer, 2008, p. 127). Dat vond Mayer niet verkeerd, maar schrijven was toch veel meer dan zoiets beperkts. Hij haalde Kellogg (1994) aan die volgens hem liet zien dat "in composing a written text, individuals . . . engage in a special form of thinking -- the making of meaning -- that may well define one of the most unique characteristics of our species" (geciteerd in Mayer, 2008, p. 127).

De geselecteerde zinnen van studenten in Tabel 1.1 lijken met deze conclusie schril te contrasteren. Moeten studenten eerst niet een correcte en betekenisvolle zin kunnen schrijven, voordat ze een betekenisvolle tekst kunnen produceren?

Een derde gevolgtrekking die werd gemaakt, is dat studenten vooral moeten leren hun publiek te beïnvloeden met hun schrijfproduct, terwijl de aandacht in het onderwijs vaak sterk ligt op 'correct' schrijven. Wie de uitspraken in Tabel 1.1 doorneemt, kan zich misschien voorstellen dat hedendaagse studenten uitermate goed zijn in het beïnvloeden van hun publiek via hun schrijfproducten, alleen niet op de manier zoals Kellogg bedoelde.

Bereiter en Scardamalia (1987) onderscheidden volgens Mayer drie schrijfstadia. In het eerste stadium hebben kinderen moeite met het genereren van ideeën. In stadium twee, kennis-

vertellen, hebben ze geen moeite meer met het genereren van ideeën, maar evalueren en organiseren ze nog niet. In stadium drie, kennis-transformatie, transformeren ze ten slotte ideeën door ze te evalueren en te organiseren. Onderwijs zou vooral gericht moeten zijn op het helpen van studenten om over te stappen van stadium twee naar stadium drie: "A major goal of writing instruction is to help students progress from a knowledge-telling approach to a knowledge-transforming approach" (Mayer, 2008, p. 128). Het probleem bij deze opvatting lijkt weer dat de veronderstelling dat studenten geen problemen hebben met het schriftelijk produceren van ideeën en informatie en alleen nog zouden moeten leren die te evalueren en opnieuw te ordenen, niet ontleend is aan de praktijk.

De belangrijkste gevolgtrekking die ten slotte uit het cognitieve model werd afgeleid, was dat studenten te weinig plannen en dat ze dus in de eerste plaats moeten leren plannen. "It follows that an important instructional intervention is to encourage students to engage in a range of planning subprocesses, including generating, evaluating, and organizing, before they begin to write" (Mayer, 2008, p. 131). Studenten moeten vooral beter leren plannen, om beter te gaan schrijven, stelde Mayer. Het gaat niet om het feitelijke schrijven, maar om de fase daarvoor. Men kan zich afvragen, waar het onderwijsprogramma is dat volgens Mayer perfect zou werken en wat de resultaten waren. Opnieuw wordt op basis van wat men denkt te weten, geconcludeerd hoe het schrijfonderwijs zou moeten zijn, zonder dat men dit daadwerkelijk getoetst heeft. Het punt hier is strikt genomen niet dat deze aanbeveling empirisch niet goed gefundeerd is. Waar het om gaat, is dat deze aanbevelingen docenten richten op de planfase van het schrijven, terwijl helemaal niet zeker is, dat dit in de praktijk daadwerkelijk goed uitwerkt.

Samenvattend, de voorgaande voorbeelden waren bedoeld om te illustreren dat onderzoek soms mogelijk een negatieve invloed heeft uitgeoefend op docenten en het oefenen met taal en schrijven in de klas. Er zijn vermoedelijk zonder lang zoeken nog veel meer voorbeelden te vinden. Het mogelijke probleem is veroorzaakt doordat onderzoekers uitspraken deden op basis van te weinig en te gebrekkig onderzoek. Uitspraken zijn te vaak niet 'evidence-based', maar projecties van ideeën die men heeft over hoe men denkt dat het is. Of zoals Jaynes het formuleerde: "we are all under an ego-driven temptation to project our private thoughts out onto the real world, by supposing that the creations of one's own imagination are real properties of Nature" (Jaynes, 1989, p. 7).

2.3 Succesvol schrijfonderwijs

In het voorgaande zijn factoren genoemd waarom schrijfonderwijs mogelijk faalt. Omgekeerd kan ook gekeken worden naar voorbeelden van succesvol schrijfonderwijs.

Sommige mensen ontwikkelen zich tot goede schrijvers, andere niet. Het is verleidelijk om aan te nemen dat mensen schrijven vooral leren op school. De onderwijsinspectie veronderstelt dat ook. "In tegenstelling tot andere taalvaardigheden leren kinderen schrijven vrijwel uitsluitend op school. Thuis leren kinderen spreken en luisteren, maar leren schrijven is doorgaans exclusief verbonden aan de school" (Inspectie van het onderwijs, 2010, p. 11). Of dat echter werkelijk zo is, is de vraag. Afkomst en het milieu blijken heel bepalend te zijn voor schoolse vaardigheden. Is het dan niet mogelijk dat dit ook opgaat voor het schrijven van teksten? Heeft de school op dit punt inderdaad invloed en hoeveel dan?

Voor het beantwoorden van deze vraag is idealiter een experiment nodig waarbij sommige kinderen naar school gaan en andere niet. Dat experiment is niet uitvoerbaar, omdat alle kinderen naar school moeten. Een minder ideaal antwoord kan gevonden worden door te kijken naar de verschillen tussen slechte en goede scholen. Op een goede school leert een kind meer en haalt ten slotte een hogere score dan op een slechte school. Het is dan duidelijk hoeveel een goede school meer oplevert dan een slechte school.

Een complicerende factor hierbij is dat kinderen nogal kunnen verschillen door aanleg en milieu. Een ruwe manier om daar rekening mee te houden is het percentage achterstandskinderen dat een school heeft. Basisscholen worden daarom door de onderwijsinspectie ingedeeld in 7 schoolgroepen op basis van het percentage achterstandskinderen dat ze hebben. Groep 1 zijn de scholen zonder achterstandskinderen met kinderen van hoogopgeleide ouders. De inspectie gaat ervan uit dat binnen een groep scholen de kinderen gemiddeld per school niet meer belangrijk verschillen.

Scholen binnen een schoolgroep blijken ongeveer 30 procentpunt van elkaar te kunnen verschillen in het percentage items dat hun leerlingen gemiddeld goed hebben op de Cito-eindtoets voor het schrijven van teksten (Inspectie van het Onderwijs, 2010, p. 32). Groep 7 met veel achterstandskinderen begint ruwweg bij 50% goed en eindigt bij 80% goed. Groep 1 begint ruwweg bij 65% en eindigt bij ongeveer 95%. Een goede of slechte school kan dan 30 procentpunt verschil maken: een leerling scoort bij school A 60% goed, maar als hij naar

school B was gegaan, had hij vermoedelijk 90% gehaald op de Cito-eindtoets voor het schrijven van teksten. Dat is een zeer groot verschil (men moet zich hierbij realiseren dat het 'nulpunt' bij de Cito-items niet 0% is; het gaat hier immers om meerkeuze items). Het zou kunnen dat sommige scholen door toeval in een bepaald jaar heel gunstig uitkomen. Er blijken echter veel scholen te zijn die systematisch over meerdere jaren hoog (21% van de scholen) of juist laag (25% van de scholen) uitkomen. Er bestaan dus schrijfsterke scholen en schrijfzwakke scholen.

De scholen met veel achterstandskinderen scoorden gemiddeld 67% goed, de scholen zonder achterstandskinderen scoorden gemiddeld 80% goed (Inspectie van Onderwijs, 2010, p. 32). Een verschil van 13 procentpunt wat ook behoorlijk is, maar aanzienlijk kleiner dan de invloed van de school binnen eenzelfde scholengroep, waarin sprake is van ongeveer 30 procentpunt verschil. Met andere woorden: het verschil binnen eenzelfde scholengroep is groter dan tussen scholengroepen. De kwaliteit van de school maakt belangrijk meer uit voor de schrijfvaardigheid van de leerling dan afkomst en milieu.

Deze conclusie wordt ook bevestigd door de resultaten van de Morningside Academy. Deze particuliere Amerikaanse basis- en middenschool die ook leraren opleidt, is volledig gebaseerd op 'evidence-based educational methods' en richt zich op kinderen met een leerachterstand en een leerhandicap. Per jaar garandeert de school een vooruitgang van minimaal twee normale schooljaren, anders ontvangen de ouders het schoolgeld retour. Sinds de oprichting in 1980 heeft men minder dan 1% van de schoolgelden retour moeten storten (Johnson & Street, 2004). Wanneer onderwijs geoptimaliseerd wordt op basis van de beschikbare kennis over wat werkt en wat niet werkt, kunnen achterstandsleerlingen niet alleen het normale programma goed doen, maar blijkt het, zelfs met dit type leerlingen, in de helft van de tijd te kunnen.

Op dezelfde manier als schrijfzwakke en schrijfsterke scholen kunnen worden onderscheiden, bestaan er ook taalsterke en taalzwakke scholen. In de praktijk overlappen beide begrippen elkaar sterk, omdat de cito-taalscores onderling hoog correleren. Op welke punten verschilden taalsterke scholen in hun onderwijs van taalzwakke? Een rapport van de onderwijsinspectie (Inspectie van het Onderwijs, 2009) vergeleek 137 willekeurig getrokken taalzwakke scholen met 142 willekeurig getrokken taalsterke scholen op 39 punten. Deze punten werden door de inspecteurs per school beoordeeld als 'voldoende' of 'onvoldoende'. Dit leverde een aantal significante verschillen tussen het onderwijs van taalsterke en taal-

zwakke scholen op. De bevestigde punten die door een meerderheid van de taalsterke scholen werden toegepast, zijn hierna vermeld.

Taalsterke scholen bleken de lesstof beter door te werken met de leerlingen. Taalzwakke scholen bleken de lesstof vaak in te korten of sommige leerlingen vrij te stellen. Taalsterke scholen planden voldoende tijd in voor het taalonderwijs. Meer leerlingen waren actief tijdens de les. Taalzwakke scholen planden vaak te weinig tijd in. Ook waren de leerlingen tijdens de les vaker passief. Op taalsterke scholen waren de docenten taakgerichter, legden de docenten duidelijker uit en gaven vaker uitleg over de te volgen strategie. Zij gaven ook effectiever feedback (Inspectie van het Onderwijs, 2009, p. 34-36). Bij zorgleerlingen pakten taalsterke scholen de zorg planmatiger aan en controleerden de effecten beter. Leerlingen met taalachterstand kregen vaker extra tijd (p. 39).

Afgaande op dit onderzoek bestaat effectief schrijfonderwijs dus uit iets simpels als de leerstof goed doorwerken, genoeg tijd uittrekken, zorgen dat de leerlingen werken. Docenten moeten taakgericht en duidelijk zijn en effectieve feedback geven. Verder moeten langzame leerlingen extra tijd krijgen of een speciaal programma waarvan het resultaat nagegaan moet worden.

Muralidharan en Sundararaman (2011) lieten experimenteel zien dat een kleine financiële extra beloning voor docenten in India die gekoppeld was aan objectieve leerlingprestaties, leidde tot betere prestaties van de leerlingen. De docenten bereikten dit verschil door meer huiswerk op te geven, in de klas extra opdrachten te geven, buiten de normale schooltijden les te geven, vaker toetsen af te nemen en door speciale aandacht aan achterblijvende kinderen te besteden (p. 68). Ook deze uitkomst duidt erop dat het gedrag van de docent in de klas van groot belang is.

2.4 Naar een aanpak van het schrijfprobleem

Hoe kan het schrijfprobleem van hbo-studenten, als dat inderdaad aantoonbaar is, aangepakt worden? Bij de onderstaande uitgangspunten is, behalve van de hiervoor behandelde literatuur, ook uitgegaan van de principes van het ABC-model (Antecedents-Behavior-Consequences-model) dat besproken wordt in 3.3.

Primair beoordelingsprobleem

Het schrijfprobleem lijkt in de eerste plaats een meetprobleem, een beoordelingsprobleem. Doordat het niveau waarop een student schrijft, moeilijk vaststelbaar is, wordt tekstkwaliteit en daarmee schrijfvaardigheid een vaag en moeilijk grijpbaar iets. Voor docenten en onderwijsinstellingen wordt het dan verleidelijk om problemen met de schrijfvaardigheid niet te zien of te negeren. Purves (1992) stelde: "The quality of school writing is what observers report they see" (p. 109). Op het moment dat een docent kan zien, dat het taalgebruik ernstig te wensen overlaat, is het eenvoudiger het probleem te negeren of te veronderstellen dat het elders opgelost zou moeten worden, dan op basis van het eigen subjectieve oordeel een lastig probleem onder de aandacht te brengen. Inspanningen om het probleem aan te pakken zouden zich daarom in de eerste plaats moeten richten op de vraag wat men precies verstaat onder een qua basale schrijfvaardigheid goed geschreven tekst.

Zodra men beschikt over een bevredigende operationalisatie van basale tekstkwaliteit, is een volgende voor de hand liggende stap om na te gaan hoe het zit met de schrijfvaardigheid van studenten. Als een duidelijke en werkbare operationalisatie van schrijfvaardigheid en tekstkwaliteit voorhanden is, zou dat meteen ook het bijbehorende schrijfonderwijs vastleggen. Het lijkt immers niet zinvol om te toetsen op het ene aspect en te trainen op een totaal ander aspect.

Focus op product in plaats van proces

Is het mogelijk om op basis van de literatuur iets op te merken over de vermoedelijke vorm die schrijfonderwijs zou moeten hebben?

Bonset & Hoogeveen (2007, p. 54) veronderstelden dat in het onderwijs meer gefocust moet worden op het planningsproces. In zekere zin gaan ze nog een stap verder, ongeveer alles is goed, zo lang er maar niet daadwerkelijk geschreven wordt. Graham & Perin (2007a) kwamen op basis van hun meta-analyse tot de conclusie dat het aanleren van een schrijfstrategie de grootste bijdrage levert aan beter schrijven en suggereerden verder dat eigenlijk alles werkt, zo lang men maar geen schrijfopdrachten geeft. Rijlaarsdam et al. (2008) veronderstelden dat schrijfopdrachten niet nodig zijn en dat men vooral schrijven leert door het communicatieproces te observeren. Mayer (2008) stelde dat studenten vooral beter en meer moeten plannen, voordat ze beginnen met het daadwerkelijke schrijven.

Al deze opvattingen gaan uit van bepaalde ideeën over schrijven en leren, terwijl de empirische onderbouwing weleens te wensen overlaat (zie 2.2). Het lijkt moeilijk voorstelbaar dat docenten zich door deze opvattingen en conclusies van onderzoekers nooit hebben laten beïnvloeden bij het frustrerende en tijdrovende schrijfonderwijs. Tegelijkertijd lijkt dat schrijfonderwijs afgaande op de vele berichten steeds minder effectief.

Een succesvolle aanpak van het schrijfprobleem zal vermoedelijk niet gebaseerd moeten zijn op ideologische bevlogenheid, maar op empirische evidentie. In plaats van te werken vanuit de eigen opvattingen en theorie, zou men moeten proberen te werken vanuit de empirie. Niet ideeën en opvattingen van docenten en onderzoekers zouden het uitgangspunt moeten zijn voor aanbevelingen en conclusies, maar concrete schrijfproducten van studenten. Een niet-geselecteerde steekproef van zulke schrijfproducten vormde de basis voor dit onderzoek.

Door het werken met een steekproef van zulke schrijfproducten wordt het schrijfprobleem tastbaar en concreet. Een kenmerk van veel van de voorgestelde schrijfbenaderingen is echter dat ze juist niet rechtstreeks resulteren in een concreet schrijfproduct, omdat ze procesgericht zijn. De processen die uiteindelijk leiden tot de definitieve tekst liggen echter niet vast, zijn niet tastbaar en vormen daarmee geen empirische basis. De definitieve tekst ligt wel vast, is wel tastbaar en vormt daarmee wel een empirische basis. Ook is die tekst daardoor een goed startpunt om een indruk te krijgen van de schrijfvaardigheid van de student en om de student feedback te geven, niet over zijn schrijfproces, maar over zijn tekst (want die doet er uiteindelijk toe). Purves (1992) stelde in dit verband: "School writing is a matter of products not processes" (p. 113). Steinberg (1980) merkte echter op: "Teaching of writing focuses too much on product, on the written paper that the student submits, and not enough on process, on how to write" (geciteerd in Mayer, 2008, p. 127). Die opmerking komt erop neer dat het eindproduct er slechts beperkt toe doet, omdat de manier waarop het tot stand kwam ook van belang is. De sollicitant die dat probeert uit te leggen aan de werkgever die hem niet uitnodigde voor een een gesprek vanwege die belabberde brief, zal vermoedelijk weinig baat hebben bij die overtuiging.

Mayer (2008, p. 124) omschreef planning als een procedure 'to establish a plan for producing text'. In de praktijk wordt echter meestal niet gevraagd een concreet plan te schrijven, maar gewoon een eindtekst. Het resultaat van de planningsfase is dan niet meer rechtstreeks waarneembaar, maar moet worden afgeleid uit de kwaliteit van de eindtekst.

Tekstkwaliteit via 'holistische' beoordeling is al een uitermate vaag begrip en het resultaat van de planfase is daarmee nog vager. Ook de tegenovergestelde benadering werkt in de praktijk niet. Stel dat men studenten een plan laat schrijven voor een nog te schrijven artikel. De beoordeling van een normale tekst is in de praktijk al uiterst moeilijk en onbetrouwbaar. Maar een plandocument voor een artikel is niet echt af, nog minder gangbaar en de criteria voor beoordeling worden daardoor nog willekeuriger en daarmee nog afhankelijker van de individuele beoordelaar.

Focus op reviseren in plaats van plannen

De constatering dat studenten niet of amper plannen bij schrijfproducten zal ongetwijfeld kloppen. Ook lijkt het duidelijk dat beter plannen kan leiden tot een beter schrijfproduct. Maar is het verstandig bij iemand die amper kan schrijven, te beginnen met uitvoerig plannen? Moet iemand die rijlessen neemt, eerst zelf een doel bepalen, dan de route vastleggen alvorens aan de feitelijke les te beginnen? Moet iemand die begint met ballet eerst leren een choreografie te schrijven of is het beter te beginnen met het oefenen van de basisbewegingen?

Mayer (2008, p. 142) presenteerde een ogenschijnlijk vreemd resultaat, gezien de strekking van zijn voorafgaande betoog. Uit onderzoek kwam namelijk consistent naar voren dat kinderen die beter waren met transcriberen (letters op papier zetten, een beter handschrift hadden) ook betere schrijvers waren, in de zin dat ze beter formuleerden. Het bleek zelfs zo te zijn, dat handschrift-training leidde tot duidelijk betere schrijfproducten bij jonge kinderen.

"Overall, these studies provide consistent support for the idea that writers are better able to use the translation process for constructing essays that express their ideas when they do not have to devote excessive attention to the mechanics of writing letters. Graham and Harris (2000) summarize the research base by noting that 'writing development is dependent on the mastery of transcription skills' . . . that is, essay writing depends on having automated handwriting skill" (Mayer, 2008, p. 142).

Hoewel bij Mayer de nadruk vooral lag op het plannen (het genereren en het ordenen), blijkt hier iets heel simpels als goed en vlot letters op papier te kunnen zetten, belangrijk voor de kwaliteit van de geproduceerde tekst. Het lijkt voor de hand te liggen, om dit door

te trekken naar werken met een toetsenbord. Iemand die vlot en blind kan typen, is vermoedelijk belangrijk in het voordeel bij het schrijven van een tekst, dan iemand die iedere letter moet opzoeken en steeds naar het toetsenbord moet kijken. Is het dan niet plausibel om te veronderstellen, dat wat geldt voor letters op papier zetten ook geldt voor het vinden van woorden? Iemand met een kleine woordenschat lijkt duidelijk in het nadeel ten opzichte van iemand met een grote woordenschat. Iemand die zich bij ieder woord moet afvragen, hoe het gespeld wordt, heeft een nadeel boven iemand die automatisch weet hoe het gespeld moet. Iemand die vertrouwd is met bepaalde standaardformuleringen, is in het voordeel boven iemand die deze formuleringen niet kent. Iemand die snel ziet waar de fout in een zin zit en die snel kan corrigeren, is in het voordeel boven iemand die de fout niet ziet of niet weet hoe deze te corrigeren.

Met andere woorden: Mayer noemde criteria voor geoefende schrijvers. Hij formuleerde een model van hoe mensen idealiter schrijven. In het ideale geval doen we het misschien volgens die fasen. Om schrijven te leren, is het echter misschien beter achteraan in het model te beginnen, bij het bewerken (de 'reviewfase') in plaats van bij de lastige en abstracte planfase. Alleen de stellige aanbevelingen aan docenten hielden vaak precies het tegenovergestelde in.

Hebben studenten inderdaad problemen met reviewen? Pianko (1979) rapporteerde dat eerstejaars 'college' studenten minder dan 9% van de tijd besteedden aan lezen en reviewen van wat ze geschreven hadden (geciteerd in Mayer, 2008, p. 143). Hayes en Flower (1986) concludeerden dat hoe kundiger de schrijver, hoe meer tijd hij besteedde aan revisie (geciteerd in Mayer, 2008, p. 145). Fitzgerald en Markman (1987) gaven kinderen via een directe instructie-aanpak dertien lessen van 45 minuten waarin ze leerden teksten te herzien. De beoordeelde kwaliteit van de teksten ging voor de experimentele groep van 1 (minimaal) naar 4 (maximaal), maar niet voor de controlegroep (geciteerd in Mayer, 2008, p. 148). Door beter te leren reviseren, gingen de schrijfproducten van minimaal naar maximaal. Mayer (2008) merkte in dit verband op: "Students need to see how revision can turn a poor paper into an excellent one" (p. 148).

Dat studenten een tekst kunnen schrijven, lijkt vast te staan. Het probleem is vooral dat de tekst vaak zo slecht geschreven is en zoveel fouten bevat, dat de lezer de tekst niet meer begrijpt en niet langer serieus neemt. Is het eventuele schrijfprobleem dan niet beter op te lossen door studenten te leren van een slechte tekst een goede tekst te maken? In de tijd dat er

nog geen tekstverwerkers waren, was dat een moeilijke optie, maar nu is het bewerken van een tekst goed mogelijk geworden.

In dit verband is er nog een ander argument. Schrijfvaardigheid blijkt op twee totaal verschillende manieren vastgesteld te kunnen worden: 'direct' via (holistische) beoordeling van de door de studenten geschreven teksten, maar ook 'indirect' via objectieve tests bestaande uit meerkeuzevragen (Godshalk et al., 1966; Cooper, 1984, Breland, 1983). Dat schrijfvaardigheid uitermate lastig meetbaar is, blijkt alleen voor de directe methode te gelden en merkwaardig genoeg niet voor de indirecte methode. In 4.1 wordt dit punt uitgebreider behandeld. Waar menselijke beoordelaars op een of andere manier vrijwel volledig vastlopen in een moeras van onbetrouwbaarheid, blijken specifieke taaltests opgebouwd uit meerkeuzevragen snel en betrouwbaar goed schrijvende studenten te kunnen onderscheiden van slecht schrijvende studenten. Objectieve tests vormen evenwel geen goed middel om te leren schrijven (neemt men aan) en zijn niet toepasbaar op teksten. Ze hebben dus belangrijke beperkingen. Het is echter mogelijk dat de specifieke vaardigheden waar de objectieve tests zich op richten (het opsporen en corrigeren van fouten) om het oordeel van beoordelaars te voorspellen, juist de doorslaggevende vaardigheden vormen die studenten ontberen, wanneer zij in de ogen van beoordelaars slechte teksten produceren (in deelstudie 5 wordt op dit punt uitgebreider ingegaan).

Focus op lezen in plaats van schrijven

Een derde principe lijkt op grond van Tabel 1.1 en het voorgaande ook plausibel. Het probleem is vermoedelijk niet het schrijven, maar het lezen. Het probleem is niet in de eerste plaats dat studenten niet kunnen schrijven, maar dat ze niet in staat zijn goed te lezen wat ze zelf (of anderen) geschreven hebben. Wie Tabel 1.1 bekijkt, ziet dat de betrokken studenten of zelf hun tekst volledig niet gelezen hebben of als ze hem wel lazen, niet gezien hebben dat er iets niet klopte. Zodra een student doorheeft dat er iets niet klopt in een zin, zal hij proberen een oplossing te verzinnen. Zo lang hij het probleem niet ziet, zal hij ook geen oplossing zoeken. Het leesprobleem komt dus vermoedelijk eerst en pas daarna het schrijfprobleem: hoe moet het dan wel? Deze veronderstelling verklaart ook waarom bepaalde multiple-choice taaltests zo uitermate effectief slechte schrijvers van goede schrijvers kunnen onderscheiden, terwijl de tests qua inhoud niet schrijfvaardigheid lijken te meten, maar eerder leesvaardigheid. Het probleem in eerste instantie is vermoedelijk niet het schrijven, maar het kritisch lezen van wat men geschreven heeft.

Het idee dat voor schrijven tekstbegrip noodzakelijk is, wordt ondersteund door Miller en McCardle (2010). Zij merkten op dat volgens het *Child Development & Behavior Branch report (NICHD, NIH, DHHS, 2009)* allochtone leerlingen die Engels leren er vaak wel in slagen Engels te leren lezen (verklankend lezen) en te leren spellen tot een niveau dat vergelijkbaar is met de andere leerlingen, maar dat het ze vaak niet lukt tekstbegrip en schrijven even goed te leren beheersen als autochtone leerlingen (Miller & McCardle, 2010, p. 125). Dit duidt erop dat een goed tekstbegrip noodzakelijk is om goed te kunnen schrijven. Dit lijkt ook plausibel. Wie slecht is in het begrijpen van teksten, zal ook problemen hebben met het begrijpen van de zelf geschreven tekst. Juist het scherp lezen van de eigen tekst vormt de basis van het eindeloos reviseren en bijstellen van de tekst. Het gegeven dat leerlingen uit achterstandsgroepen vaak lager scoren bij schrijftaken zou erop kunnen wijzen dat een culturele achtergrond waarin de leerling kritisch heeft leren omgaan met verbale informatie, tekstbegrip dus, de basis vormt bij het zelf schrijven.

In het 'behavioral' schrijfmodel van Gilbert (2004) ontbreekt de planfase volledig. De schrijffase bestaat in haar model uit een 'topic': het onderwerp dat de leraar opgeeft. Dit leidt tot schrijfgedrag bij de student en resulteert in de eerste 'draft'. De eerste draft wordt vervolgens gelezen door de student en fouten worden opgemerkt ('noted alarms'). De passages met fouten worden vervolgens herzien. Dit resulteert in de eindtekst ('edited draft'). In dit model vormt scherp lezen de basis van het herschrijven. De eigen ervaring met schrijven leert dat in de praktijk herschrijven een terugkerend proces is. Gilbert bevestigt dit ook: "Anyone who has ever written anything will tell you this: Writing, like any art, can always be improved" (p. 365).

Focus op fouten

Een vierde principe kan nu ook geformuleerd worden. Het gaat niet om wat er goed is in de eindtekst, maar om wat er nog beter zou kunnen. Het doel is een tekst die zo perfect mogelijk is. Dat betekent dat iedere afwijking van het ideaal in beginsel een afwijking te veel is. Het streven is niet een tekst die de boodschap wel communiceert als de lezer zich voldoende inspanning getroost, maar het doel van het schrijfonderwijs is te leren een zo perfect mogelijke tekst te schrijven. Aan iedere tekst valt uiteindelijk altijd wel iets te verbeteren. Een goede schrijver blijft streven naar perfectie. Iedere fout, iedere gemiste verbetering, is er daarom één te veel.

Teksten zijn sociale boodschappen. Op het moment dat een dergelijke boodschap afwijkt van het ideaal, komt de boodschap minder effectief over. Een sollicitant kan nog zo goed zijn, op het moment dat de sollicitatiebrief een fout bevat, kan dat net een fout te veel zijn. Een verstandige sollicitant zal dat risico liever niet lopen. Of positief geformuleerd: een verstandige sollicitant zal zijn sollicitatiebrief op ieder mogelijk verbeterpunt verbeteren alvorens hem te versturen.

Dat wil zeggen dat iedere afwijking van het ideaal een verbeterpunt vormt en dus 'fout' is. Dit principe lijkt haaks te staan op hoe we willen omgaan met studenten. Het is belangrijk hun activiteiten te stimuleren en te bekrachtigen. Dit laatste blijft onverminderd van belang, maar hoeft het ander niet altijd uit te sluiten. Voor de beoordeling van het schrijfproduct heeft dit principe belangrijke gevolgen. De beoordelaar hoeft niet langer aan te geven hoe goed hij de tekst vindt, waarbij hij zich snel laat leiden door inhoudelijke overwegingen. Het is voldoende om aan te geven wat hij qua taalgebruik niet goed vindt aan de tekst en wat volgens hem verbeterd zou kunnen worden. Strikt genomen moet iedere 'fout' die een beoordelaar signaleert, vertaald kunnen worden in een tekstwijziging waardoor de tekst beter wordt. Of een gesignaleerde 'fout' inderdaad leidt tot een verbetering van de tekst zou in beginsel empirisch uitgezocht kunnen worden. Voor de praktijk is dat natuurlijk te omslachtig en zal moeten worden afgegaan op het oordeel van ervaren schrijvers en lezers.

Het streven naar perfectie lijkt aan te sluiten bij het idee van mastery-learning. De keuze voor fouten als uitgangspunt bij de beoordeling van basale schrijfvaardigheid sluit ook aan bij het wonderlijke, empirische gegeven dat multiple-choice taaltests voor taalgebruik en zinscorrectie beter dan de menselijke beoordelaar onderscheid kunnen maken tussen goede en slechte schrijvers (Godshalk et al., 1966; Cooper, 1984, Breland, 1983).

Focus op oefenen in plaats van doceren

In de hiervoor besproken literatuur is laten oefenen met schrijven niet de meest populaire optie. Studenten vinden schrijven frustrerend en docenten ook. Onderzoekers zien het nut er niet van in. Toch is het de vraag of die weerstand tegen daadwerkelijk schrijven terecht is. Wel is die weerstand begrijpelijk. Schrijven kost veel tijd, de beoordeling van de schrijfproducten kost veel tijd en de beoordeling en het commentaar van de docent komen pas veel later. Om oefenen zinvol te maken, is duidelijke en snelle feedback wenselijk.

Als het meetprobleem opgelost is en duidelijk is hoe men schrijfvaardigheid denkt vast te stellen aan de hand van een tekst, kan op basis hiervan schrijfvaardigheid getoetst worden. De meest effectieve onderwijsvorm om in deze vaardigheid vervolgens beter te worden, is daadwerkelijk oefenen met schrijven, waarna het geschrevene beoordeeld en kort besproken wordt. Deze overtuiging is de basis van het in paragraaf 3.3 te bespreken leermodel. Wie autorijden wil leren, moet niet naast de bestuurder gaan zitten, maar zelf achter het stuur gaan zitten. Wie een groot pianist wil worden, kan niet volstaan met eindeloos naar pianomuziek te luisteren, maar zal ook zelf eindeloos veel moeten oefenen met pianospelen. Onderwijs in de vorm van doceren (uitleg van een docent) kan die eigen oefening nooit vervangen is het uitgangspunt van dat leermodel. Uiteindelijk is echter de vraag wel of niet oefenen een empirische kwestie. Op het moment dat tien uur uitleg meer resultaat oplevert dan tien uur oefenen, zou het onverstandig zijn van dit gegeven geen gebruik te maken.

Focus op kleine in plaats van grote schrijfopdrachten

Wie kiest voor een schrijfgerichte aanpak, zal zich moeten afvragen wat goede schrijfopdrachten zijn. Voor een belangrijk deel wordt dit vastgelegd door de manier waarop men basale schrijfvaardigheid precies operationaliseert. De omvang van de schrijftaak staat hier echter los van. Studenten moeten doorgaans werkstukken inleveren van meerdere pagina's, terwijl ze vaak moeite hebben met het formuleren van een enkele zin. Gilbert (2004) merkte op: "Mastery of any discipline requires fluency in the basic skills, yet we educators allow many students to advance before ensuring they have acquired this degree of expertise" (p. 362).

De basis van het in 3.3 uitgebreider te bespreken ABC-leermodel is dat leren activiteit vereist bij de student die gevolgd wordt door feedback. Een effectief oefenprogramma werkt met kleine opdrachten die weinig tijd van de student vergen om uit te voeren die vervolgens onmiddellijk gevolgd worden door feedback. De mate van interactiviteit moet bij voorkeur hoog zijn.

Met de 'interactiviteit' wordt het aantal responsen (de B in de ABC sequentie) per tijdseenheid bedoeld. Tegelijkertijd komt dit aantal ook overeen met het aantal keren dat feedback gegeven moet worden (de C in de ABC-sequentie) en met het totaal aantal ABC-sequenties per tijdseenheid. Voor een effectieve onderwijsopzet moet bij voorkeur gewerkt worden met korte taken waarop vervolgens snel en duidelijk feedback wordt gegeven.

Wanneer de student een literatuurlijst krijgt en (zonder andere feedback) drie maanden later de toetsuitslag, bestond het onderwijs uit één ABC-sequentie in drie maanden. Dit komt overeen met viermaal feedback per jaar. Wanneer dezelfde stof wordt ondergebracht in een individueel studiesysteem waar de studietaak is opgesplitst in weektaken met daarna een toets, wordt de feedbacklus een week. Dit komt overeen met ongeveer vijftigmaal feedback per jaar, wat belangrijk interactiever is. Wanneer dezelfde stof als geprogrammeerde instructie via bijvoorbeeld de computer wordt aangeboden, kan onder ideale omstandigheden enkele malen per minuut feedback gegeven worden. De tijd per feedbacklus daalt naar 10 à 15 seconden. Uitgaande van 1600 uur per jaar, komt dit overeen met meer dan 300 000 maal feedback per jaar. Indien dit realiseerbaar zou zijn: een toename met een factor van bijna 100 000.

In een flightsimulator en in een computergame is de tijd die verstrijkt tussen het aanbieden van de 'situatie' (opdracht), het reageren daarop door de 'student' waarna het programma reageert met een bijgestelde versie van de situatie, zo snel dat men zich normaal niet meer realiseert dat het in feite om een opeenvolging van statische beelden gaat. De responsetijd van het systeem ligt doorgaans belangrijk onder de 0.1 seconde. Het systeem is zeer interactief.

Focus op duidelijke in plaats van vage schrijfopdrachten

Een leerling die de opdracht krijgt een werkstuk van enige omvang in te leveren over een onderwerp naar keuze, wordt niet alleen geconfronteerd met een schrijfprobleem. Welk onderwerp moet hij kiezen? Wat is een geschikt onderwerp? Waar vindt hij informatie? Vervolgens moet de informatie gelezen en bestudeerd worden. Om basaal schrijven te leren zijn zulke opdrachten onnodig complex. Ook wanneer het onderwerp wordt voorgeschreven zit hij met het probleem van het vinden van informatie en die bestuderen. In plaats van een schrijfopdracht is de opdracht evenzeer een leesopdracht. Als men toch een leesopdracht wil combineren met een schrijfopdracht, is het veel helderder om de te lezen informatie mee te leveren. Een probleem dat zich dan voordoet, is dat de meegeleverde informatie wordt overgeschreven. Het is in de gegeven tekst al goed geformuleerd, zodat hij er weinig meer aan toe kan voegen. Om basaal schrijven te leren is dit mogelijk helemaal niet een slechte strategie. Zeker wanneer de leerling daarna de tekst opnieuw schrijft, zonder dat het voorbeeld aanwezig is. De leerling maakt zich daardoor de oorspronkelijke tekst op een productieve manier eigen en oefent daardoor veel basisvaardigheden effectief in. Men kan

dit vergelijken met ballet. De juf laat zien wat de bedoeling is. De leerlingen zijn vervolgens eindeloos bezig hetzelfde resultaat te bereiken. Pas nadat de leerling 'fluency' heeft bereikt in nadansen, is de basis aanwezig om grotere stukken te dansen.

Voor het inoefenen van basale schrijfvaardigheid lijkt het daarom beter de benodigde informatie bij de schrijfopdracht mee te leveren. Zo'n schrijfopdracht kan de vorm hebben van een stukje informatie met de opdracht die om te zetten naar een goed stukje tekst. Schrijven op basaal niveau is niet speciaal informatie genereren, maar vooral informatie bewerken.

2.5 Samenvatting

In dit hoofdstuk is ingegaan op het probleem van het falende schrijfonderwijs. Schrijfonderwijs is moeilijk te evalueren op resultaat, doordat er weinig overeenstemming is over wat een goede tekst is. Door dit beoordelingsprobleem is de vooruitgang van leerlingen en het tekortschietende schrijfniveau niet duidelijk. Een tweede probleem is dat schrijven geleerd moet worden door daadwerkelijk te schrijven, maar schrijfopdrachten kosten de docent buitensporig veel nakijktijd, terwijl de feedback te laat komt om veel effect te hebben. Docenten veronderstellen ook dat hun leerlingen het al geleerd hebben of wanneer het duidelijk is dat ze dit nog niet kunnen, dat het niet hun verantwoordelijkheid is. Leren schrijven heeft in de praktijk een lage prioriteit en schrijfvaardigheid is geen onderdeel van centraal afgenomen examens.

Mogelijk heeft ook het nieuwe schrijfonderwijs, waarin het accent wordt gelegd op het schrijfproces in plaats van op de geproduceerde tekst, niet goed uitgewerkt. De doelstelling kan hierdoor verschoven zijn van het produceren van een goed geschreven tekst, naar kunnen praten over het schrijfproces. Verder wordt hierdoor veel tijd besteed aan het plannen van de schrijftaak, waardoor er minder daadwerkelijk geschreven wordt en nog minder geoefend wordt met het herzien van de tekst. Onderzoek heeft waarschijnlijk vaker een negatieve rol gespeeld bij het tekortschieten van het schrijfonderwijs. Zo werd in de meta-analyse van Graham en Perin (2007a) iedere vorm van schrijfonderwijs voorgesteld als effectief, behalve daadwerkelijk laten schrijven, zonder dat men dit kon onderbouwen met deugdelijk onderzoek.

Schrijfsterke scholen zijn scholen waarvan de leerlingen voortdurend (veel) beter dan gemiddeld scoren op de CITO-schrijfitems. Een schrijfsterke school kan ten opzichte van een schrijfzwakke school resulteren in 30 procentpunt meer goed beantwoorde schrijfitems. Dit verschil is veel groter dan het verschil ten gevolge van ouderlijk milieu. De school heeft daarmee een zeer grote invloed op het schrijfniveau van de leerlingen.

Op een soortgelijke manier als schrijfsterke scholen blijken er ook taalsterke scholen te zijn. Wat doen taalsterke scholen anders dan taalzwakke? Taalsterke scholen bleken de lesstof beter door te werken met de leerlingen, terwijl taalzwakke scholen de lesstof vaak bleken in te korten of sommige leerlingen vrijstelden. Taalsterke scholen planden voldoende tijd in voor het taalonderwijs. Meer leerlingen waren actief tijdens de les. Taalzwakke scholen planden vaak te weinig tijd in. Ook waren de leerlingen tijdens de les vaker passief. Bij zorgleerlingen pakten taalsterke scholen de zorg planmatiger aan en controleerden de effecten beter. Leerlingen met taalachterstand kregen vaker extra tijd.

Hoe zou effectief schrijfonderwijs eruit moeten zien? Allereerst zou er een duidelijke en werkbare definitie moeten komen van wat verstaan moet worden onder tekstkwaliteit. Een tweede punt is dat de nadruk zou moeten liggen op het product en niet op het proces. Ten derde zou de aandacht niet vooral moeten uitgaan naar het abstracte planproces, maar naar het concrete reviseren. Slecht schrijven is niet alleen een schrijfprobleem, maar vooral ook een leesprobleem. Slechte schrijvers zien niet wat er precies fout zit in hun zin. In plaats van de aandacht te richten op de betekenis, zou er vooral gefocust moeten worden op fouten in de tekst. Verder zou schrijfonderwijs zich vooral moeten richten op daadwerkelijk oefenen met schrijven. Dat kan het beste gebeuren via veel kleine opdrachten en onmiddellijke en duidelijke feedback.

3

Onderwijs: evaluatie, constructie en methode

In paragraaf 2.1 werden een aantal oorzaken geopperd voor het mogelijk tekortschietende schrijfonderwijs. Een algemener punt in dit verband is het probleem van de (summatieve) onderwijsevaluatie. Doordat onderwijs niet of niet op de juiste criteria geëvalueerd wordt, is onderwijs vaak niet optimaal. Bij evaluatie van onderwijs wordt doorgaans naar kenmerken van het gegeven onderwijs gekeken in plaats van naar de gerealiseerde leerwinst. Onderwijs zou echter niet beoordeeld moeten worden op de vorm of inhoud, maar op de daadwerkelijk gerealiseerde leerwinst. De grondslag voor deze vorm van onderwijsevaluatie is empirisch: het verschil tussen begin- en eindmeting. In 3.1 wordt hier uitgebreider op ingegaan. Deze paragraaf vormde de basis voor het onderzoek naar de effectiviteit van het nieuw geconstrueerde programma.

Bij de constructie van onderwijs wordt in de praktijk vaak eerst de leerstof gekozen en het onderwijsprogramma gemaakt en pas als laatste de toets geconstrueerd. Deze wordt vervolgens zo samengesteld dat de behandelde stof zo goed mogelijk afgevraagd wordt. Het resultaat van deze werkwijze is dat het middel (het onderwijsprogramma) het einddoel (een voldoende resultaat op de toets) bepaalt in plaats van omgekeerd. Een betere werkwijze lijkt te zijn: eerst nagaan wat men precies wil bereiken (de onderwijsdoelstelling formuleren) en hoe men dat precies gaat vaststellen (de toets volledig vastleggen) en pas daarna het onderwijsprogramma zo kiezen en construeren dat studenten optimaal voorbereid worden op de toetsing.

De focus op leerstof is overigens vanuit de docent gezien, wel begrijpelijk. Zijn taak is onderwijs te verzorgen en in de praktijk zal hij ook op het gegeven onderwijs worden beoordeeld. Of studenten na afloop wel of niet een bepaalde vaardigheid beheersen, is normaal niet waarop de docent beoordeeld wordt. Men kan de situatie vergelijken met een boek. Uitgevers zullen zich inspannen om het boek er optimaal uit te laten zien. Of het boek in de praktijk vervolgens inderdaad bruikbaar is om bijvoorbeeld te leren programmeren, onttrekt zich aan de waarneming. Misschien zal een enkele lezer later ooit verzuchten dat het boek er mooi uitzag, maar niet erg hielp. Dat is echter geruime tijd na de aankoop.

Dit onderwijsconstructieprobleem komt aan de orde in 3.2. Bij de constructie van het nieuwe programma is getracht dit probleem te vermijden.

Een derde probleem dat zich bij onderwijs voordoet, is de vraag naar de meest optimale onderwijsmethode. Wanneer men duidelijkheid heeft over de toetsing, moet daarna nog de vraag beantwoord worden wat vermoedelijk de beste methode is om de student op de toets voor te bereiden. Uitgangspunt voor het antwoord op deze vraag is de gekozen onderwijskundige theorie (het leermodel). Op basis van het gekozen leermodel is het ook mogelijk bestaande programma's te beoordelen op hun verwachte effectiviteit. Deze beoordeling kan nooit meer zijn dan een voorspelling van de verwachte effectiviteit. Pas een empirische evaluatie van de leerwinst van een programma kan uitwijzen of de verwachte effectiviteit ook daadwerkelijk gerealiseerd wordt.

De bestaande taalmethodes in deelstudie 2 zijn voor zover bekend is nooit empirisch geëvalueerd en moesten daarom voor hun te verwachten effectiviteit beoordeeld worden op basis van een onderwijskundig model. Voor het leermodel is uitgegaan van het ABC-model dat was uitgewerkt in een beoordelingsschema. Dit model vormde ook het uitgangspunt voor de constructie van het nieuwe taalprogramma TAVAN. De achtergronden van het ABC-model worden besproken in 3.3.

3.1 Onderwijsevaluatie-probleem

Wanneer is onderwijs goed?

Hudson (2001) constateerde toen kinderen niets bleken te leren van grammatica-onderwijs: "but this is hardly surprising - the same is surely true of any subject" (p. 3). Hij ging er vanuit dat in het meeste onderwijs niets geleerd werd. Burt stelde op WISE 2011 (World Innovation Summit for Education 2011): "The problem is not access to education, but quality of education." Lambay beweerde op dezelfde conferentie: "The issue is about quality." Niet de toegang tot onderwijs is het probleem, maar de kwaliteit van onderwijs.

Wanneer bij onderzoek naar de economische effecten van onderwijs wordt uitgegaan van toetsscores in plaats van aantal jaren onderwijs, zijn sterke verbanden gevonden met economische groei. Minne, Van der Steeg en Webbink (2007) rapporteerden: "Nog recenter is de

replicatie van Hanushek en Woessman (2007) met nog meer gegevens. Zij vinden dat een toename van de gemiddelde toetsscore van een land met één standaarddeviatie samenhangt met 2 procentpunt meer groei [per jaar] over een periode van 40 jaar" (Minne et al., 2007, p. 12). Over een periode van 40 jaar resulteert dat in meer dan een verdubbeling van de productie. De auteurs stelden: "De gedachte is dat vanwege verschillen in onderwijskwaliteit een jaar onderwijs moeilijk vergelijkbaar is tussen landen" (p. 12).

Kwaliteit van onderwijs is een wat vaag begrip, omdat mensen verschillende opvattingen hanteren over wat onderwijs is of zou moeten zijn en over de vraag op welke criteria het precies beoordeeld moet worden. Als onderwijs niet of niet op de juiste criteria beoordeeld wordt of kan worden, is er echter geen reden en geen mogelijkheid om de kwaliteit van het onderwijs systematisch te verbeteren. Het onderwijsevaluatie-probleem heeft betrekking op de vraag hoe de kwaliteit van onderwijs vastgesteld en gekwantificeerd moet worden (Van Es, 1980). Wanneer is onderwijs 'goed'? Het volgende voorbeeld, ontleend aan Bons (2011), is bedoeld om het probleem te concretiseren.

"Onlangs verbleef ik met tien anderen drie dagen in een kasteel in Zeist. Na een ochtend theorie over persoonlijke ontwikkeling op de werkvloer kregen we de opdracht al onze eventuele teleurstellingen en frustraties over onze banen op te schrijven. Nadat iedereen klaar was, gingen we met z'n allen naar buiten. De begeleider nam de formulieren in en zei dat we rond een ton moesten gaan staan. Hij pakte een aansteker, stak de fik in de formulieren, gooide de papieren in de ton en zei: 'Dát doen we met alle frustraties, wég er mee.' De cursus kostte 2.495 euro. De cursusleider zal zich ongetwijfeld goed hebben voorbereid, veel hebben gelezen en hard hebben gewerkt. Maar wat is de zin hiervan?"

In dit voorbeeld velde de schrijver impliciet een negatief oordeel over het gevolgde onderwijsprogramma. Het onderwijs kwam niet overeen met zijn opvattingen over hoe een onderwijsprogramma eruit behoort te zien en het was volgens hem daarmee geen goed onderwijs. Een harde basis voor een dergelijke conclusie ontbreekt echter. Mogelijk ziet een ander dit onderwijs als een eyeopener. Op dezelfde wijze zullen ook docenten, instellingen en politici hun ideeën hebben over hoe goed onderwijs eruit behoort te zien en kunnen de waardeoordelen over specifiek onderwijs volledig uiteen lopen. Uitgaan van dit soort opvattingen levert geen empirische basis op voor evaluatie, maar een uitgangspunt dat per persoon, per cultuur en per periode kan verschillen.

Wat bedoelen we met 'onderwijs'?

Om de evaluatie van onderwijs een empirische basis te geven, is het belangrijk het eerst eens te worden over wat onderwijs precies is of moet zijn.

Een opmerking over de terminologie: de term 'onderwijs' kan overkoepelend gebruikt worden om het geheel van onderwijsdoel, begingedrag, onderwijsprogramma en -methode, eindgedrag (gerealiseerd leerresultaat) en evaluatie gezamenlijk aan te duiden. Daarom zijn de termen 'onderwijsprogramma' of 'onderwijsmethode' bij mogelijke onduidelijkheid gebruikt voor dat deel van het onderwijs, dat studenten moet helpen bij het bereiken van een bepaald doel.

Richards (2010) onderscheidde vier verschillende soorten opvattingen over het geven van taalonderwijs: wetenschapsresearch opvattingen, theoretisch-filosofische opvattingen, waardegebaseerde opvattingen en kunst-vakmanschap opvattingen. Wetenschapsresearch opvattingen veronderstellen dat in onderzoek naar leren en geheugen ontdekte leerprincipes belangrijk zijn. De docent ontwikkelt taken en activiteiten uitgaande van deze principes. Hij 'monitort' de prestaties van zijn studenten op taken om te zien of het gewenste doel bereikt is. Theoretisch-filosofische opvattingen gaan uit van een bepaalde theorie en principes. Deze theorie ziet de docent als belangrijk. Vervolgens selecteert men materiaal en taken uitgaande van deze theorie. De docent 'monitort' vervolgens zijn onderwijs om te kijken of het nog voldoet aan zijn theoretische principes. Waardegebaseerde opvattingen gaan uit van bepaalde waarden. Deze waarden ziet men als belangrijk. Alleen onderwijsmiddelen en taken die voldoen aan deze waarden zijn toegestaan. De evaluatie bestaat er uit dat men nagaat of het onderwijs nog voldoet aan de waarden. Kunst-vakmanschap opvattingen zien iedere onderwijssituatie als uniek. De docent moet de bijzondere kenmerken van de situatie identificeren. Vervolgens probeert de docent verschillende onderwijsstrategieën uit. Ten slotte gaat het erom dat de docent een persoonlijke onderwijsbenadering ontwikkelt.

In de wetenschapsresearch opvattingen is er een kennis- of vaardigheidsdoel dat de docent uiteindelijk bij de student wil bereiken. In de theoretisch-filosofische en de waardegebaseerde opvattingen gaat het erom dat het onderwijs overeenkomstig de theorie, filosofie of waarde gegeven wordt. In de kunst-vakmanschap opvattingen gaat het om het ontwikkelen van een eigen stijl van onderwijs.

Van de vier soorten opvattingen die Richards onderscheidde, hebben twee soorten betrekking op het onderwijsprogramma. Daarbij gaat het er niet om wat het onderwijs bereikt bij de student, maar of het onderwijsprogramma overeenkomt met vooraf aangenomen normen. Eén soort opvatting is docentgericht. Het gaat er in die opvatting niet om of de student iets opsteekt of dat het onderwijs aan bepaalde eisen voldoet, maar dat de docent zich ontwikkelt. Slechts één van de vier soorten opvattingen is gericht op het bereiken van een doel met betrekking tot het gedrag van de student.

Deze vier verschillende opvattingen laten zien dat het begrip 'onderwijs' heel verschillend kan worden opgevat. Eén opvatting legde het accent op de student: die moet profiteren. Eén opvatting legde openlijk het accent op de docent: die moet zich ontwikkelen en kunnen doen wat hij wil. Twee gingen uit van a priori opvattingen en stelden vanuit die opvattingen eisen aan het onderwijsprogramma. Zahorik (1986) merkte over deze laatste opvattingen op: "Their truth is not based on a posteriori conditions or on what works. Rather, their truth is based on what ought to work or what is morally right" (geciteerd in Richards, 2010, p. 21).

Volgens het Van Dale online-woordenboek is 'onderwijs': 'Het systematisch overbrengen van kennis en vaardigheden door bevoegde leraren'. In deze 'klassieke' opvatting van onderwijs is de docent actief en komt de student niet voor. Er wordt niet getwijfeld aan wat de docent doet of probeert te doen, want in dat geval zou men het hebben over: systematisch *proberen* over te brengen van kennis en vaardigheden. Tenslotte moet de leraar bevoegd zijn. Deze opvatting leidt tot maximaal vijf onderwijscomponenten: de leraar, zijn bevoegdheid, het aantal uren onderwijs, de stof die behandeld wordt en de wijze waarop de stof behandeld wordt.

Evaluatie op proces of op resultaat?

Het probleem met deze 'klassieke' benadering is dat men wel eisen stelt aan het onderwijsprogramma dan wel de docent, maar ten slotte nog steeds niet zeker weet of de kennis en vaardigheden die de docent probeerde over te brengen op de student, ook inderdaad overgekomen zijn bij de student. Bij onderwijsevaluatie vanuit deze opvatting horen criteria als de bevoegdheid van de docent (wel of niet bevoegd), het aantal uren onderwijs dat gegeven of gevolgd werd (de aanname is vaak, dat meer uren leidt tot meer overgebrachte kennis) of

alle leerstof wel behandeld is (als het onderwerp behandeld is, ligt het probleem niet bij het onderwijs, maar bij de student) en de wijze waarop de stof behandeld is (duidelijk uitgelegd, levendig, goed te verstaan, inspirerend). In veel gevallen wordt op scholen bij evalueren ook het ordecriterium gebruikt: kon de docent wel orde houden? De gedachte daarachter is dat orde in de klas noodzakelijk is voor de leerlingen om te kunnen leren.

Hogescholen en universiteiten gaan in hun evaluaties vaak nog een stap verder. Het gaat er niet alleen om of de stof duidelijk is uitgelegd, maar de student moet ook tevreden zijn over het onderwijs en voor het vak gemotiveerd zijn door de docent. De aanname daarbij is mogelijk dat ontevreden studenten slechte reclame zijn en gemakkelijk elders kunnen gaan studeren en dat gemotiveerde studenten beter leren. Of de studenten daadwerkelijk iets geleerd hebben, is dan nog steeds de vraag.

Ook in het basisonderwijs is procesevaluatie vaak de norm. Zo vermeldde Scheerens (1997): "Gezien het sterke accent op procesevaluatie in de werkwijze van de Inspectie van het Onderwijs, vormt een output-georiënteerd (zelf)evaluatie-instrumentarium, geënt op eindtermen en tussendoelen een essentiële aanvulling. Bij een centralistische procesevaluatie is het gevaar van een bureaucratisering en 'goal displacement' levensgroot aanwezig" (p. 51).

Popham (1983) had het in dit verband over evaluatie op basis van intrinsieke en extrinsieke criteria. Hij gaf als voorbeeld dat men een boormachine kan beoordelen op vormgeving en kleur (intrinsieke criteria). "Wie wil een lelijke machine?", merkte hij op (p. 6). Men kan echter ook kijken naar de snelheid en zuiverheid waarmee boormachines gaten boren. Dit zijn extrinsieke criteria, omdat ze niet betrekking op de boormachine zelf, maar op het effect dat de boormachine heeft. Popham stelde: "Een beoordelende aanpak van onderwijsevaluatie met de nadruk op intrinsieke criteria is in het onderwijs gemeengoed" (p. 6).

Driscoll (2000) definieerde onderwijs als volgt: "By instruction I mean any deliberate arrangement of events to facilitate a learner's acquisition of some goal" (p. 25). In vergelijking met de vorige opvatting is de docent verdwenen en de actieve student ervoor in de plaats gekomen. Daarmee is ook de bevoegdheid van de docent niet langer een punt van belang. De stelligheid dat kennis en vaardigheden automatisch overkomen, is eveneens verdwenen. Het onderwijs is alleen nog een welbewuste poging de student te ondersteunen bij het bereiken van een doel.

In het onderwijsmodel van Van Gelder (Van Gelder, Oudkerk Pool, Peters & Sixma, 1973; Joling, 2001) is onderwijs ook een doelgerichte activiteit. In dit model bestaat onderwijs uit vier hoofdbestanddelen: het onderwijsdoel, de beginsituatie (het begingedrag), de onderwijsleersituatie (het onderwijsprogramma) en het onderwijsresultaat (het eindgedrag). De bedoeling van het onderwijsprogramma is dat de student aan het einde van het programma met zijn eindgedrag het doel bereikt heeft. Het onderwijsdoel moet gesteld zijn in termen van het eindgedrag van de student en niet in termen van wat de docent moet behandelen.

Het verschil tussen de opvatting van Van Gelder en de 'klassieke' opvatting is duidelijk. Bij Van Gelder gaat het erom dat de student na afloop bepaalde kennis of een bepaalde vaardigheid heeft. In de 'klassieke' opvatting gaat het erom dat de docent na afloop alle stof behandeld heeft. De focus is verschoven van het gedrag van de docent naar het gedrag van de student.

Onderwijsevaluatie op resultaat

Wanneer we uitgaan van de opvattingen van Driscoll en Van Gelder en onderwijs beschouwen als een middel om de student kennis en vaardigheden bij te brengen, is de consequentie dat onderwijs niet valide beoordeeld kan worden via aspecten van het onderwijsprogramma. Het onderwijsprogramma is immers alleen middel en geen doel. Het bijbelse 'aan de vruchten kent men de boom' behoort dus ook voor onderwijs te gelden.

Ook volgens De Groot (1983a) moest onderwijskwaliteit blijken uit de onderwijsresultaten. "Bij onderwijs gaat het uiteindelijk niet om de vraag hoe mooi we het geven maar om wat het uiteindelijk uithaalt, om wat leerlingen ervan meenemen. . . . Het gaat uiteindelijk om de afhankelijke variabelen, om de resultaten, om de leereffecten" (p. 60).

Hoewel het idee dat onderwijs geëvalueerd moet worden op de leerresultaten plausibel mag lijken, onderschrijft niet iedereen die stelling. Karstanje (1983) stelde in een reactie op De Groot dat er geen overeenstemming bestond over de doelstellingen van onderwijs. Het 'harde' doeldenken zou een fictie zijn en kwaliteitsbepaling was onmogelijk. De doelstelling van onderwijs moest niet alleen op het gebied van leereffecten liggen. De maatschappelijke functie omvatte in zijn ogen meer, wat volgens hem duidelijk werd zodra men het onderwijs wilde veranderen. Daarbij zou het er niet om gaan dat de leerlingen beter onderwijs

krijgen. "Het gaat . . . om rechtvaardiger onderwijs, om een bredere ontwikkeling van de leerlingen. Het gaat erom uit de leerling te halen wat erin zit. Dat kan men niet evalueren via toetsen" (p. 75).

Onderwijsresultaat meten

Wanneer we wel uitgaan van de veronderstelling dat onderwijskwaliteit moet blijken uit de leerresultaten, ligt het voor de hand om naar de toetsresultaten te kijken. Het percentage geslaagden op de toets zegt echter in de praktijk weinig, omdat de docent meestal zelf de toets construeerde en aan de hand van de resultaten de norm koos.

Om te kunnen evalueren op leerresultaat is daarom een vastliggend doel waarvan het bereiken van dat doel onafhankelijk van de docent wordt vastgesteld, een belangrijke eis. Een voorbeeld van een onafhankelijke toetsprocedure is het theoretische en praktische deel van het rij-examen. De rijinstructeur kan hier geen invloed op uitoefenen anders dan door de kandidaat zo goed mogelijk op te leiden.

Een tweede punt om te kunnen evalueren op leerresultaat, is dat het werken met een doelvariabele de voorkeur verdient boven het werken met een doel dat al dan niet gehaald wordt. Veronderstel dat het doel is dat een student een korte brief van een A4 kan schrijven zonder fouten. De student krijgen vijf lessen om dit te leren. Na afloop kan hij nog steeds niet een foutloos briefje schrijven. De doelstelling is niet bereikt en je zou dan kunnen concluderen dat het onderwijs niet goed was. Deze conclusie is echter wat voorbarig, omdat het heel goed kan zijn dat de student van bijvoorbeeld 20 naar 10 fouten per A4 is gegaan. In dat geval had het onderwijs wel een duidelijk effect, maar leverde dit effect in combinatie met het beginniveau niet het gewenste eindresultaat. Een duidelijk doel is weliswaar nodig om het onderwijs te sturen, maar een variabele die de afstand tot het doel meet, levert meer informatie over de vooruitgang van de student dan een dichotome maat als wel/niet bereikt.

Een derde punt is dat de afstand tot het doel niets zegt over de vooruitgang, de leerwinst, van de student. Stel dat de student nog dertig fouten per A4 van het doel verwijderd is. Kan hieruit geconcludeerd worden dat die vijf lessen slecht onderwijs vormden? Die conclusie is niet zinvol, omdat het beginniveau van de student niet bekend is. Het is dus mogelijk dat de student juist aanzienlijk vooruit is gegaan. Om de leerwinst (het effect van het onder-

wijsprogramma) te kunnen meten, is een voormeting noodzakelijk. Zonder voormeting is het wel mogelijk op basis van de eindmeting een indruk te krijgen van het eindniveau van de studenten. Maar de bijdrage van het onderwijsprogramma aan dat eindniveau, het effect van het onderwijsprogramma, valt daaruit normaal niet af te leiden.

Een voorbeeld om dit te illustreren. Bosbrandweer moet 10 basisregels voor de veiligheid kennen en kunnen toepassen. Er wordt een cursus georganiseerd en na afloop blijkt iedere brandweerman de 10 regels te weten en te kunnen toepassen. Aan de veiligheidsvoorschriften op dit punt is dus voldaan. Op basis van alleen deze informatie valt echter geen hard oordeel te vellen over de waarde van de cursus. Het kan zijn dat iedere brandweerman de principes al kende en toepaste. In dat geval heeft de cursus niets bijgedragen, hoewel het dan nog steeds mogelijk is dat de cursus met andere groepen wel effectief kan zijn. Aan de andere kant kan ook niet geconcludeerd worden dat de cursus heel goed werkte. Het is immers mogelijk dat de cursus bij mensen met een lage score helemaal geen verbetering oplevert. Het eindniveau van de cursisten zegt niet automatisch iets over de kwaliteit van het gegeven onderwijs.

Het omgekeerde gaat echter wel op. Stel dat na afloop van de cursus de cursisten laag scoren. In dat geval is het zeker dat het desbetreffende onderwijs voor hen nog niet goed genoeg gewerkt heeft. Zo bestempelde een visitatiecommissie van een steekproef van 23 afstudeerscripties van de opleiding Journalistiek van Hogeschool Windesheim er 11 als onvoldoende (De Stentor, 2011). Wanneer studenten veel fouten in een tekst maken, kan geconcludeerd worden dat het onderwijs verbetering behoeft, terwijl wanneer studenten goed scoren niet automatisch kan worden aangenomen dat dit komt door het onderwijs.

In de praktijk is het niet de gewoonte om voormetingen af te nemen. Peters (2010) vermeldde bij de verschillende toepassingsmogelijkheden voor taaltests in het hoger onderwijs niet het gebruik als voormeting voor onderwijsevaluatie. Ook Deygers en Kanobana (2010) dachten in hun overzichtsartikel niet aan de mogelijkheid om een taaltoets te gebruiken als voormeting.

De Groot (1983a) ging ervan uit dat onderwijsevaluatie mogelijk was op grond van vergelijking. "Kwaliteitsuitspraken zijn altijd vergelijkend; bijvoorbeeld: bij leraar A leer je meer, of betere dingen, dan bij leraar B" (p. 60). Met de mogelijkheid van een voor- en nameting hield hij geen rekening. In een overzicht van de mogelijkheden om studietoetsen te gebrui-

ken voor onderwijsevaluatie noemde hij wel meerdere vormen van vergelijkend onderzoek, maar niet de mogelijkheid een beginmeting uit te voeren. Ook de mogelijkheid een enkele cursus qua effect te evalueren werd niet vermeld (De Groot, 1983b).

In het meer op de toegepaste gedragsanalyse gebaseerde studieboek van Holland, Solomon, Doran en Frezza (1976) over onderwijsconstructie werd daarentegen een concreet voorbeeld vermeld daterend uit 1962 van een evaluatie met voor- en nameting. Een chemieprogramma werd getest met 12 leerlingen. De eindtest werd ook als pretest gebruikt. De resultaten werden gepresenteerd in een tabel per leerling, niet als gemiddelde. Bij de voormeting was de hoogste score 15% goed, bij de nameting was de laagste score 87% goed. De meeste scores waren 100% goed (p. 238).

Scheerens (1997) merkte op dat onderwijskwaliteit moet blijken uit de opbrengsten van het onderwijs en dat (bij correlationeel onderzoek naar schooleffectiviteit) de leerprestaties van de leerlingen gecorrigeerd moeten worden voor het aanvangsniveau. Hij noemde de evaluatie- en feedbackfunctie het 'kernmechanisme voor effectiviteitsverbetering'. "Een nauwgezet volgen van de vorderingen van leerlingen door de hele schoolloopbaan vormt hierbij de basis" (p. 49).

Bezwaren tegen evaluatie op resultaat

In de onderwijspraktijk bestaan echter vaak bezwaren tegen toetsen. Boes (2002) stelde in een beschouwing over de zin en onzin van toetsen, dat op scholen vaak onnodig wordt getoetst, wat ten koste gaat van de onderwijstijd. De resultaten van toetsen zouden worden aangewend voor oneigenlijke doelen, zoals het vergelijken van de kwaliteit van scholen. Bij veel toetsen wordt volgens Boes het oorspronkelijke doel uit het oog verloren, namelijk nagaan of het beoogde doel bereikt is. Bovendien wordt uitgegaan van het gemiddelde, zonder rekening te houden met de verschillen tussen kinderen.

Een belangrijk argument dat Boes niet vermeldt, maar dat vermoedelijk wel een grote rol speelt, is dat er in de praktijk vaak geen duidelijk alternatief is. Wanneer een leerling onvoldoende of te laag scoort op een toets, is er in de praktijk doorgaans geen duidelijk en effectief programma om te zorgen dat hij de volgende keer belangrijk beter scoort (dit punt wordt in deelstudie 5 uitgebreider behandeld). Het gevolg is dat men wel toetst, maar in feite niets met de uitslag kan.

Een praktische reden om als docent geen belang te hebben bij een gestandaardiseerde toets is dat eventuele slechte resultaten op die toets gemakkelijk toegeschreven kunnen worden op de docent, ook al treft hem mogelijk geen enkele blaam. Tegenvallende resultaten kunnen gemakkelijk leiden tot een situatie waarin de docent maar moet aantonen dat het niet door zijn lesgeven komt. Iets dat in de praktijk vrijwel niet valt aan te tonen.

Gegeven deze problemen zullen onderwijsgevenden het idee dat standaard een voormeting verricht moet worden, vaak niet enthousiast ontvangen. Verder brengt een voormeting extra werk mee. Er is een extra toets nodig die volledig vergelijkbaar moet zijn met de eindtoets (dit kan gerealiseerd worden door uit een itempool aselect vragen te trekken). Vervolgens moet die toets ook nog worden afgenomen. Het werken met een voormeting vormt bovendien een risico. De informatie die men verzamelt als docent kan gemakkelijk aantonen dat het onderwijs niet effectief was. Vaak zal ook niet bekend zijn, dat evaluatie op resultaat zonder voormeting problematisch is. Ten slotte is de docent vaak meer gericht op de eigen rol in het onderwijsprogramma en het eindniveau van de studenten, dan op de vraag wat het onderwijsprogramma daaraan bijdroeg.

Als evaluatie op leerwinst gebruikelijk zou zijn, was te verwachten dat van onderwijsprogramma's bekend is met hoeveel de doelvariabele gemiddeld verbetert. Dit gegeven ontbreekt echter vrijwel altijd, wat betekent dat er geen evaluatie op programma-effect heeft plaatsgevonden.

Ook vergelijking van de eindmeting met de uitkomsten van een soortgelijk, ander programma geeft geen uitsluitsel over de gemiddelde leerwinst. Wel kan men door vergelijking van programma's op de eindmeting, als de studenten strikt aselect zijn toegewezen aan de groepen, proberen vast te stellen welk programma beter werkte. Ook voor dat doel levert een voormeting echter nog waardevolle informatie op, doordat per student de leerwinst vastgesteld kan worden.

3.2 Onderwijsconstructie-probleem

De hiervoor geschetste onderwijsbenadering heeft niet alleen consequenties voor het evalueren van onderwijs, maar ook voor de constructie van onderwijs. Deze benadering gaat er-

van uit dat begonnen wordt met de formulering van een doel en een doelvariabele, daarna met het ontwikkelen van twee meetinstrumenten (een voor de pretest en een voor de post-test) en ten slotte volgt de constructie van het onderwijsprogramma.

Het is goed om hier bij stil te staan, want in de onderwijspraktijk wordt meestal precies andersom gewerkt. De focus is op het onderwijsprogramma dat gevuld moet worden. De vulling bestaat uit de stof die men wil behandelen. De docent of de opleiding kiest dus eerst de onderwerpen die behandeld moeten worden. Vervolgens wordt de werkvorm bedacht ofwel de manier waarop men de stof wil behandelen. Als laatste wordt een toets geconstrueerd voor de eindmeting. Uitgangspunt bij de toetsconstructie is dan dat de vragen betrekking moeten hebben op de behandelde stof.

In de praktijk begint men dus bij het programma in plaats van bij de doelvariabele. Daarna wordt op basis van het programma de doelvariabele ingevuld en gekozen. Men werkt niet vanuit het doel naar een geschikt middel, maar men construeert een middel waar vervolgens een passend doel bij geformuleerd wordt.

Zo stelde TLC Seminars (2009) in haar *Basic Instructor Training Course*: "Prior to putting a pen to paper to develop a lesson plan, you should have already completed and researched your subject material and carefully reviewed and developed your training objectives, and determined which method or methods of instruction you will use." Men begint bij de leerstof, daarna formuleert men doelstellingen en kiest men de instructiemethode. Over de toetsing wordt niets vermeld. Interessant in dit verband is de opmerking: "Attempting to write a lesson plan without prior planning would be the same as the old classic example of 'Getting the cart before the horse'." Men realiseerde zich kennelijk niet, dat dat precies is, wat men doet. Hoewel Tyler (1949) wel het belang zag van het vaststellen of de doelen van het onderwijs bereikt werden, is volgens de 'Tyler Rationale' deze fase de laatste in het ontwikkelproces. Ook Richards (2001) en Taba (1962) plaatsten de evaluatiefase als laatste in het constructieproces.

De gebruikelijke werkwijze bij onderwijsconstructie is als volgt te illustreren. Voor het schrijfonderwijs aan beginnende studenten, kiezen we de stof die we willen behandelen uit boek X. In de tien beschikbare lessen, behandelen we een aantal punten uit X. Op basis van deze behandelde punten construeren we vervolgens een toets. Of de student na afloop van het studieonderdeel inderdaad beter scoort dan voorheen, valt door het ontbreken van de be-

ginmeting niet te zeggen. Maar zelfs wanneer de student perfect scoort op de toets, betekent dat nog niet dat hij ook beter is gaan schrijven. De toets meet wel de stofbeheersing, maar vermoedelijk niet de schrijfvaardigheid. De validiteit van de toets is niet aangetoond voor de vaardigheid waar het in de praktijk om gaat.

3.3 Onderwijsmethode-probleem

Zodra de toets vastligt, rijst de vraag wat de beste onderwijsmethode is om studenten voor te bereiden op de toets. Hoe moet het onderwijsprogramma er idealiter uitzien om het beste resultaat te bereiken? Deze vraag kan nooit definitief beantwoord worden. Het is altijd mogelijk dat een ander ontwikkeld programma een nog groter effect blijkt te sorteren. Op basis van de beschikbare theorie over hoe mensen leren, kan wel aangegeven worden wat vermoedelijk belangrijke punten zijn voor een effectief onderwijsprogramma.

ABC-model

De belangrijkste punten voor een effectief onderwijsprogramma voor het verbeteren van basale schrijfvaardigheid kunnen worden ontleend aan het ABC-model uit de toegepaste gedragsanalyse. Het ABC-model is een andere term voor de 'three term contingencies' van operante conditionering (Cooper, Heron & Heward, 2007, p. 42). De drie componenten van het model zijn:

A. antecedents (omgeving, studietekst, vragen/opdrachten);
B. behavior (gedrag/antwoord);
C. consequences (gevolgen/feedback).

Deze drie componenten vormen de basis voor leren (en in stand houden) van operant gedrag: gedrag waarmee de student inwerkt, opereert op de omgeving. Door ervaring leert de student welke antwoord bij een bepaalde vraag een voor hem positief gevolg heeft (bijvoorbeeld juist gerekend wordt). Positieve gevolgen zijn hierbij gebeurtenissen die de student via zijn gedrag vaker probeert te laten optreden. De student probeert bijvoorbeeld meer juiste antwoorden te geven om zijn score te verhogen.

Tegenover operant gedrag staat respondent gedrag. Bij respondent gedrag werkt de omgeving in op de student en zijn gedrag. Dit gedrag wordt aangeleerd via respondente (klassieke) conditionering waarbij een neutrale prikkel gecombineerd wordt met een al geladen prikkel. Een parfumreclame werkt bijvoorbeeld door het onbekende en nog niet emotioneel geladen parfummerk te koppelen aan een beroemd, mooi model.

Het verschil tussen respondent en operant gedrag is dat het eerste onwillekeurig is en automatisch verloopt, terwijl het laatste door de student zelf bepaald wordt. Na de respondente conditionering roept de merknaam van het parfum automatisch een emotionele reactie op of men wil of niet, terwijl activiteiten als schaken en het beantwoorden van een vraag over de lesstof niet automatisch gaan, maar aandacht en doelgerichte activiteit van de student vragen.

In onderwijssettings gaan operant en respondent leren samen. Het goede antwoord wordt gevolgd door positief geladen feedback. In eerste instantie worden goede antwoorden op die manier positief geladen. In tweede instantie worden daardoor ook de vragen, opdrachten en de stof positief geladen. De student merkt dat hij het kan, de opgaven krijgen een positieve emotionele lading en de student ontwikkelt zelfvertrouwen. Het omgekeerde is echter ook mogelijk. De student heeft bijvoorbeeld de ervaring wiskundeopgaven vaak niet te kunnen. Het zien van een wiskundeopgave is dan voldoende voor een negatieve emotionele reactie die blokkerend kan werken.

Het ABC-model gaat ervan uit dat iemand leert door te doen. Wil er sprake zijn van leren, dan moet de student actief zijn. Er wordt daarom gewacht tot de student antwoord geeft. De student bepaalt het tempo. De activiteit van de student (B) wordt uitgelokt en gestuurd met vragen en opdrachten (A). Belangrijk is dat antwoorden snel gevolgd worden door duidelijke en positief geladen feedback (C) en dat feedback die voor de student negatief geladen is ('fout'), achterwege blijft of niet benadrukt wordt. Een student die de (Nederlandse) betekenis van een woord in een vreemde taal leert, moet dat woord zien of horen en daarna de betekenis geven. Na een goed antwoord volgt positieve feedback, dat wil zeggen feedback die het gewenste gedrag versterkt. Negatieve feedback werkt niet tegenovergesteld aan positieve feedback zoals vaak wordt aangenomen, maar verstorend in op het leerproces.

Testeffect

Op basis van dit leermodel is bijvoorbeeld te verwachten dat een testitem waarbij de student zelf - na eenmalige aanbieding van het woordenpaar - de betekenis formuleert, een groter leereffect oplevert dan een leeritem, waarbij de student het woordenpaar krijgt aangeboden en dit moet inprenten. Karpicke en Roediger (2008) lieten zien dat het aantal leeritems geen invloed heeft, maar het aantal testitems een zeer grote invloed heeft op het percentage woordbetekenissen dat onthouden wordt. Verder bleken docenten en studenten zich dit verschil in effectiviteit tussen leer- en testitems niet te realiseren.

Bij het bestuderen van studieteksten bleek herhaald proberen zich zo veel mogelijk te herinneren en te noteren van de bestudeerde tekst te leiden tot meer dan twee keer zo veel goed beantwoorde items bij de eindtest een week later (67% tegen 30%) dan het zelf bestuderen van de stof. Ook in dit geval was dat tegenovergesteld aan wat de studenten verwachtten (Karpicke & Blunt, 2011). Butler, Karpicke en Roediger (2008) merkten op: "Testing of information can have a powerful positive effect on future retention of the tested material, a phenomenon known as the *testing effect*" (p. 918).

Effect van feedback

In de voorgaande onderzoeken speelde feedback geen rol. Butler et al. (2008) onderzochten in twee experimenten het leereffect van een 4-keuze multiple-choice test en de invloed van feedback daarop. Items die niet getest waren in de eerste test, scoorden gemiddeld 29% goede antwoorden. Items die getest waren, scoorden 44% goede antwoorden in de eindtest. Wat was nu de invloed van feedback? Items die gevolgd waren door feedback scoorden 85% goede antwoorden bij de eindtest: een winst van 41 procentpunt. Feedback bleek een grote invloed te hebben bij items die aanvankelijk fout werden beantwoord (van 6% naar 78%), maar ook bij aanvankelijk goed beantwoorde items nog uit te maken (van 79% naar 93%). De auteurs stelden: "As shown in many studies, feedback is a critical aspect to learning, but instructors' policies in providing it vary considerably, ranging from comprehensive feedback after each testing occassion to little or no feedback at all" (p. 927).

Volgens het ABC-model moet niet de docent actief zijn, maar de student. De taak van de docent is beperkt tot het structureren van de omgeving via opdrachten en feedbackvoor-

zieningen. In afwijking van traditioneel onderwijs ligt het accent niet primair op het geven van informatie, maar primair op het prikkelen tot activiteit, door vragen en opdrachten, en motiveren door het geven van bekrachtigende feedback. De student werkt, de docent managet (Cooper, Heron & Heward, 2007; Heward, 2005; Holland, 1960; Jenson, Sloane & Young, 1988; Keller, 1968; Lindvall & Bolvin, 1967; Malott, 2008; Skinner, 1954; Skinner, 1958; Skinner, 1968; Vargas, 2009).

Opdracht als basis

Onderwijs waarbij de student moet luisteren naar de uitleg van een docent, is vanuit deze opvatting weinig zinvol, omdat er geen garantie is dat de student inderdaad de juiste activiteit ontwikkelt met betrekking tot de informatie. Luisteren is een essentieel andere taak dan de informatie oproepen of de informatie gebruiken zoals normaal bij de toets wordt gevraagd (Karpicke & Roediger, 2008; Karpicke & Blunt, 2011; Rohrer & Pashler, 2010). Volgens Holland (1960) kun je niet verwachten dat studenten zomaar het gewenste leergedrag gaan vertonen. "Behavior is learned only when it is emitted and reinforced. But in the classroom the student performs very little, verbally. . . . Not only is reinforcement needed for learning, but a high density of correct items is necessary" (p. 278-279).

Een onderwijsprogramma bestaat uit een reeks opdrachten die de student moet doorwerken. In het ideale geval zijn er een groot aantal opdrachten die veel goede antwoorden genereren. De structurering van de opdrachten is zo dat de opdrachten opklimmen qua moeilijkheidsgraad. Geleidelijk aan, zonder dat de student zich dit realiseert, komt hij op een hoger niveau. Studenten die vastlopen kunnen snel gelocaliseerd worden. Zodra er iets mis gaat bij het doorwerken van de methode, wordt dit via de antwoorden onmiddellijk zichtbaar.

Keller Plan of PSI

Keller (1968) heeft deze leertheoretische principes gebruikt voor het zogenaamde Keller Plan. Hij splitste leertaken op in kleine units, die bijvoorbeeld bestonden uit tekst met vragen. De student kon dit materiaal in eigen tempo doornemen. Iedere unit werd afgesloten met een toets en een gesprek met een tutor dat werkte als bekrachtiger en sociale controle. De norm voor de toets lag hoog. Wanneer de student de toets niet haalde, moest dezelfde eenheid opnieuw bestudeerd en getoetst worden tot deze ten slotte gehaald werd.

Het Keller Plan bleek de eerste onderwijsvernieuwing te zijn met een aantoonbaar positief effect in vergelijking met traditioneel onderwijs. Uit de meta-analyse van Kulik, Kulik en Cohen (1979) kwam naar voren dat het Keller Plan ofwel PSI (personalized system of instruction) een gemiddeld effect van 0.5 SD heeft. Een student die bij een test anders als vijftigste eindigt in een groep van honderd studenten, eindigt dankzij PSI als dertigste. Bij toetsing een aantal maanden na afloop van een cursus worden die verschillen nog groter. Het geleerde beklijft dus beter met PSI. Ook profiteren zwakkere studenten in dezelfde mate van dit onderwijs als betere studenten (Tyree, 1997).

Bij het Keller Plan ligt de nadruk op de actieve student, wat overigens niet verward moet worden met de de actieve student in het sociaal-constructivistische onderwijsmodel, waarbij ernaar gestreefd wordt dat de student zijn eigen doelen en leerproces vaststelt (Simons, 2000; Simons, 2006). De verantwoordelijkheid voor het leerproces wordt in dit constructivistische model bij de student gelegd. Emeritus-hoogleraar orthopedagogiek Stevens formuleerde het zo: "Hij [de leerling] neemt de verantwoordelijkheid op zich voor zijn eigen leren" (Zuidweg, 2006, p. 48). Bij Keller ligt die verantwoordelijkheid voor het leren juist bij de docent.

Er zijn wel overeenkomsten met andere sturende onderwijsmodellen, zoals het model van active teaching (Lowyck, 1994) en het model van directe instructie. Ook deze modellen kenmerken zich door duidelijke en expliciete doelen, hoge verwachtingen van de leerlingen, kleine stappen, opbouw in de lesstof en oefeningen met onmiddellijke feedback (Creemers, 1991; Veenman, 2001). Deze principes bleken niet alleen zinvol bij het aanleren van elementaire vaardigheden, maar ze bleken ook effectief bij het leren van complexe vaardigheden. Dit kwam naar voren uit onderzoek van Rosenshine (1997), die tegenstanders uitdaagde om met bewijzen te komen dat het niet werkt: "To those who discard teacher-led cognitive strategy instruction for discovery learning, I have a simple quote from a recent movie, modified slightly to fit education: 'Show me the data'."

Onderwijsfactoren

Uit observationeel onderzoek waarbij de bestaande onderwijssituatie gekwantificeerd en gerelateerd werd aan de gemeten onderwijsuitkomsten, komen factoren naar voren die in dezelfde richting wijzen. Van der Werf en Weide (1991) kwamen op basis van een vergelij-

kend onderzoek tussen 124 scholen uit op twee kenmerken die van belang zijn voor effectief onderwijs aan leerlingen die het Nederlands niet als moedertaal hebben. Op scholen waarbij het gemiddelde prestatieniveau hoger lag en goed gecompenseerd werd voor de nadelige effecten van een allochtone achtergrond, besteedde de leerkracht veel tijd aan basisvaardigheden en werden hoge eisen gesteld aan de doelen die voor deze vaardigheden bereikt moesten worden.

Onderzoek dat verband legt tussen kenmerken van nationale context, scholen, klassen, leerlingen aan de ene kant en onderwijsuitkomsten aan de andere kant laat zien dat het belang van onderwijsfactoren in verhouding tot leerlingfactoren in de praktijk beperkt is. Creemers (1994, p. 13) stelde in dit verband dat het percentage van de totale variantie in de leerresultaten dat verklaard wordt door onderwijsfactoren doorgaans ongeveer twintig procent is. Dit percentage is op grond van later onderzoek bijgesteld naar tien procent (Creemers & Kyriakydes, 2008; Scheerens & Bosker, 1997). Volgens Steenbergen (2009) toonde onderzoek van onder andere Opdenakker en Van Damme uit 2000 aan dat het nettoschooleffect nog veel minder is (minder dan twee procent) als gelet wordt op de invloed van de school op niet-cognitieve criteria als academisch zelfbeeld en zelfbewustzijn.

Bij de onderwijsfactoren blijken factoren op het niveau van de klas meer te verklaren dan die op schoolniveau. Gestructureerd onderwijs en effectieve leertijd zijn twee factoren die samenhangen met de effectiviteit van het onderwijs. Op het gebied van instructie blijken de kwaliteit en de hoeveelheid van de instructie belangrijke variabelen te zijn en op het gebied van leerstrategieën zijn dat bekrachtiging en feedback. Onderwijs met kenmerken als een ordelijk leerklimaat, een hoge actieve leertijd, hoge verwachtingen, een beperkt aantal duidelijke doelen, nadruk op basisvaardigheden en duidelijke presentatie van de stof met daarna oefeningen en feedback lijkt vaak beter te werken. Een zorgvuldig opgezet onderwijsprogramma met geleidelijke progressie is daarbij onmisbaar voor een goede leeromgeving (Binder & Watkins, 1990; Creemers, 1991; Creemers, 1994; Kirschner, Sweller & Clark, 2006; Parsons & Polson, 2000; Rosenshine 1997; Van der Werf & Weide, 1991; Veenman, 2001).

Invoeringsproblemen

Ruim dertig jaar geleden was bekend dat PSI een belangrijk effect sorteerde, maar dit heeft niet geleid tot grootschalige invoering. Bij de keuze voor een bepaald soort onderwijs viel

het juist op dat de vraag naar het effect ervan niet eens een rol speelde. Eind jaren negentig waren onderwijsvernieuwingen in Nederland niet gebaseerd op empirisch onderzoek naar wat wel en niet werkt (Van der Werf, 2005; Van der Werf, 2008). Dat lijkt overeen te komen met de eerdere opmerkingen in paragraaf 2.1 over onderwijsevaluatie, waarbij iedereen de eigen opvattingen als basis voor de beoordeling van het onderwijs neemt of is geneigd te nemen.

Dit is niet alleen het geval in Nederland. In de Verenigde Staten werden besluiten over onderwijsvernieuwingen evenmin genomen op basis van empirisch onderzoek. Het leereffect van verschillende onderwijsmethodes werd weliswaar in het grootschalige *Project Follow Through* over een periode van bijna dertig jaar onderzocht (1967-1995), maar vervolgens werd de directe instructie methode van Engelmann die veruit het meest effectief was, genegeerd (Coombs, 1998; Matthews, 2003; Nadler, 1998).

3.4 Implicaties voor deelstudies 1, 2 en 3

Het onderwijsevaluatie-probleem heeft betrekking op de vraag hoe de kwaliteit van onderwijs vastgesteld en gekwantificeerd moet worden. Niet de toegang tot onderwijs is het probleem, maar de kwaliteit ervan. Om de kwaliteit van onderwijs te kunnen meten en te verbeteren, moet eerst duidelijk zijn wat onderwijs is.

Onderwijs is een middel om studenten een vaardigheid te leren (kennis is de vaardigheid bepaalde vragen goed te beantwoorden). Wie deze opvatting accepteert, moet onderwijs niet evalueren op kenmerken van het programma of de mening van de student, maar op het resultaat (de leerwinst) bij de student. Om de leerwinst van een student te kunnen vaststellen is een voor- en nameting noodzakelijk om na te gaan in hoeverre het onderwijsprogramma bijdraagt aan het bereiken van de beoogde vaardigheid.

Voor de evaluatie (de meting van het effect) van het nieuwe onderwijsprogramma (onderzoeksvraag 4) is daarom uitgegaan van de volgende zes stappen:

1. er is een doel geformuleerd (foutloos schrijven);
2. er is een doelvariabele geformuleerd (het aantal fouten per honderd woorden);
3. deze is uitgewerkt in twee gelijkwaardige meetinstrumenten;

4. met deze instrumenten is een voormeting uitgevoerd;
5. de studenten hebben een van beide onderwijsprogramma's gevolgd;
6. met beide instrumenten is een nameting uitgevoerd.

Deze opzet maakte het mogelijk om de leerwinst die ieder onderwijsprogramma leverde te bepalen en beide onderwijsprogramma's met elkaar te vergelijken op leerwinst. Verder werd door deze opzet automatisch rekening gehouden met eventuele onbedoelde niveauverschillen tussen de groepen.

Bij het construeren van onderwijs wordt doorgaans eerst het programma ingevuld en daarna de toets ontwikkeld. Men construeert eerst het middel om vervolgens het doel aan het middel aan te passen. Bij de opzet van het nieuwe onderwijsprogramma (onderzoeksvraag 3) is precies andersom gewerkt: eerst is een doelvariabele geformuleerd, vervolgens zijn twee meetinstrumenten opgesteld, terwijl het programma pas als laatste is geconstrueerd.

Onderzoeksvraag 2 - wat is de waarde van bestaande methodes om basale schrijfvaardigheid te verbeteren bij eerstejaars hbo-studenten? - zou eenvoudig te beantwoorden geweest zijn, wanneer de verschillende methodes dezelfde doelvariabele hadden en wanneer het effect van het desbetreffende onderwijsprogramma op de doelvariabele bekend zou zijn. In de praktijk hebben bestaande methodes doorgaans geen duidelijk geformuleerde doelvariabele en is er geen informatie welke leerwinst het doorwerken van het desbetreffende programma oplevert. Een hard antwoord op deze vraag viel met de beschikbare gegevens daardoor niet te geven.

Voor de beoordeling van bestaande taalmethodes is daarom uitgegaan van het beoordelingsschema voor studieteksten van Van Es (1985). De basis voor dit schema is het ABC-model van operant gedrag. Het ABC-model gaat ervan uit dat we leren door te doen. De nadruk ligt op de activiteit van de student. Die wordt door vragen en opdrachten opgewekt en gericht. Door snelle en duidelijke feedback bij goede antwoorden wordt de student gemotiveerd. Onderwijsprogramma's worden gezien als een reeks vragen en opdrachten. Dit schema is vervolgens ook als uitgangspunt gebruikt voor de constructie van het nieuwe onderwijsprogramma.

4

Deelstudie 1

Foutenonderzoek*

* Resultaten van het foutenonderzoek werden gepresenteerd op de 10th ABC Conference (Association for Business Communication) in mei 2010 te Antwerpen (Van Eerden & Van Es, 2010), op ICSEI (International Congress for School Effectiveness and Improvement) in januari 2011 te Limassol, Cyprus (Van Eerden, Van Es & Van der Werf, 2011a), op de Rhetoric in Society lll Conference in januari 2011 te Antwerpen (Van Eerden & Van Es, 2011b) en op de ORD (Onderwijs Research Dagen) in juni 2011 te Maastricht (Van Eerden, Van Es & Van der Werf, 2011c).

4.1 Inleiding

In het foutenonderzoek zijn Nederlandstalige schrijfproducten van studenten beoordeeld op fouten om de vraag te beantwoorden wat het niveau van basale schrijfvaardigheid voor Nederlands is bij eerstejaarsstudenten in het hoger onderwijs. Deze onderzoeksvraag leidde tot drie vragen, namelijk tot de vraag of fouten objectief vaststelbaar zijn en tot de vraag hoeveel fouten, objectief gedefinieerd, in teksten van eerstejaarsstudenten voorkomen. De derde vraag was welke soorten fouten voorkomen.

4.1.1 Fouten in teksten

Wat precies onder taalfouten moet worden verstaan, is niet altijd duidelijk. Renkema (2005) heeft het in de Schrijfwijzer (zonder het daar expliciet over taalfouten te hebben) over het afkeuren of goedkeuren van een bepaald taalverschijnsel en koppelt dit aan zeven verschillende normen waaraan de taalgebruiker zich zou moeten houden. Hij merkt er echter onmiddellijk bij op dat vaak niet duidelijk is welke norm van toepassing is en dat de normen soms tot verschillende resultaten leiden (p. 16). In navolging van Renkema zou men daarmee een taalfout kunnen definiëren als een passage in een tekst die door een beoordelaar wordt afgekeurd.

De Nederlandstalige Wikipedia (2013) gaf als omschrijving: "Een taalfout is iedere inbreuk op het taalgebruik dat door een dominante gemeenschap als juist wordt ervaren." Deze omschrijving lijkt aan te sluiten bij de voorgaande, wanneer men zich realiseert dat men in de praktijk meestal met 'expert-beoordelaars' zal proberen te werken: mensen die geacht worden goed te kunnen schrijven. Een fout is dan wat een expert-beoordelaar signaleert als fout.

Williams (1981) onderscheidde bij een fout de ervaring van de student die de fout maakte, de ervaring van de taalkundige die de regel had bedacht die met de fout overtreden werd en de ervaring van de docent die de fout signaleerde. In deze opvatting is er pas sprake van een fout wanneer er een expliciete regel wordt overtreden. Volgens Wall en Hull (1989, p. 264) wees eerder onderzoek van Hull uit 1987 erop dat schrijvers fouten ruimer opvatten dan alleen het overtreden van regels. De veronderstelling dat er altijd een expliciete regel overtreden zou moeten zijn, lijkt daarmee onnodig beperkend.

Loerts (2012, p. 24-25) vermeldde dat uit onderzoek gebleken is dat mensen op zinnen met fouten reageren met specifieke reactiepatronen in het EEG. De N400 treedt op 400 milliseconden na het aanbieden van de prikkel bij semantische fouten en is een negatieve spanningspiek. De LAN (left anterior negativity) treedt op 300 tot 500 milliseconden na het begin van de prikkel bij syntaxfouten en is ook een negatieve spanningspiek. De P600 begint 500 milliseconden na het begin van de prikkel, piekt bij 600 milliseconden en kan doorlopen tot 1500 milliseconden en is een positieve spanningsgolf. De P600 treedt op bij een veelheid van syntaxfouten. Dit laat zien dat mensen onder bepaalde omstandigheden op taalfouten reageren met specifieke en kenmerkende reacties in het EEG.

Op grond van verschillende onderzoeken bleek dat een taalfout niet altijd een duidelijk vaststaand iets is (Anson, 2000; Connors & Lunsford, 1988; Lunsford & Lunsford, 2008; Robinson, 1998; Rose, 1985; Wall & Hull, 1989; Williams, 1981). Het criterium voor wat als fout wordt gezien, kan tussen beoordelaars verschillen. Een taalfout is ook tijdgebonden. Dit bleek uit een vergelijking van de lijst met meest voorkomende fouten van Connors en Lunsford (1988) en Lunsford en Lunsford (2008). Anson stelde: "the underlying rules that define error are themselves part of an organic and changing system" (Anson, 2000, p. 6/7).

Hoewel Connors en Lunsford het begrip 'fout' nuanceerden: "every teacher has his or her ideas what errors are common and important" (Connors & Lunsford, 1988, p. 396), stelden zij ook dat taalfouten niet genegeerd konden worden. Zij citeerden Shaughnessy (1977) die fouten zag als: "unintentional and unprofitable intrusions upon the consciousness of the reader. . . . They demand energy without giving back any return in meaning" (Connors & Lunsford, 1988, p. 396). Fouten zijn onvolkomenheden in de tekst die het voor de lezer lastiger maken de tekst te lezen.

Vermoedelijk om deze reden staan uitgevers, redacteuren, schrijvers en personen uit het bedrijfsleven vaak zeer negatief tegenover fouten in teksten. Iedere fout ziet men als een fout te veel. Men streeft naar optimale teksten. Voor docenten schrijfvaardigheid ligt de zaak mogelijk gecompliceerder. Een docent die iedere fout aanstreept in het werk van zijn studenten raakt veel tijd kwijt, terwijl het didactische effect op zijn studenten twijfelachtig is. Docenten schrijfvaardigheid zijn mogelijk mede daardoor soms geneigd fouten te zien als een normaal onderdeel van het schrijf- en ontwikkelingsproces waar ze zich verder niet al te druk over hoeven maken. Men kan niet leren schrijven, zonder fouten te maken.

Over de invloed van taalfouten bestaan veel meningen, maar empirisch onderzoek op dit gebied is veel schaarser. In deelstudie 4 (Effect van fouten) wordt hier verder op ingegaan. In het kader van dit foutenonderzoek (deelstudie 1) wordt uitgegaan van de veronderstelling dat fouten in een tekst ongewenst zijn en dat minder fouten (na correctie voor de lengte van de tekst) in het algemeen zal overeenkomen met een betere tekstkwaliteit.

Kwantificeren van fouten

Uitgaande van de omschrijving 'een fout is wat een expert-beoordelaar signaleert als fout', is de vraag of een volgende expert-beoordelaar onafhankelijk van de eerste beoordelaar, de fout ook zal signaleren. Zijn fouten objectief aantoonbaar? In de praktijk lijkt dit niet altijd zonder meer het geval te zijn. Indien sommige fouten wel objectief aantoonbaar zijn, betreft de volgende vraag het kwantificeringsprobleem. Het kwantificeringsprobleem bestaat uit twee gerelateerde deelvragen. De eerste deelvraag betreft de beste methode om de fouten die door verschillende beoordelaars gesignaleerd zijn, samen te nemen tot een stabiele score. De tweede gerelateerde deelvraag betreft het berekenen van de overeenstemming tussen beoordelaars.

In het beschikbare onderzoek naar fouten in teksten worden twee verschillende benaderingen gebruikt. In de ene benadering is een fout iets absoluuts dat zonder problemen door een expert-beoordelaar kan worden vastgesteld. In deze benadering wordt per tekst slechts één beoordelaar gebruikt, zodat niet duidelijk is of de geconstateerde fouten inderdaad door een andere beoordelaar teruggevonden zullen worden. In de andere benadering, die minder frequent voorkomt, worden per tekst meerdere beoordelaars gebruikt.

Bij deze laatste benadering (meerdere beoordelaars voor dezelfde teksten) doen zich drie problemen voor. Het eerste probleem is dat van de foutenexplosie. Naarmate meer beoordelaars worden ingezet, groeit het aantal fouten dat in de tekst(en) wordt gevonden. Williams bracht de foutenexplosie als volgt onder woorden: "So while it may seem useful for us to ask one another whether we think X is an error, we have to be skeptical about our answers, because we will invariably end up with more errors than we began with" (Williams, 1981, p. 154).

Wall en Hull (1989, p. 268) lieten 55 docenten een door een student geschreven tekst van ongeveer 400 woorden nakijken waarin volgens hen 35 fouten zaten. Dit leverde in totaal 1800 gesignaleerde fouten op of 32.7 per docent. Van alle plaatsen in de tekst die als fout waren aangestreept, was bijna twee derde (63.5%) aangestreept door minder dan 10% van de docenten. Over deze grote groep fouten bestond daarmee amper of geen overeenstemming.

Wall en Hull (1989) vonden ook fouten waarover wel meer overeenstemming bestond. Op 25 plaatsen in de tekst waren fouten gesignaleerd door ten minste 41% van de docenten, de zogenaamde 'high consensus errors' (Wall & Hull, 1989, p. 269). Deze groep fouten was daarmee relatief klein ten opzichte van de totale groep fouten.

De foutenexplosie lijkt te ontstaan doordat iedere beoordelaar fouten signaleert, die verder door geen enkele beoordelaar worden gesignaleerd. De ruis in het oordeel van de afzonderlijke beoordelaars wordt als het ware niet uitgefilterd, maar juist opgenomen in het eindsignaal, waardoor het aantal fouten waarover overeenstemming bestaat, steeds kleiner lijkt te worden. Hierna volgt een uitgebreidere verklaring.

Wanneer twee beoordelaars dezelfde tekst(en) op fouten analyseren lijkt het plausibel dat ze bepaalde fouten gemeenschappelijk zullen signaleren, terwijl iedere beoordelaar verder ook fouten zal signaleren die alleen hij zag. Bij drie beoordelaars zal het totale aantal niet-bevestigde fouten (fouten die slechts gesignaleerd werden door een enkele beoordelaar) nog groter worden en zo verder. Stel dat drie beoordelaars steeds 50 fouten in de tekst signaleren waarvan per beoordelaar 25 niet-bevestigde en 25 fouten die ook door de andere beoordelaars worden gevonden. In totaal signaleren de beoordelaars dan 150 fouten die resulteren in 100 tekstfouten. Van deze 100 tekstfouten zijn 75 niet-bevestigd en over 25 bestaat perfecte overeenstemming. Wanneer men nu uitgaat van alle 100 gevonden tekstfouten lijkt de overeenstemming minimaal. De beoordelaars zijn het immers slechts over 1 van de 4 fouten eens. Het toevoegen van een extra beoordelaar lijkt de situatie nog te verergeren. Nu worden 200 fouten gesignaleerd, wat resulteert in 125 tekstfouten waarvan er maar liefst 100 niet-bevestigd zijn. Men lijkt het nu nog maar eens te zijn over 1 van de 5 fouten. Wanneer de niet-bevestigde fouten echter worden uitgefilterd, bevat de tekst 25 fouten en zijn de beoordelaars het over die fouten onderling perfect eens. Op deze manier bekeken is de foutenexplosie een te verwachten statistisch verschijnsel, dat kan worden vermeden door uit te gaan van de fouten waarover beoordelaars het in zekere mate eens zijn.

Het is daarbij niet nodig te eisen dat de fouten door alle beoordelaars zijn gesignaleerd. Doordat beoordelaars lang niet alle in de tekst aanwezige fouten plegen te signaleren, zouden er in dat geval bij gebruik van veel beoordelaars amper fouten overblijven waarover volledige overeenstemming bestaat. Het zal in de praktijk voldoende zijn om als bevestigd te gelden, wanneer een fout door een onafhankelijke tweede beoordelaar is gesignaleerd. De kans dat een door een beoordelaar gesignaleerde fout wanneer die in de tekst niet aanwezig is, bevestigd wordt door een andere beoordelaar (op die specifieke plaats en met dezelfde omschrijving) is te verwaarlozen.

Een tweede probleem dat zich voordoet bij meerdere beoordelaars is dat men de betrouwbaarheid van de beoordeling of de overeenstemming tussen de beoordelaars zou willen kwantificeren. Dit is echter minder eenvoudig dan het lijkt. Stel dat twee beoordelaars alle fouten in een tekst proberen aan te strepen. Allereerst kunnen de strepen qua positie of lengte van elkaar verschillen, terwijl men misschien toch dezelfde fout bedoelt. Een tweede mogelijkheid is dat men dezelfde plek aanstreept, maar niet dezelfde fout bedoelt. Alleen aanstrepen is dus niet voldoende, men zal de fout ook moeten omschrijven. Vervolgens zal een nieuwe beoordelaar moeten beslissen of twee fouten wel of niet hetzelfde zijn.

Op basis van deze beoordeling ontstaat een lijst van alle tekstfouten waarin kan worden aangegeven of beoordelaar A de fout signaleerde (0=niet; 1=wel) en in een volgende kolom of beoordelaar B deze fout signaleerde. Voor het bepalen van de overeenstemming is men nu geneigd de correlatie te berekenen tussen de variabele van beoordelaar A en die van beoordelaar B. Deze werkwijze levert echter een negatieve correlatie op, doordat men niet aan beide beoordelaars dezelfde serie zinnen heeft voorgelegd met het verzoek aan te geven of de zin goed/fout is, maar doordat de foutenlijst is geconstrueerd op basis van de door de beoordelaars gesignaleerde fouten. Wanneer men een fout niet heeft gesignaleerd (een 0), betekent dit automatisch dat de andere beoordelaar die fout wel heeft gesignaleerd. Een correlatie om de overeenstemming uit te drukken, werkt daardoor in dit geval niet goed.

Hetzelfde probleem doet zich voor wanneer men als maat voor de overeenstemming tussen twee beoordelaars Cohens kappa zou willen gebruiken. Volgens de omschrijving meet Cohens kappa: "the agreement between two raters who each classify *N* items into *C* mutually exclusive categories" (Engelstalige Wikipedia, 2014). In dit geval worden echter geen N items geklassificeerd. Wanneer men de teksten zou opsplitsen in zinnen en iedere zin afzonderlijk aan iedere beoordelaar zou voorleggen, met het verzoek aan te geven of er

wel of niet een of meer fouten in de zin zitten, zou Cohens kappa wel berekend kunnen worden. In dat geval zou echter ook de correlatie (Pearsons r) berekend kunnen worden.

Nagaan hoeveel procent van de door een beoordelaar gesignaleerde fouten worden gesignaleerd door een andere beoordelaar, heeft ook beperkingen. Wanneer een beoordelaar in dit opzicht op veilig speelt, signaleert hij alleen zeer duidelijke fouten. Het resultaat is dat bijna alle gesignaleerde fouten bevestigd worden door de andere beoordelaar, maar dat veel fouten niet gesignaleerd worden. De omgekeerde benadering, nagaan hoeveel procent de beoordelaar heeft van de fouten van de andere beoordelaar, werkt dan mogelijk beter. Het probleem blijft dan dat dit vooral iets zegt over de desbetreffende beoordelaars en weinig over een specifieke fout. Die was immers of bevestigd of niet-bevestigd. De kans dat een bevestigde fout bevestigd zal worden door een derde beoordelaar, lijkt belangrijk groter dan de kans dat een niet-bevestigde fout bevestigd zal worden. De ene fout is daarmee kennelijk de andere niet.

In deze deelstudie is op verschillende manieren getracht het kwantificeringsprobleem - hoe meet je de overeenstemming tussen beoordelaars die fouten signaleren? - op te lossen. Het probleem is ten slotte opgelost door niet per fout te werken, maar met het aantal fouten per tekst (uitgaande van meerdere teksten). Beoordelaars vinden per tekst een bepaald aantal fouten. Na correctie voor de lengte van de teksten kan vervolgens tussen twee beoordelaars een correlatie berekend worden als maat voor de overeenstemming. Bij meer dan twee beoordelaars kan met de gemiddelde onderlinge correlatie tussen de beoordelaars worden gewerkt. Dit kan zowel gedaan worden voor alle gesignaleerde fouten per honderd woorden als voor alleen de bevestigde fouten per honderd woorden.

Een derde probleem wanneer meerdere beoordelaars teksten op fouten beoordelen, is dat de aantallen fouten tussen beoordelaars sterk kunnen verschillen. De ene beoordelaars is veel kritischer (signaleert meer fouten) dan de andere. Wall en Hull (1989, p. 268) rapporteerden dat de minst kritische docent 9 fouten in de tekst signaleerde en de meest kritische 56, dus meer dan 6 keer zo veel. Of dit echter daadwerkelijk een probleem is, is de vraag. Beoordelaars verschillen in gemiddelde en spreiding, maar statistisch gezien is het eenvoudig hiervoor te corrigeren door te standaardiseren op hetzelfde gemiddelde en dezelfde standaarddeviatie. Door indexen als de (productmoment) correlatie-coëfficiënt te gebruiken, gebeurt dit automatisch. De rangordening van de teksten en de (productmoment) correlatie veranderen door dit standaardiseren niet.

4.1.2 Aanzet operationalisering schrijfvaardigheid

Het basisprobleem bij het schrijfvaardigheidsonderwijs is het meetprobleem: beoordelaars zijn het niet eens over de kwaliteit van door studenten geschreven teksten. In 2.1 werd op dit punt al ingegaan. Voor het beoordelen van schrijfvaardigheid kijken docenten naar de geproduceerde tekst en geven daar een globaal (holistisch) oordeel over. Beoordelaars zijn het onderling meer oneens dan eens. De waarde van het holistische oordeel is daardoor in de praktijk beperkt. Het oordeel van de ene beoordelaar wordt niet of nauwelijks bevestigd door een volgende beoordelaar (Charney, 1984; Cooper, 1984; Gamaroff, 2000, Hyslop, 1990; Jansen & Wesdorp, 1973).

Pogingen om teksten te laten beoordelen op meerdere specifieke punten in plaats van af te gaan op een enkele algemene indruk (de zogenaamde analytische beoordeling) bleken niet tot een grotere overeenstemming tussen beoordelaars te leiden (Cooper, 1984; Rijlaarsdam & Wesdorp, 1984; Vanmaele, 2000). Wel bleken de beoordelingen van de verschillende punten door dezelfde beoordelaar onderling sterk te correleren, wat hun waarde beperkte, terwijl de beoordeling meer tijd ging vergen. Een beoordelingsvorm tussen analytische en holistische beoordeling in was 'primary trait' beoordeling. De beoordelaar kreeg opdracht een enkel specifiek punt te beoordelen, bijvoorbeeld spelling of interpunctie. Primary trait beoordeling was beperkter, maar vaak niet veel betrouwbaarder (Cooper, 1984).

De hierboven aangeduide vormen van beoordeling staan bekend als directe methodes. Het gebruik van specifieke objectieve toetsen bestaande uit meerkeuzevragen om de schrijfvaardigheid van studenten vast te stellen staat bekend als de indirecte methode (Cooper, 1984). Het voordeel van objectieve toetsen is dat ze betrouwbaar en - in tegenstelling tot de verwachting van veel docenten - valide blijken te zijn. Ze voorspellen de schrijfvaardigheid gemeten via het holistische oordeel van een groot aantal beoordelaars, vrijwel perfect. Specifieke objectieve tests gericht op het opsporen en corrigeren van fouten in zinnen bleken het holistische oordeel van een groot aantal beoordelaars zeer goed te kunnen voorspellen (Cooper, 1984).

Indirecte methodes hebben echter ook belangrijke nadelen. Een eerste nadeel is dat ze niet gebaseerd zijn op een tekst, maar op een toets. Er wordt dus geen specifieke tekst beoordeeld, maar een student. Voor de onderwijspraktijk zijn ze daardoor als instructiemiddel niet bruikbaar. Een tweede nadeel is dat de score niet absoluut interpreteerbaar is. Een stu-

dent die 80% goed scoort op een multiple-choice taaltest is misschien een goede schrijver, maar kan ook een slechte schrijver zijn. Dit hangt af van de moeilijkheid van de gebruikte test. Als meetmethode vormen multiple-choice taaltests daarom geen ideale oplossing voor het meetprobleem.

Het idee om schrijfvaardigheid te meten via meerkeuzevragen lijkt in eerste instantie nogal merkwaardig. Studenten schrijven dan niet, maar hoeven slechts te kiezen. Bij wijze van voorbeeld volgt hier een item ontleend aan Van Schooten (1988, Bijlage 3, Toets: Idioom en Stijl, deel C, vraag 3). De vraag gaat over zin 8 in een kort 'opstel' dat op een eerder blad is afgedrukt. Kortheidshalve geven we hier alleen zin 8 van het opstel.

> *8 Verder was er ook een zwembad, waar we, als het warm was, heengingen.*
>
> *---*
>
> *Hoe kun je zin acht het best verbeteren? (Als de zin goed is, kruis dan "A" aan.)*
>
> *A De zin is goed.*
>
> *B Verder was er ook een zwembad als het warm was waar we heengingen.*
>
> *C Verder als het warm was, waar we heengingen, was een zwembad.*
>
> *D Verder was er ook een zwembad, waar we heengingen als het warm was.*

De aanleiding voor dit onderzoek waren de vele fouten in teksten van studenten die een van ons onder ogen kreeg. Berichten uit de pers en verhalen van collega's bevestigden die indruk. Op grond van dit uitgangspunt lag het daardoor voor de hand om te kijken naar het aantal fouten in de teksten die studenten produceerden. Op het moment dat het mogelijk blijkt het aantal fouten in een tekst met een redelijke mate van betrouwbaarheid te bepalen, heeft men naar verwachting ook een maat voor basale schrijfvaardigheid. Het lijkt moeilijk voor te stellen dat schrijvers qua taalgebruik zo zouden fluctueren, dat ze de ene maal erg veel fouten in een tekst maken en de andere keer heel weinig. Verder bleken juist multiple-choice taaltests die zich richten op gangbaar taalgebruik en zinscorrectie goede en betrouwbare voorspellers te zijn van tekstkwaliteit gemeten via het holistische oordeel. Het aantal fouten in een tekst hangt daarmee mogelijk sterk samen met het holistische oordeel.

Een procedure waarin een beoordelaar de opdracht krijgt zo veel mogelijk fouten in een tekst aan te strepen en iedere fout te omschrijven, is te beschouwen als een vorm van primary trait beoordeling. De beoordeling is niet globaal: de beoordelaar kan niet volstaan met een enkele score. De beoordeling is ook niet analytisch: het is voldoende de tekst eenmaal door te werken. Door de aandacht van de beoordelaar volledig te richten op fouten, krijgt de beoordelaar aan de ene kant een eenvoudige opdracht en aan de andere kant wordt hij door die opdracht gedwongen de tekst nauwkeurig te lezen en door te werken. Tijdens het doorwerken produceert hij een schriftelijke neerslag: de plaatsen in de tekst waar hij een fout denkt te zien en zijn omschrijving van die fout. De beoordelingstaak resulteert daardoor in een concreet product, dat naderhand op verschillende manieren kan worden gecheckt.

Een eenvoudige manier is te kijken naar hoeveel fouten de beoordelaar gesignaleerd heeft. Een beoordelaar die erg weinig fouten heeft gesignaleerd, was vermoedelijk snel klaar, maar heeft mogelijk ook veel fouten gemist. Een meer verfijnde manier is te kijken naar het percentage bevestigde fouten dat hij gemist heeft. Een bevestigde fout is een fout die door twee of meer onafhankelijke beoordelaars is gesignaleerd. Een 'goede' beoordelaar die de instructie krijgt alle fouten te signaleren, zal weinig bevestigde fouten missen.

Een andere manier om de beoordeling te checken is een volgende beoordelaar achteraf de terechtheid van een fout te laten nagaan. Veel fouten zullen bij een achterafbeoordeling als 'terecht' beoordeeld worden, maar van sommige gesignaleerde fouten zal de achterafbeoordelaar het nut niet inzien.

Een meer statistische manier om een beoordelaar te evalueren, is te kijken naar de correlatie tussen het aantal fouten dat hij signaleert per tekst en het aantal dat een andere beoordelaar signaleert. Komt zijn oordeel over de teksten overeen (in termen van correlatie) met dat van de andere beoordelaar of wijkt het volledig af?

Een fout die een beoordelaar signaleert, moet vertaald kunnen worden in een tekstwijziging die een betere tekst oplevert. Een fout is dus niet iets vaags, maar iets vrij concreets. In beginsel zou via onderzoek aangetoond moeten kunnen worden dat de verbeterde tekst inderdaad beter is. In de praktijk is dat omslachtig en zal meestal afgegaan worden op het oordeel van ervaren taalgebruikers. Dit betekent echter wel dat een beoordeling op fouten een goed uitgangspunt vormt om een tekst daadwerkelijk te verbeteren. De beoordeling resulteert in gerichte feedback op basis waarvan de tekst herschreven zou moeten kunnen wor-

den. Een bezwaar van holistische beoordeling, dat de feedback van deze beoordeling weinig specifiek is, wordt daarmee ondervangen.

Door toepassing van meerdere onafhankelijke beoordelaars werd het mogelijk het holistische oordeel betrouwbaar te meten. Het oordeel van een enkele holistische beoordelaar was onbetrouwbaar, maar door een hele serie beoordelaars dezelfde teksten te laten beoordelen, werd de som van de oordelen betrouwbaar. Met 'betrouwbaar' wordt dan bedoeld dat bij herhaalde beoordelingen ongeveer dezelfde uitkomst wordt gevonden. De correlatie tussen beide beoordelingen wordt dan gebruikt als maat voor de betrouwbaarheid. Als meer dan twee metingen beschikbaar zijn, wordt gewerkt met de gemiddelde onderlinge correlatie of de daarvan afgeleide coëfficiënt alfa (Nunnally, 1967).

Willen fouten in een tekst bruikbaar zijn als maat voor tekstkwaliteit, dan moeten fouten in een tekst objectief vaststelbaar zijn. Dat wil zeggen dat verschillende beoordelaars op dit punt tot soortgelijke uitkomsten komen. Dit lijkt echter niet altijd het geval te zijn. Wat de ene beoordelaar ziet als een ernstige fout, ziet de andere beoordelaar soms als goed (Wall & Hull, 1989). Wanneer echter gewerkt wordt met de aantallen fouten per tekst, blijken beoordelaars het (in termen van correlatie) wel behoorlijk eens te zijn. Om op die manier naar fouten te kunnen kijken, was het noodzakelijk dat gewerkt werd met meerdere beoordelaars die onafhankelijk van elkaar werkten.

Een mogelijk bezwaar van het werken met fouten in onderwijssettings is dat het een foute didactische benadering inhoudt: een fout is iets negatiefs. Dit is echter ook een kwestie van hoe een fout opgevat en geïnterpreteerd wordt. Men kan een fout opvatten als een mogelijk verbeterpunt. Een fout is dan een punt in de tekst waarop de tekst mogelijk niet in orde is en verbeterd kan worden. Een taalfout valt te vergelijken met een cirkel die niet helemaal gesloten is. De open plek verstoort het ideaal en leidt de aandacht af. Wie een goede tekst wil schrijven, zal streven naar perfectie en zich niet afvragen of een fout meer of minder erg is. In het onderwijs wordt een fout ook vaak gezien als iets absoluuts waarover geen discussie mogelijk is. Een fout is fout, mag niet, is verkeerd en wordt de leerling aangerekend. Fouten zijn echter niet zo absoluut als soms gesuggereerd wordt. De ene beoordelaar zal veel meer fouten zien, dan de andere. Wat de ene beoordelaar 'fout' vindt, zal de ander mogelijk goed vinden. Men kan een fout dus ook wat relatiever interpreteren. Het doel van dit deelonderzoek was in eerste instantie basale tekstkwaliteit te meten en daarom leek de negatieve associatie met 'fout' voor dit doel geen probleem.

Verwachtingen

Het eerste doel van het foutenonderzoek was na te gaan of fouten objectief vaststelbaar zijn. De meest vergaande verwachting hierbij was dat fouten gemakkelijk te herkennen zouden zijn en dat er verder geen discussie over mogelijk zou zijn. Een fout is een fout. Wall en Hull (1989) formuleerden het als volgt: "Sentence-level error seems to be one part of written language upon whose nature everyone agrees" (p. 262). Teksten zouden lastig te beoordelen zijn, maar niet op het niveau van basale fouten. Een schrijfvakman zou zonder veel problemen de fouten uit een tekst kunnen halen; een andere vakman zou precies dezelfde fouten gevonden hebben.

Het tweede doel van het foutenonderzoek was na te gaan hoeveel fouten voorkomen in teksten van eerstejaarsstudenten. Hierbij was de verwachting dat het in slechte teksten moeilijker zou zijn om fouten te lokaliseren, omdat een beoordelaar dan als het ware door de bomen het bos niet meer zou zien.

Het derde doel was na te gaan welke soorten fouten gemaakt werden. De derde verwachting was - op grond van artikelen over taalfouten, de mening van collega's en eigen ervaring met teksten van studenten - dat fouten in werkwoordsvormen (d/t-fouten) het meest frequent zouden voorkomen.

4.2 Methode

Om de methode voor het foutenonderzoek helder te krijgen is eerst een pilotonderzoek uitgevoerd.

4.2.1 Pilotonderzoek

Het doel van het pilotonderzoek was de methode voor het foutenonderzoek te ontwikkelen, te testen en bij te stellen. Een ander doel van de pilot was na te gaan hoe groot de onderlinge overeenstemming tussen de beoordelaars was.

Methode

Tien teksten afkomstig van hbo-studenten (5) en van universitaire studenten (5) werden voorgelegd aan 3 beoordelaars die zelf goed konden schrijven. Iedere beoordelaar werkte volledig onafhankelijk van de andere beoordelaars. De beoordelaars kregen de instructie de in het oog lopende fouten te onderstrepen en te nummeren en op een apart vel de fout kort te omschrijven.

Resultaten

De beoordelaars bleken zeer te verschillen in de percentages van alle bevestigde fouten die ze signaleerden. Per beoordelaar was dit percentage respectievelijk 66, 35 en 30. Niet iedere beoordelaar was even goed in dit opzicht.

Wanneer de bevestigde fouten gezien werden als de items van een test waarbij de beoordelaars de taak hadden alle bevestigde fouten te signaleren, was de coëfficiënt alfa van deze test 0.93. Het percentage bevestigde fouten dat een beoordelaar signaleerde kon daarmee zeer betrouwbaar worden vastgesteld.

De gemiddelde onderlinge correlatie voor het aantal gesignaleerde fouten per honderd woorden tussen de beoordelaars voor de 10 teksten bedroeg 0.67. De overeenkomstige alfa-standardized was 0.86. Voor de bevestigde fouten bedroeg de gemiddelde onderlinge correlatie 0.90 en de alfa 0.95.

De correlatie tussen tekstlengte en het aantal bevestigde fouten per honderd woorden bleek -0.65 te zijn. Dit was significant (N=10, p=0.04, tweezijdig). Met andere woorden: hoe langer de tekst, des te minder fouten per honderd woorden. Studenten met een gebrekkige schrijfvaardigheid schrijven kortere teksten.

Conclusies

Wat kan de pilot bijdragen aan verbeteringen voor de aanpak van het foutenonderzoek?

De instructie was dat men 'in het oog lopende' fouten moest aangeven. Dat is iets anders dan 'alle' fouten. Het idee achter de instructie was nu in feite dat de beoordelaar twee taken tegelijk uitvoerde. De fouten lokaliseren en tegelijkertijd beoordelen of het een in het oog lopende fout was. Deze combinaties van taken maakt de beoordeling onnodig complex.

In de tweede plaats was het nakijken van de teksten op fouten een lastige en open taak. In beginsel kon op ieder punt in de tekst een groot aantal verschillende fouten zitten. Wanneer in een tekst veel fouten tegelijk en achterelkaar worden gemaakt, is het gemakkelijk om te zien dat er iets mis is, maar het is lastig om te achterhalen welke fouten er allemaal precies aanwezig zijn. Na de tweede tekst blijkt in de praktijk al vermoeidheid op te treden, terwijl beoordelaars juist geneigd waren de klus af te maken en zoveel mogelijk teksten achterelkaar te doen. Men zag dan als het ware de fouten niet meer.

Het nakijken kostte verder veel tijd, doordat voor het nakijken naar schatting ongeveer 1 minuut per fout nodig was per beoordelaar. Een beoordelaar die 90% van de fouten vond, was dus .90 x 466 (het aantal fouten in de 10 teksten) = 419 minuten kwijt met nakijken of 7 uur voor deze 10 teksten. Per tekst kostte het nakijken in dat geval gemiddeld een kleine drie kwartier (42 minuten).

Beoordelaars werden in de pilot niet uitgedaagd om zo veel mogelijk fouten op te sporen, omdat ze er niet van op de hoogte waren dat per beoordelaar een succespercentage te berekenen viel.

Ook is de ene beoordelaar de andere niet. Het percentage bevestigde fouten dat een beoordelaar signaleerde, liep voor de drie beoordelaars sterk uiteen. Deze verschillen kunnen zijn ontstaan, doordat de ene beoordelaar veel meer ieder foutje rapporteerde dan de andere beoordelaars. De werkwijze van de beoordelaars was ook verschillend: slechts een paar teksten per keer of een grotere serie teksten. De taalvaardigheid van de beoordelaar zelf is vermoedelijk eveneens van belang.

Op basis van de bovengenoemde factoren was het mogelijk de aanpak en de instructie voor het foutenonderzoek bij te stellen.

De instructie was dat men 'in het oog lopende' fouten moest aangeven. Dit bleek een complexe taak te zijn. In de praktijk ging de ene beoordelaar alle fouten benoemen, terwijl de

andere beoordelaar zich probeerde te richten op de in het oog lopende fouten. De nieuwe instructie zal daarom vragen om alle fouten te lokaliseren. Hierdoor is de beoordelingstaak vereenvoudigd, waardoor de kans kleiner wordt dat men fouten niet meldt of mist.

Ook zal de herziene instructie expliciet stellen dat men per keer niet meer dan twee teksten tegelijk moet nakijken. Het bleek dat na de tweede tekst vermoeidheid optreedt, zodat men fouten over het hoofd zag.

Het was achteraf mogelijk om het succespercentage van een beoordelaar te berekenen. Dit zal eveneens in de instructie vermeld worden, zodat het voor de beoordelaars motiverender is om optimaal te presteren.

Uit de resultaten bleek dat beoordelaars behoorlijk van elkaar kunnen verschillen en dat het van belang is om beoordelaars te gebruiken die hoog scoren, in die zin dat zij een groot aantal bevestigde fouten opsporen. Na afloop of op basis van een pretest wordt dit overigens pas duidelijk. De geschiktheid van een beoordelaar kan vooraf getest worden aan de hand van enkele proefteksten. De kwaliteit van de beoordelaars bleek namelijk betrouwbaar gemeten te kunnen worden via een beperkt aantal teksten. Uitgaande van de 10 teksten en deze minimale steekproef van 3 'proefpersonen' was de coëfficiënt alfa 0,93.

Verder is het aantal bevestigde fouten dat gevonden wordt sterk afhankelijk van het aantal beoordelaars. Door meer beoordelaars van een bepaalde kwaliteit in te zetten, neemt niet het aantal gevonden fouten in de tekst toe, maar vooral ook het aantal gevonden bevestigde fouten.

4.2.2 Methode Foutenonderzoek

Selectie materiaal

Het basismateriaal bestond uit 159 teksten die door vijf groepen eerstejaarsstudenten geschreven werden: 127 teksten waren afkomstig van drie groepen eerstejaars hbo-studenten

van de Hanzehogeschool Groningen die de opleiding International Business and Languages (IBL) volgden en 32 teksten waren afkomstig van twee groepen universitaire eerstejaarsstudenten van de studie Communicatie- en Informatiewetenschappen (CIW) van de Rijksuniversiteit Groningen. De teksten van de universitaire studenten zijn opgenomen om de uitkomsten van de hbo-studenten in perspectief te kunnen plaatsen. Om selectie te voorkomen, zijn alle ingeleverde teksten van de studenten van deze vijf groepen gebruikt.

De teksten hadden een omvang van maximaal één A4. Ze zijn in het cursusjaar 2007-2008 geproduceerd in het kader van reguliere schrijfopdrachten. De CIW-studenten kregen bij het vak Taalbeheersing de opdracht om een toegankelijke tekst te schrijven over het onderwerp: waarom kiezen studenten Groningen om te gaan studeren. IBL-studenten moesten bij het vak Bedrijfscommunicatie/Nederlands een prettig leesbare tekst maken in correct Nederlands voor een brede doelgroep over de voor- en nadelen van light producten (71 teksten). Verder bestond het hbo-materiaal uit Inleidingen die geschreven werden in het kader van projectverslagen (56 teksten).

De werkstukken werden thuis uitgewerkt en geprint ingeleverd met uitzondering van 31 teksten over light producten die tijdens het werkcollege werden geschreven. Indien deze teksten handgeschreven waren en in de steekproef bleken te vallen, zijn ze daarna ingevoerd en geprint, waarna de printversie gecheckt is op afwijkingen van de handgeschreven versie.

Steekproeftrekking

Alle teksten werden genummerd. De trekking van de steekproef vond plaats via de randomfunctie van de rekenmachine. Wanneer een toevalsgetal getrokken werd, dat niet (meer) in de desbetreffende groep voorkwam, is opnieuw een getal getrokken. De opbouw van de steekproef was daarbij als volgt:

- 5 thuisgemaakte teksten van hbo-studenten over light producten;
- 5 tijdens het werkcollege gemaakte teksten van hbo-studenten over light producten;
- 10 Inleidingen van hbo-projectverslagen;
- 10 teksten van universitaire studenten.

Wanneer bij de trekking uit de Inleidingen een tekst werd getrokken van een student die al in de steekproef vertegenwoordigd bleek te zijn bij de teksten over light producten, werd deze tekst terzijde gelegd en werd opnieuw getrokken. Een enkele keer bleek een tekst getrokken te worden die afkomstig was van een Duitse student die nog maar net in het Nederlands schreef. Deze teksten leken niet representatief te zijn voor hoe een gemiddelde student schreef en werden daarom terzijde gelegd, waarna opnieuw een tekst werd getrokken.

In totaal werden 30 teksten getrokken, 20 van hbo-studenten en 10 van universitaire studenten. Deze verdeling werd zo gekozen, omdat het foutenonderzoek primair tot doel had fouten die hbo-studenten maken in beeld te brengen. De universitaire studenten werden in het onderzoek betrokken om na te gaan of er verschillen bestonden met de hbo-studenten.

Selectie beoordelaars

Getracht werd om beoordelaars te vinden met een goede schrijfvaardigheid. De veronderstelling was dat naarmate beoordelaars beter schrijven, zij ook beter kunnen oordelen over de schrijfproducten van anderen. Voor het beoordelen van de 30 teksten werden een docent Nederlands, een docent Engels en een journalist gevraagd. De beoordelaars ontvingen een vergoeding voor het beoordelingswerk. Verder heeft een van beide onderzoekers gefungeerd als beoordelaar. Alle vier beoordelaars schreven zelf beroepsmatig of doceerden taalvaardigheid of hadden dat gedaan. De leeftijd varieerde van begin twintig tot eind vijftig.

Instructie

De beoordelaars kregen de instructie alle fouten in de 30 teksten te onderstrepen, te nummeren en op een afzonderlijk vel de fout te omschrijven (zie bijlage 2). Het was niet mogelijk te volstaan met alleen het onderstrepen van een fout. De omschrijving van een fout was belangrijk. Door een fout te omschrijven werd duidelijk of verschillende beoordelaars het over dezelfde fout hadden. De fout moest op een apart vel omschreven worden in woorden naar eigen keuze, daardoor had de beoordelaar letterlijk de ruimte om commentaar te geven.

Bij het omschrijven van een fout was het mogelijk om de beoordelaars een standaardlijst met fouten te geven waaruit men kon kiezen. Bij iedere fout moet dan steeds opnieuw de lijst afgelopen worden op zoek naar de meest passende omschrijving. Uit de beschrijving van Lunsford en Lunsford (2008) viel af te leiden dat het werken met een dergelijke lijst problemen geeft: volgorde effecten, ontbrekende categorieën en vage, onduidelijke categorieën. Om deze redenen werd ervoor gekozen om de beschrijving van de fout aan de beoordelaar over te laten.

De instructie week af van de instructie in het pilotonderzoek. Daar moesten de beoordelaars 'in het oog lopende fouten' aangeven. Dit bleek een onduidelijke opdracht te zijn. In de praktijk ging de ene beoordelaar alle fouten benoemen, terwijl de andere beoordelaar zich probeerde te richten op de in het oog lopende fouten. De instructie in het foutenonderzoek vroeg daarom om 'alle fouten' te lokaliseren. Hierdoor werd het ook mogelijk per beoordelaar na te gaan, welke fouten hij gemist had.

Iedere beoordelaar beoordeelde alle 30 teksten. De beoordelaars werkten zelfstandig en onafhankelijk van elkaar, in hun eigen tempo en zonder expliciete tijdsdruk. De teksten werden via toevalsgetallen in een willekeurige volgorde geordend om volgorde-effecten tussen de verschillende soorten teksten te vermijden. De instructie (zie Tabel 4.1 voor een ingekorte versie) aan de beoordelaars stelde expliciet dat per keer niet meer dan twee teksten tegelijk moesten worden nagekeken. Uit de pilot was gebleken dat het beoordelen van een tekst een half uur tot een uur kon kosten. Beoordelaars waren in de pilot geneigd om de klus af te maken en zoveel mogelijk teksten achterelkaar te doen. In de praktijk bleek na de tweede tekst al vermoeidheid op te treden.

Bij de pilot bleek dat het mogelijk was om achteraf het succespercentage van een beoordelaar te berekenen. Deze mogelijkheid werd expliciet in de instructie vermeld om de beoordelaars te motiveren tot een optimale inzet.

Verwerking

De beoordelingen van de vier beoordelaars werden verwerkt door op een schone kopie van de tekst te noteren waar de fout zat via een foutnummer boven het woord of de passage. Dit werd gedaan om de fouten eenvoudig te kunnen linken aan de tekst. Op een vel papier werd

het foutnummer genoteerd en het nummer dat de fout van beoordelaar 1, 2, 3 en 4 had gekregen. Per tekst heeft de onderzoeker die niet als beoordelaar fungeerde dus voor iedere aangestreepte fout vastgesteld of deze ook door de andere drie beoordelaars was aangestreept met een soortgelijke omschrijving. Door deze procedure werden de vier beschrijvingen samengevat tot één definitieve foutenlijst.

De gegevens werden daarna ingevoerd in een SPSS-datafile waarbij iedere in de tekst aanwezige fout een regel ('case') vormde met het foutnummer voor identificatie van de fout en de beoordelaars die de fout wel (1) of niet (0) gesignaleerd hadden. Iedere beoordelaar vormde hierbij een variabele (een kolom). De minimale score van een fout was hierbij 1 (slechts één beoordelaar signaleerde de fout), de maximale score 4 (alle beoordelaars signaleerden de fout).

Tabel 4.1 Ingekorte instructie beoordelaars foutenonderzoek

Boordeel de teksten in de volgorde waarin ze in de map zitten.

Beoordeel niet meer dan twee teksten achter elkaar om het missen van fouten door vermoeidheid te voorkomen.

Als dezelfde fout vaker voorkomt, deze steeds opnieuw onderstrepen en nummeren.

Voor iedere beoordelaar wordt achteraf het percentage gevonden fouten berekend van fouten die ook door andere beoordelaars gevonden zijn.

=====

0. Vermeld op het bijgaande schrijfpapier het nummer van de tekst die u beoordeelt.

1. Geef in de tekst alle fouten aan door die te onderstrepen. Vermeld hierbij ook een nummer (bij iedere volgende fout een volgend nummer gebruiken, ook al gaat het om dezelfde soort fout).

2. Zet vervolgens op het bijgaande papier het nummer van de fout en geef een korte omschrijving van de fout.

'Bevestigde' fouten waren fouten die door twee of meer beoordelaars werden gesignaleerd. Wanneer een fout gesignaleerd was door een beoordelaar, waren er twee mogelijkheden. De eerste was dat de fout alleen gesignaleerd werd door deze ene beoordelaar. De andere mogelijkheid was dat de fout ook door ten minste één andere beoordelaar was gesignaleerd. In het eerste geval, de fout werd niet bevestigd, was de desbetreffende beoordelaar kennelijk de enige die op dat punt een specifiek probleem zag in de tekst en het is daarmee de vraag of er op die plaats in de tekst werkelijk een probleem is. In het tweede geval is er ook een onafhankelijke andere beoordelaar die op hetzelfde punt in de tekst een soortgelijke omschrijving geeft. De kans dat op dit punt in de tekst werkelijk een probleem aanwezig is, lijkt daarmee belangrijk groter te zijn. Verder kan een beoordelaar relatief gemakkelijk eindeloos veel fouten signaleren, maar is het aantal bevestigde fouten dat gesignaleerd kan worden in een tekst beperkt.

Alle fouten zijn achteraf opnieuw beoordeeld door de onderzoeker en het resultaat van deze beoordeling is als aparte variabele in de datafile opgenomen**.** Sommige fouten bleken namelijk niet echt fout te zijn en andere waren vatbaar voor discussie. De uitkomsten van deze achterafbeoordeling waren beperkt tot 'klopt' of 'klopt niet'. In het geval dat een bevestigde fout werd beoordeeld als niet-kloppend, werd een tweede achterafoordeel gevraagd aan een andere beoordelaar. Wanneer deze de fout ook beoordeelde als niet-kloppend, werd het eindoordeel 'klopt niet'. Anders werd het eindoordeel 'klopt'.

De bevestigde, kloppende fouten zijn op basis van de letterlijke tekst en het commentaar van de beoordelaars, zoals weergegeven is in bijlage 3, ingedeeld in twintig soorten fouten. Hierbij was de onderzoeker vrij in de keuze van het aantal categorieën en in de omschrijvingen. Het doel hiervan was een indruk te krijgen van de fouten die studenten in doorsnee maken.

4.3 Resultaten

4.3.1 Is een 'fout' echt een fout?

In totaal werd 3980 maal een fout door een beoordelaar gesignaleerd. Deze 3980 signaleringen bleken na analyse betrekking te hebben op 2400 verschillende fouten. Van die fou-

ten waren 1411 idiosyncratische fouten en 989 fouten werden bevestigd door een andere beoordelaar.

De eerst geformuleerde verwachting was, dat een fout eenvoudig vastgesteld kon worden en dat er over het bestaan en de juistheid geen discussie mogelijk zou zijn. Een andere expert-beoordelaar zou precies dezelfde fouten signaleren. Aan het bestaan van een (gesignaleerde) fout zou niet getwijfeld kunnen worden. Indien de vier beoordelaars het voortdurend eens waren geweest, hadden de 3980 foutsignaleringen geresulteerd in 995 (3980:4) fouten die door alle vier beoordelaars zouden zijn gesignaleerd. Indien de vier beoordelaars het voortdurend volledig oneens waren geweest, hadden de 3980 signaleringen geresulteerd in 3980 verschillende fouten die door steeds slechts een beoordelaar waren gesignaleerd. Beide stellingen werden niet bevestigd. De waarde van 2400 gevonden verschillende fouten lag tussen beide waarden in.

Gemiddeld signaleerde iedere beoordelaar 995 fouten (zie Tabel 4.2). Daarvan werden 642 bevestigd door een andere beoordelaar. Dat komt overeen met 65%. Ruwweg 2 van de 3 fouten die een beoordelaar signaleerde, werden daarmee bevestigd. Ongeveer 1 van de 3 gesignaleerde fouten werd niet bevestigd. De veronderstelling dat een gesignaleerde fout een soort absoluut gegeven is, waaraan niet getwijfeld kan worden, bleek daarmee niet te kloppen. Het tegenovergestelde bleek echter ook onjuist: van de gesignaleerde fouten werd ongeveer twee derde wel bevestigd. De meeste gesignaleerde fouten zijn daarmee 'echte' fouten.

Tabel 4.2 Gemiddelde aantallen fouten per beoordelaar

Gesignaleerde bevestigde fouten	642	65%
Gesignaleerde niet-bevestigde fouten	353	35%
Totaal gesignaleerde fouten per beoordelaar	995	100%

Gesignaleerde bevestigde fouten	642	65%
Gemiste bevestigde fouten	347	35%
Totaal bevestigde fouten (alle 4 beoordelaars)	989	100%

Dat een fout niet bevestigd werd, betekende overigens nog niet dat de fout niet-kloppend zou zijn. Bij de achterafbeoordeling werd van de niet-bevestigde fouten in totaal 67% beoordeeld als kloppend en 33% als niet-kloppend. Van de niet-bevestigde fouten waren volgens de achteraf-beoordeling dus twee van de drie fouten kloppend. Voor de bevestigde fouten lagen deze percentages op respectievelijk 92% en 8%.

Volgens de eerst geformuleerde verwachting zouden fouten niet alleen altijd bevestigd worden door een tweede beoordelaar, maar ook altijd gevonden en gesignaleerd worden. Ook dit deel van de verwachting bleek niet te kloppen. Beoordelaars bleken veel bevestigde fouten te missen (niet te signaleren). Van alle 989 bevestigde fouten die via de vier beoordelaars gevonden werden gevonden, werd per beoordelaar gemiddeld ruim een derde (35%) gemist (zie Tabel 4.2).

Bij beoordelingstaken is het vaak gebruikelijk de uitkomsten onder te brengen in een 2x2 tabel. Een arts krijgt bijvoorbeeld 1000 röntgenfoto's te beoordelen. Bij 120 foto's signaleert hij een probleem. Wanneer een tweede arts nu dezelfde foto's beoordeelt, kunnen de uitkomsten worden ondergebracht in een 2x2 tabel. Deze tweede arts zal bij bijvoorbeeld 180 foto's een probleem constateren, waarbij ze het over 50 probleemgevallen eens zijn. Op basis van deze getallen valt de tabel te construeren en valt af te leiden dat ze het in totaal over 800 foto's eens zijn en over 200 niet. Op basis hiervan kan een correlatie of een andere index voor de overeenstemming berekend worden. Men is geneigd te veronderstellen dat dit ook mogelijk zou moeten zijn, bij fouten die door beoordelaars in teksten worden gesignaleerd. Dit blijkt echter niet op te gaan doordat in het laatste geval geen sprake is van afzonderlijke items. De beide artsen beoordeelden 1000 foto's die wel of niet een probleem konden bevatten. In het geval van een taaltekst kan men echter niet vooraf stellen dat er bepaald aantal fouten aanwezig is. De situatie zou anders worden, wanneer men de beoordelaars een bepaald aantal afzonderlijke zinnen liet beoordelen op correctheid.

Wanneer we veronderstellen dat de ideale beoordelaar alle bevestigde fouten zou vinden en geen onbevestigde fouten zou signaleren, dan zou hij 989 fouten rapporteren. In werkelijkheid miste de gemiddelde beoordelaar 347 bevestigde fouten, terwijl hij 353 fouten rapporteerde die niet-bevestigd werden. Het aantal gemiste fouten was daarmee gemiddeld ongeveer even groot als het aantal gesignaleerde fouten dat niet bevestigd werd. Gemiddeld genomen kwam het aantal gesignaleerde fouten (995) daarmee vrijwel perfect overeen met

het aantal bevestigde fouten (989). Wat men aan de ene kant over het hoofd zag, maakte men goed door aan de andere kant wat extra fouten te 'verzinnen'.

Het percentage gesignaleerde bevestigde fouten varieerde sterk tussen de beoordelaars (zie Tabel 4.3). De slechtste beoordelaar in dit opzicht (b4) miste meer dan de helft van alle bevestigde fouten. De beste beoordelaar (b1) signaleerde 80%, maar miste toch nog steeds 20% van alle bevestigde fouten. Zelfs de beste beoordelaar signaleerde lang niet alle bevestigde fouten.

Tabel 4.3 Percentage van de bevestigde fouten (bf) dat een beoordelaar signaleerde en het percentage dat bevestigd werd van de door hem gesignaleerde fouten (gf). Tevens de correlatie tussen het aantal door hem gesignaleerde fouten en het aantal door de overige beoordelaars gesignaleerde fouten per honderd woorden voor 30 teksten

beoordelaar	gesignaleerde bf	bevestigde gf	correlatie
b1	80%	49%	.90
b3	56%	81%	.92
b4	46%	87%	.92
b5	77%	66%	.85
b1+b3+b4+b5	100%	65%	

4.3.2 Is een 'slechte' tekst echt een slechte tekst?

Uit Tabel 4.3 blijkt dat de beoordelaars een bepaalde benadering volgden. De beoordelaars b3 en b4 volgden een zuinige benadering. Men was terughoudend met het signaleren van fouten, dit in afwijking van het verzoek in de instructie om 'alle' fouten te signaleren. Hierdoor werd het percentage niet-bevestigde fouten laag gehouden, maar miste men veel bevestigde fouten. De beoordelaars b1 en b5 volgden een royale benadering. Ze probeerden zoveel mogelijk fouten te signaleren zonder zich druk te maken over de vraag of een andere beoordelaar die ook zou zien. Het resultaat was dat zij veel bevestigde fouten vonden, maar ook veel niet-bevestigde fouten signaleerden.

In Figuur 4.1 zijn de percentages uit Tabel 4.3 voor de vier beoordelaars grafisch weergegeven. Hoewel het aantal beoordelaars klein was, lijkt het verband tussen beide variabelen onmiskenbaar. De correlatie bedroeg 0.94 en de p-waarde was 0.03 bij eenzijdige toetsing. Een beoordelaar die erin slaagde veel van alle bevestigde fouten te signaleren, signaleerde ook veel fouten die niet bevestigd werden.

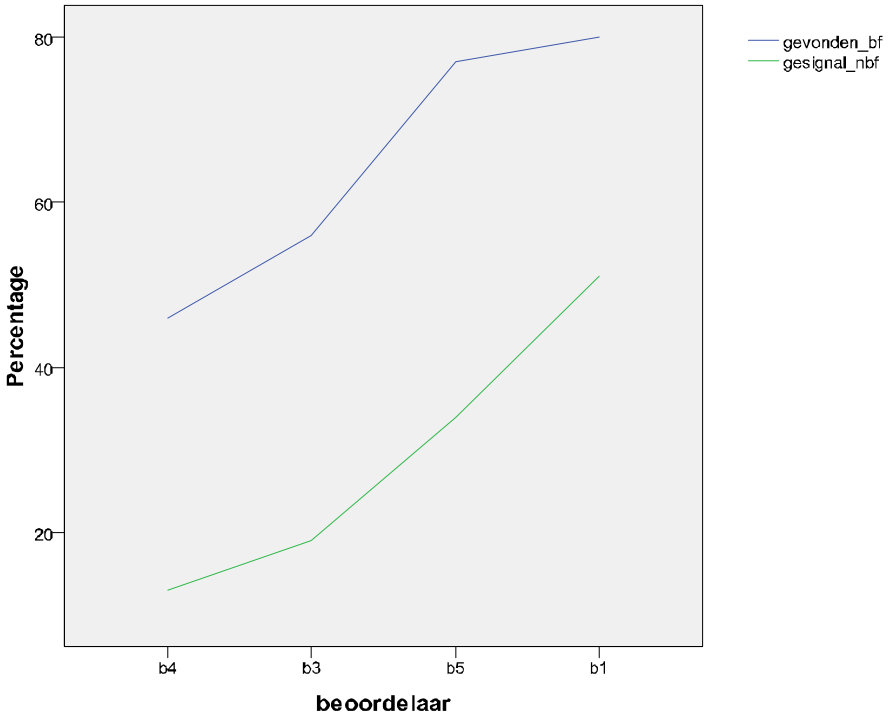
De beoordelaars gebruikten, ondanks de instructie, twee verschillende benaderingen. Dit verschil in benadering bleek geen invloed te hebben op de voorspellende waarde van het aantal bevestigde fouten dat een beoordelaar in een tekst signaleerde.

In de laatste kolom van Tabel 4.3 is de correlatie vermeld tussen het aantal fouten dat een beoordelaar signaleerde in een tekst (per honderd woorden) en dat wat de overige beoordelaars samen op dit punt signaleerden. Alle correlaties zijn uitermate hoog en b3 en b4 doen het niet slechter of beter dan b1 en b5. De door de beoordelaar gevolgde benadering had dus geen invloed op de waarde van het aantal door hem gesignaleerde, bevestigde fouten als maat voor basale tekstkwaliteit. De verklaring is vermoedelijk dat iedere beoordelaar zijn eigen strategie volgde, zodat deze steeds op dezelfde manier doorwerkte bij iedere tekst waardoor het verschil in beoordeling uiteindelijk geen invloed had op de rangordening van de teksten.

Hoewel de beoordelaars het per fout niet altijd eens waren, gold dit niet voor het aantal fouten per tekst (per honderd woorden) dat ze signaleerden. Het aantal fouten (per honderd woorden) dat iedere beoordelaar signaleerde in de dertig teksten, leverde een gemiddelde correlatie tussen de beoordelaars op van .85. Voor de vier beoordelaars samen leverde dit een beoordelaarsbetrouwbaarheid (coëfficiënt alfa) op van .95. Dit betekent dat vier andere expert-beoordelaars tot vrijwel dezelfde beoordeling (rangordening) van de 30 teksten zouden komen. Een 'slechte' tekst met veel fouten per honderd woorden volgens de ene beoordelaar was dus ook volgens een willekeurige andere expert-beoordelaar in dit opzicht een 'slechte' tekst. Tekstkwaliteit gedefinieerd als aantal fouten per honderd woorden viel door de expert-beoordelaars goed vast te stellen.

Voor het aantal bevestigde fouten (per honderd woorden) was de gemiddelde correlatie tussen de beoordelaars met .93 nog hoger en coëfficiënt alfa was zelfs .98. Door alleen bevestigde fouten te gebruiken, waren de beoordelaars het onderling nog iets meer eens dan wanneer alle gesignaleerde fouten werden gebruikt.

Figuur 4.1 Het verband tussen het percentage van alle bevestigde fouten dat een beoordelaar vond (gevonden_bf, bovenste lijn) en het percentage van zijn niet-bevestigde fouten (gesignal_nbf, onderste lijn)



Het voordeel van bevestigde fouten was dat ze eenvoudig interpreteerbaar waren. Iedere bevestigde fout was er één. Aan hun bestaan kan moeilijk getwijfeld worden, doordat ze door minstens twee onafhankelijk werkende beoordelaars waren gesignaleerd. Voor de rangordening van de teksten bood het werken met bevestigde fouten geen duidelijk voordeel: de aantallen van beide soorten fouten PHW correleerden zeer hoog (.93).

4.3.3 Correctie voor tekstlengte

De teksten verschilden aanzienlijk in lengte. De studenten hadden de instructie gekregen om een tekst te schrijven van 1 A4 (ongeveer 500 woorden), maar in werkelijkheid varieerde de lengte sterk. De kortste tekst telde 135 woorden en de langste 565 (de gemiddelde tekstlengte was 279 woorden met een standaarddeviatie van 104 woorden).

De lengte van de teksten (het aantal woorden) correleerde met het aantal fouten (r=.58,

p=.001, tweezijdig). Langere teksten bevatten in doorsnee meer fouten. Wanneer niet gecorrigeerd werd voor tekstlengte, zou een korte tekst beter zijn. Het lijkt echter duidelijk dat een lange tekst meer kans levert op een fout dan een korte tekst. Er moet dus rekening gehouden worden met de lengte van de tekst.

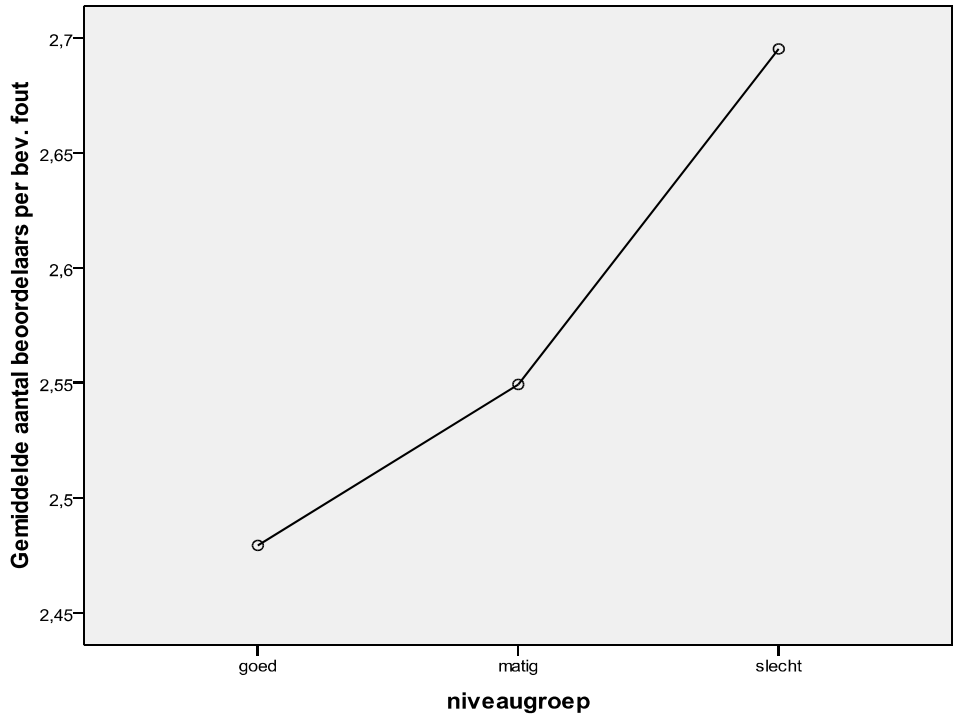
Na correctie voor de lengte van de teksten, door te werken met het aantal fouten per honderd woorden (PHW), bleek het aantal woorden dat de tekst telde significant negatief te correleren met het aantal fouten PHW (r= -0.65, p=0.000). Korte teksten bevatten in doorsnee juist meer fouten PHW. Door deze correctie sloegen de uitkomsten om. Wat eerst slechte teksten waren, omdat ze veel fouten bevatten, werden nu goede teksten, omdat ze weinig fouten PHW bevatten. Zonder correctie voor tekstlengte kan het aantal fouten in een tekst dus een misleidende maat zijn.

4.3.4 Tekstkwaliteit en zichtbaarheid bevestigde fouten

De tweede veronderstelling was dat fouten in slechte teksten (veel fouten PHW) slechter zichtbaar zouden zijn. De beoordelaar zou door de bomen het bos niet meer zien. Fouten in slechte teksten bleken echter zichtbaarder te zijn dan fouten in goede teksten. Voor de toetsing werden de teksten gerangordend naar het aantal bevestigde fouten PHW. Het aantal beoordelaars per fout bleek significant samen te hangen met de rang van de tekst. Dit gold voor gesignaleerde fouten (r=.23, p=.000, N=2400) en voor bevestigde fouten (r=.13, p=.000, N=989). De fouten in de slechte teksten werden door meer beoordelaars gesignaleerd dan de fouten in de goede teksten. De fouten in de slechte teksten vielen niet minder op, maar juist iets meer.

In Figuur 4.2 zijn de 30 teksten op basis van hun rangscore onderverdeeld in drie groepen teksten: goed, matig, slecht. Te zien valt dat vooral in de groep 'slecht', bevestigde fouten door iets meer beoordelaars gesignaleerd worden dan in de groepen 'goed' en 'matig'.

Figuur 4.2 Gemiddelde aantal beoordelaars voor de bevestigde fouten per niveaugroep teksten



4.3.5 Soort fouten

In totaal bleken er 76 bevestigde fouten te zijn die beoordeeld werden als niet-kloppend (zie bijlage 4). Van die fouten hadden 61 betrekking op het aaneenschrijven of los schrijven van woorden. Wanneer van de in totaal 989 bevestigde fouten alleen de kloppende werden genomen, bleven er 913 kloppende, bevestigde fouten over.

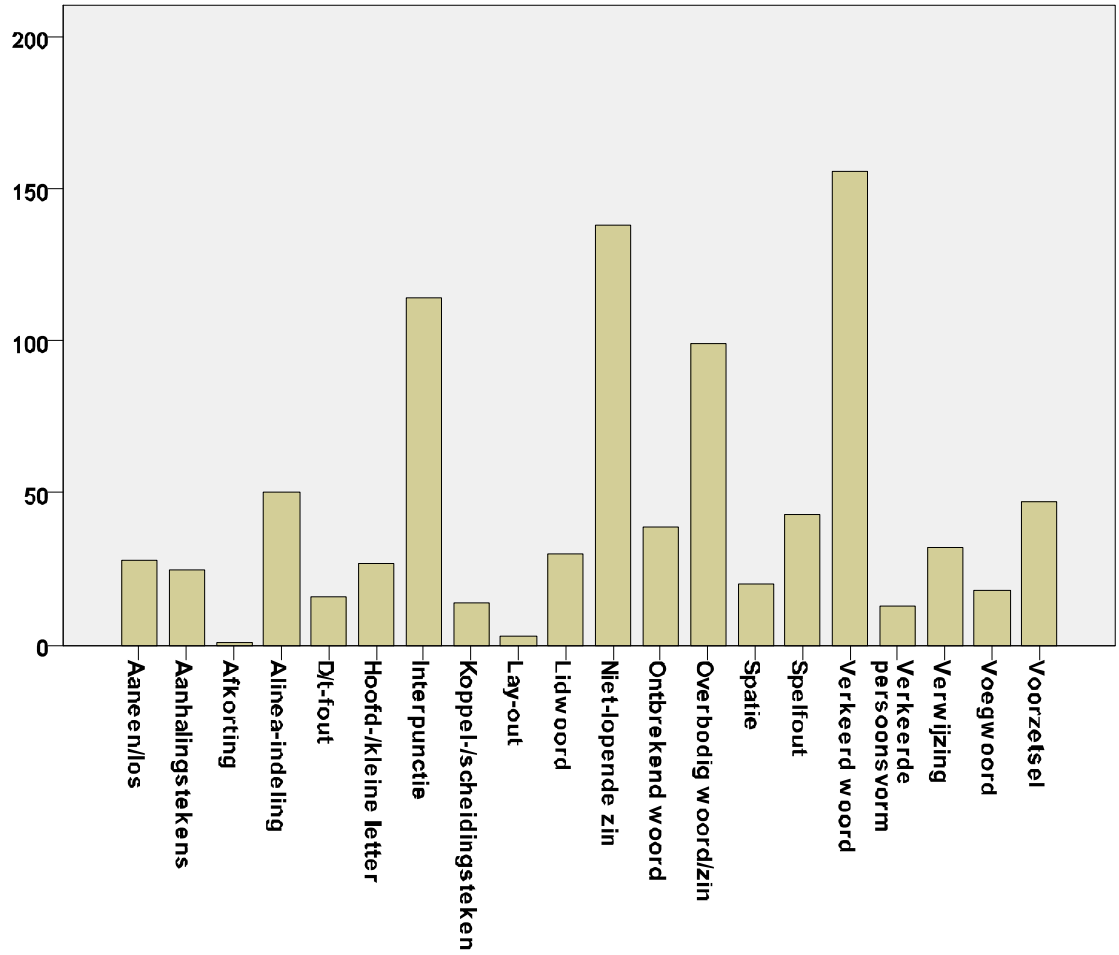
Tabel 4.4 geeft een overzicht van het aantal bevestigde, kloppende fouten en het percentage per foutcategorie. De tabel is geordend op basis van het aantal fouten per categorie. Uit de tabel blijkt dat de vier meest voorkomende categorieën, namelijk 'Verkeerd woord', 'Niet-lopende zin', 'Interpunctie', 'Overbodig woord/overbodige zin' samen meer dan de helft (55.5%) van alle bevestigde, kloppende fouten bevatten.

Tabel 4.4 Foutcategorieën gerangordend naar het aantal bevestigde, kloppende fouten

cat.nr.	aantal	perc.	cumm. perc.	omschrijving
1	156	17.1	17.1	Verkeerd woord
2	138	15.1	32.2	Niet-lopende zin
3	114	12.5	44.7	Interpunctie
4	99	10.8	55.5	Overbodig woord/overbodige zin
5	50	5.5	61.0	Alinea-indeling
6	47	5.1	66.2	Voorzetsel
7	43	4.7	70.9	Spelfout
8	39	4.3	75.1	Ontbrekend woord
9	32	3.5	78.6	Verwijzing
10	30	3.3	81.9	Lidwoord
11	28	3.1	85.0	Aaneen/los
12	27	3.0	88.0	Hoofdletter/kleine letter
13	25	2.7	90.7	Aanhalingstekens
14	20	2.2	92.9	Spatie
15	18	2.0	94.9	Voegwoord
16	16	1.8	96.6	D/t-fout
17	14	1.5	98.1	Koppel-/scheidingsteken
18	13	1.4	99.6	Verkeerde persoonsvorm
19	3	0.3	99.9	Lay-out
20	1	0.1	100.0	Afkorting
totaal	**913**	**100.0**		

De eerste acht foutcategorieën lijken erop te duiden dat op zijn minst een deel van de fouten niet nodig was geweest als de student de eigen tekst goed nagelezen had. Het gaat met name om de catgeorieën 'Niet-lopende zin', 'Overbodig woord/overbodige zin', 'Alinea-indeling' en 'Ontbrekend woord'. Kennelijk hebben studenten niet geleerd het eigen schrijfproduct kritisch na te lezen en te checken op fouten en gebreken.

Figuur 4.3 Het aantal bevestigde, kloppende fouten per foutcategorie (alfabetisch geordend)

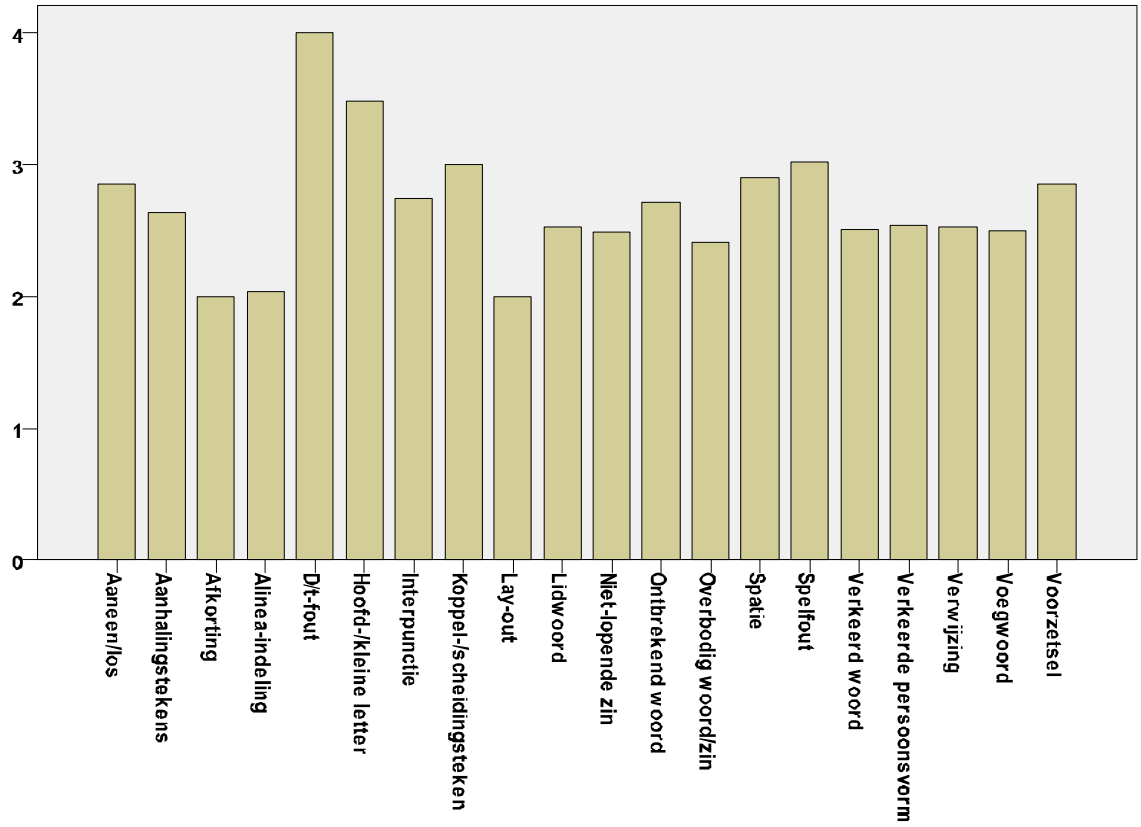


Figuur 4.3 geeft ook een overzicht van het aantal bevestigde, kloppende fouten per categorie, maar is alfabetisch geordend.

De derde verwachting van het foutenonderzoek was dat de 'D/t-fout' het meest zou voorkomen. Tabel 4.4 laat zien dat deze verwachting niet werd bevestigd: d/t-fouten blijken relatief weinig voor te komen (1.8%). Onder de gesignaleerde, niet-bevestigde fouten kwamen in het geheel geen d/t-fouten voor.

In Figuur 4.4 valt te zien dat de aanwezige d/t-fouten volstrekt zichtbaar waren voor de beoordelaars. Zodra een d/t-fout voorkwam in een tekst werd die door iedere beoordelaar gesignaleerd. Geen enkele foutcategorie scoorde qua zichtbaarheid zo hoog.

Figuur 4.4 Zichtbaarheid van de verschillende foutcategorieën (doordat gewerkt werd met bevestigde fouten zijn minimaal 2 beoordelaars nodig om fouten te 'zien')

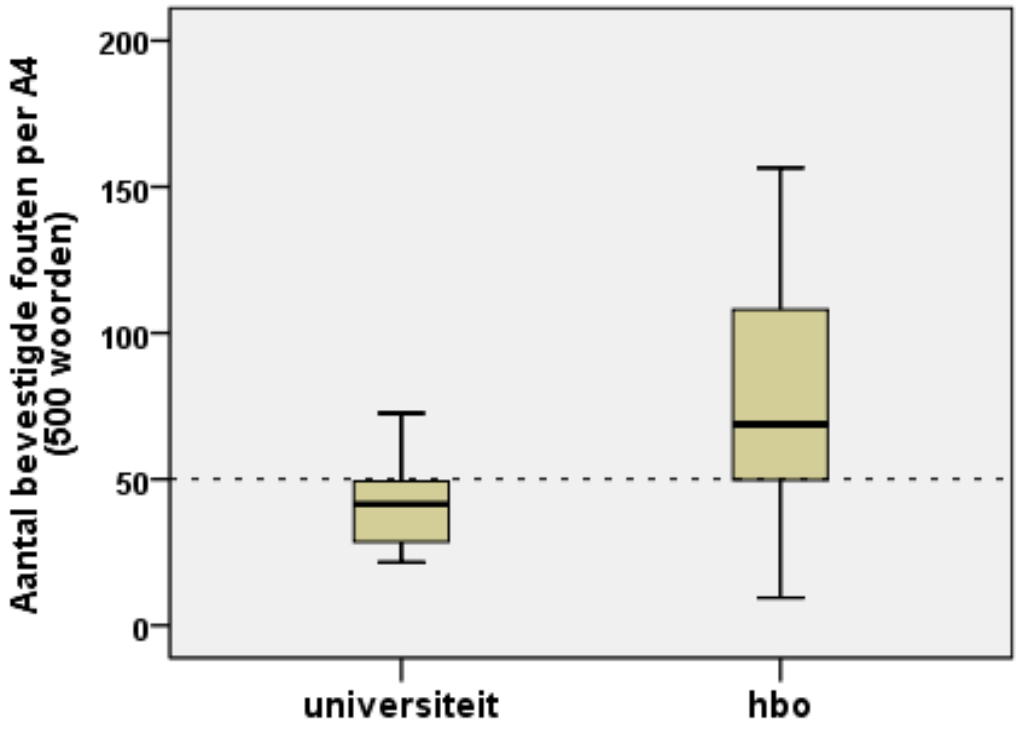


4.3.6 Uitkomsten hbo-studenten en universitaire studenten

Het aantal bevestigde fouten per honderd woorden bedroeg voor universitaire studenten gemiddeld 8,4 en voor hbo-studenten 16,1. Het aantal fouten per honderd woorden lag daarmee voor hbo-studenten ongeveer tweemaal zo hoog als voor universitaire studenten. Een t-toets onafhankelijke groepen leverde een significant verschil ($p<0.001$). Dit hield in dat een A4 (500 woorden) van de hbo-studenten gemiddeld 81 fouten bevatte (met een standaardfout van 9). Bij de universitaire studenten kwam dit neer op 42 fouten (met een standaardfout van 5).

Figuur 4.5 laat het verschil zien tussen de beide groepen in aantal bevestigde fouten per A4. Van de universitaire studenten scoorde ongeveer een vierde slechter dan 50 bevestigde fouten per A4; van de hbo-studenten drie vierde. Een vierde van alle hbo-studenten schreef zelfs met meer dan 100 bevestigde fouten per A4.

Figuur 4.5 Boxplots van het aantal bevestigde fouten per A4 (500 woorden) bij eerstejaars universitaire studenten en eerstejaars hbo-studenten



In Figuur 4.5 valt verder te zien dat de beste studenten (universiteit en hbo) rond de 25 bevestigde fouten per A4 scoorden. Dit komt overeen met 5 bevestigde fouten per honderd woorden of 1 bevestigde fout per 20 woorden. De slechtste universitaire studenten maakten ongeveer 75 fouten per A4. De slechtst schrijvende hbo-studenten maakten rond de 150 bevestigde fouten per A4. Dit komt overeen met 30 bevestigde fouten per honderd woorden of ruwweg 1 bevestigde fout per 3 woorden.

Uit figuur 4.5 blijkt ook dat de spreiding in het aantal bevestigde fouten PHW in het hbo zeer groot was. De slechtst schrijvende hbo-studenten maakten ongeveer 6 keer zoveel fouten PHW als de best schrijvende hbo-studenten. De slechtste universitaire studenten maakten ongeveer 3 keer zoveel bevestigde fouten PHW als de best schrijvende studenten. De standaarddeviatie voor het hbo bedroeg 8.2 tegen 3.2 voor de universiteit. Een Levene test op gelijkheid van varianties leverde p=0.004. Het verschil in spreiding is dus geen toeval. Tussen de hbo-studenten bestaan zeer grote verschillen in het niveau van de basale schrijfvaardigheid.

4.4 Conclusies en discussie

Om een kwantitatieve onderbouwing te krijgen voor de vele berichten over tekortschietende schrijfvaardigheid van studenten werden 20 teksten van eerstejaars hbo-studenten en 10 teksten van eerstejaars universitaire studenten voorgelegd aan vier beoordelaars die zelf goed konden schrijven. De beoordelaars kregen het verzoek alle fouten in de teksten aan te strepen en te omschrijven.

Gemiddeld per beoordelaar werd ongeveer een derde (35%) van de gesignaleerde fouten niet bevestigd door een andere beoordelaar. De eerste verwachting dat een gesignaleerde fout altijd door andere expert-beoordelaars bevestigd zou worden, bleek dus onjuist. Beoordelaars bleken verder, ondanks de instructie alle fouten te signaleren, gemiddeld 35% (van 20% tot 54%) van de bevestigde fouten niet te signaleren. Volgens de verwachting zouden fouten altijd zichtbaar zijn voor expert-beoordelaars. Ook dit deel van de verwachting bleek daarmee niet te kloppen. Deze uitkomsten relativeren het begrip 'fout'. Lang niet iedere gesignaleerde fout werd bevestigd door een andere beoordelaar, terwijl iedere beoordelaar behoorlijk wat bevestigde fouten miste.

Tegelijkertijd ondersteunen deze uitkomsten ook de bruikbaarheid van (aantallen) fouten om teksten te beoordelen. De meeste fouten werden immers wel bevestigd en de meeste bevestigde fouten werden wel gesignaleerd door de gemiddelde expert-beoordelaar. Beoordelaars bleken het onderling inderdaad zeer eens te zijn over het aantal fouten per honderd woorden (PHW) in de teksten met een gemiddelde onderlinge correlatie van .85. Uitgaande van alleen de bevestigde fouten PHW bleek de overeenstemming nog iets hoger te zijn met een gemiddelde onderlinge correlatie van .93.

Deze uitkomsten stemmen globaal overeen met de resultaten van het pilotonderzoek. Daar werden voor 10 teksten en 3 expert-beoordelaars gemiddelde onderlinge correlaties tussen de beoordelaars gevonden van .67 voor gesignaleerde fouten PHW en .90 voor bevestigde fouten PHW.

Teksten kunnen dus op basis van het aantal (bevestigde) fouten PHW gesignaleerd door enkele expert-beoordelaars betrouwbaar worden ingedeeld en gerangordend. Eventueel kan hiervoor zelfs een enkele expert-beoordelaar volstaan. Het probleem dat beoordelaars on-

derling weinig overeenstemmen in hun oordeel over de kwaliteit van teksten, lijkt daarmee opgelost te zijn, wanneer het aantal (bevestigde) fouten PHW gebruikt wordt als maat voor tekstkwaliteit.

Het voordeel van bevestigde fouten boven gesignaleerde fouten is dat het aantal bevestigde fouten PHW absoluut interpreteerbaar is: aan het bestaan van een bevestigde fout kan moeilijk getwijfeld worden. Het aantal gesignaleerde fouten PHW is echter ook afhankelijk van de desbetreffende beoordelaar: de ene beoordelaar signaleert meer fouten dan de andere. Voor de rangordening van de teksten maakt dit niet uit, maar de interpretatie van een bepaald aantal fouten PHW wordt hierdoor veel lastiger: schreven de studenten zo slecht of waren de beoordelaars zo kritisch? Bevestigde fouten PHW hebben dit probleem niet. Voor het overige maakt het in de praktijk weinig uit of men werkt met het aantal fouten PHW of met het aantal bevestigde fouten PHW, omdat beide maten hoog bleken te correleren (r=.93).

Lange teksten bleken meer fouten te bevatten dan korte teksten (r=.58), maar als gewerkt werd met het aantal fouten PHW bleken lange teksten juist minder fouten PHW te bevatten dan korte teksten (r= -.65). Ook in het pilotonderzoek werd een vergelijkbare correlatie gevonden (-.65). Rekening houden met de tekstlengte door te werken met het aantal fouten PHW kan dus leiden tot belangrijk andere uitkomsten dan zonder deze correctie.

Dit verband bevestigt echter ook de bruikbaarheid van het aantal fouten PHW als maat voor basale schrijfvaardigheid. De studenten die beter schrijven qua aantal fouten PHW blijken ook 'gemakkelijker' te schrijven.

De verwachting dat fouten in slechte teksten (veel fouten PHW) moeilijker zichtbaar zouden zijn, bleek onjuist. Zowel gesignaleerde als bevestigde fouten bleken in slechte teksten door (iets) meer beoordelaars gesignaleerd te worden.

De vijf meest voorkomende foutcategorieën waren 'Verkeerd woord', 'Niet-lopende zin', 'Interpunctie', 'Overbodig woord/overbodige zin' en 'Alinea-indeling'. Samen waren deze vijf categorieën goed voor meer dan 60% van alle bevestigde, kloppende fouten. Op basis van de fouten die gemaakt werden, lijkt het aannemelijk dat eerstejaarsstudenten hun geschreven tekst niet kritisch nalezen op fouten en gebreken. De lijst foutcategorieën is verder gebruikt voor de constructie van het nieuwe programma.

De verwachting dat de 'D/t-fout' het meest zou voorkomen, bleek niet te kloppen: d/t-fouten bleken relatief weinig voor te komen. D/t-fouten maakten minder dan 2% van het totaal uit. Hoewel ze weinig voorkwamen, bleken ze zeer zichtbaar te zijn en nooit aan de aandacht van een beoordelaar te ontsnappen. Beoordelaars zijn door hun opleiding kennelijk sterk gericht op het signaleren van d/t-fouten. Mogelijk geldt dit ook voor de als/dan-fout. Dit is een punt voor verder onderzoek.

Universitaire eerstejaarsstudenten maakten belangrijk minder bevestigde fouten PHW dan hbo-studenten. Universitaire studenten bleken in een A4 (500 woorden) gemiddeld 42 bevestigde fouten te maken, hbo-studenten 81. De beste studenten (hbo en universiteit) scoorden 25 fouten per A4 of nog iets minder. De slechtste studenten (hbo) scoorden per A4 125 bevestigde fouten of meer.

Deze uitkomsten bevestigen eerdere berichten over de tekortschietende schrijfvaardigheid van eerstejaarsstudenten. Door dit onderzoek krijgen die berichten echter een kwantitatieve onderbouwing die eenvoudig interpreteerbaar is. Op basis van de hoge aantallen bevestigde fouten die geconstateerd werden, lijkt de conclusie gerechtvaardigd dat het eerstejaarsstudenten ontbreekt aan basale schrijfvaardigheid. Hoewel eerstejaarsstudenten circa dertien jaar onderwijs hebben gevolgd met in totaal meer dan 13.300 lesuren, waarvan de kosten voor het de door de overheid gefinancierde deel per student meer dan 99.400 euro bedroegen (Elbers, 2011) was het gevolgde onderwijs met betrekking tot basale schrijfvaardigheid kennelijk weinig effectief.

Het foutenonderzoek was niet bedoeld een precieze schatting te leveren van het aantal bevestigde fouten bij eerstejaarsstudenten. Verschillende groepen zullen op dit punt altijd iets verschillen. Het doel was enerzijds een globale schatting te geven van het huidige niveau van basale schrijfvaardigheid bij eerstejaarsstudenten en anderzijds het begrip basale tekstkwaliteit te operationaliseren.

Hoewel de uitkomsten van dit onderzoek in grote lijnen overeenstemmen met de uitkomsten van het eerder verrichte pilotonderzoek, zou replicatie van het onderzoek de uitkomsten verder kunnen ondersteunen. Bij replicatie is de kwaliteit van de beoordelaars, het aantal beoordelaars en de instructie van groot belang. De kwaliteit van beoordelaars valt relatief eenvoudig te bepalen. Goede beoordelaars signaleren een hoog percentage van alle bevestigde fouten.

Beoordelaars misten gemiddeld relatief veel bevestigde fouten en voor een bevestigde fout waren minimaal twee beoordelaars nodig die de fout signaleerden, waardoor naar verwachting niet alle bevestigde fouten gevonden zijn die in de teksten aanwezig waren. Doordat gewerkt werd met vier beoordelaars van wie twee weinig bevestigde fouten misten, gaat het vermoedelijk slechts om een bescheiden percentage bevestigde fouten dat niet gevonden is. Dit betekent echter wel dat de gevonden aantallen bevestigde fouten PHW een onderschatting vormen van de werkelijke aantallen bevestigde fouten PHW die in de teksten aanwezig waren.

De constatering dat de beoordelaars zeer overeenstemden over de tekstkwaliteit gedefinieerd als het aantal fouten PHW, is gebaseerd op de gemiddelde onderlinge correlaties tussen de beoordelaars van .85. Dit gemiddelde is gebaseerd op zes correlaties die allemaal tussen 0.78 en 0.92 lagen. Verder werden deze uitkomsten bevestigd door de uitkomsten van het pilotonderzoek. Ook de beoordeling van de begin- en eindtekst op aantallen fouten in het te bespreken onderzoek naar het nieuwe schrijfvaardigheidsprogramma (hoofdstuk 6) leverde soortgelijke uitkomsten op. De hoge overeenstemming tussen expert-beoordelaars op het punt van aantallen fouten PHW lijkt dus niet uitzonderlijk, maar een algemeen verschijnsel.

De betrouwbaarheid van de indeling in foutcategorieën is niet onderzocht. Fouten konden soms in verschillende categorieën ondergebracht worden. Verder zullen verschillende beoordelaars verschillende categorieën definiëren. De uitkomsten over de aantallen fouten per categorie zijn daarom indicatief bedoeld.

Het belangrijkste resultaat van dit foutenonderzoek is dat expert-beoordelaars blijken overeen te stemmen over de kwaliteit van teksten (in termen van de onderlinge correlatie) wanneer het aantal (bevestigde) fouten PHW als maat voor tekstkwaliteit wordt gebruikt. Tekstkwaliteit (en daarmee basale schrijfvaardigheid) kan daardoor eenvoudig en betrouwbaar worden vastgesteld door eventueel zelfs een enkele expert-beoordelaar.

Het tweede belangrijke resultaat van het foutenonderzoek is dat Nederlandse eerstejaarsstudenten zeer grote aantallen bevestigde fouten PHW in hun Nederlandstalige schrijfproducten blijken te produceren. Kennelijk is het voorafgaande onderwijs op dit punt niet effectief geweest. Als we ervan uitgaan dat 10% van het voorafgaande onderwijstraject bestemd is om goed te leren schrijven, bedragen alleen al de kosten voor de overheid van dit niet-functionerende onderwijs voor de ruim 150.000 eerstejaarsstudenten in Nederland (hbo

en universiteit) 1,5 miljard euro per jaar. De feitelijke economische kosten van dit falende schrijfonderwijs (hoog opgeleide studenten die gedurende hun gehele loopbaan problemen hebben met schrijven) zullen echter vermoedelijk dit jaarlijkse bedrag nog vele malen overtreffen.

5

Deelstudie 2

Beoordeling taalmethodes

5.1 Inleiding

Voor het hoger beroepsonderwijs bestaat een groot aantal papieren en digitale methodes die beogen deficiënties op het gebied van schriftelijke taalvaardigheid weg te werken. In dit deel van het onderzoek stond de volgende vraag centraal (onderzoeksvraag 2): wat is de waarde van bestaande methodes om basale schrijfvaardigheid te verbeteren bij eerstejaars hbo-studenten? In totaal werden 17 papieren methodes en 9 digitale methodes verzameld en beoordeeld op geschiktheid.

In 3.1 is uiteengezet dat evaluatie op leerwinst onmisbaar is bij kwaliteitsverbetering van het onderwijs. Dit kan via zes stappen gebeuren. Eerst wordt een doel geformuleerd. Dit doel wordt omgezet naar een doelvariabele. Uitgaande van die doelvariabele worden twee gelijkwaardige toetsen ontwikkeld. Vervolgens vindt de voormeting plaats, nemen de studenten deel aan het programma en ten slotte volgt de nameting. De verbetering tussen nameting en voormeting vormt de leerwinst. Deze manier van beoordelen leidt tot een duidelijk en valide oordeel over een bepaalde methode. Voor bestaande taalmethodes is echter doorgaans geen empirische informatie beschikbaar waaruit blijkt of ze effectief zijn. Evaluatie van de beschikbare methodes op waargenomen leerwinst was daardoor niet mogelijk.

Een andere mogelijkheid om bestaande taalmethodes te evalueren is de mening over de methodes aan een panel van docenten te vragen. Deze benadering is gekozen bij een overzicht op het gebied van taalcursussen, dat is samengesteld in opdracht van het Ministerie van OCW (Bal, Berger, De Jonge, Oudmaijer & Tan, 2007). Het was de bedoeling om uit te zoeken of er geschikt materiaal was om pabo-studenten op taalgebied te remediëren. In totaal werden in dit overzicht zeven methodes genoemd, vier daarvan werden beoordeeld.

Iedere methode is eerst beschreven op een aantal punten, zoals doelgroep, materiaal (boek en/of software), mogelijkheden interactief leren, mogelijkheden zelfstandig leren, mogelijkheden om alleen bepaalde delen te behandelen en tot slot werden beknopt sterke en zwakke punten van de methode genoemd. Daarna volgde per methode een beoordeling door een docentenpanel. Het oordeel van het panel kwam er bijvoorbeeld op neer dat geschikte onderwerpen aan bod kwamen in de methode, dat de didactiek overzichtelijk was en dat de methode geschikt was voor zelfstudie. Impliciet bleef hierbij op basis van welke visie op onderwijs het panel tot dit oordeel was gekomen, ook was de methode waarmee het panel tot zijn conclusies kwam niet gestructureerd via een beoordelingsschema.

Het panel bestond uit ongeveer twintig docenten (het exacte aantal werd niet vermeld) die afkomstig waren uit het vo, ho en mbo. De beoordeling van de vier taalmethodes werd uitgevoerd tijdens een enkele bijeenkomst. Het verzoek aan de docenten was hierbij een beoordeling te geven met het oog op de eigen doelgroepen (Bal et al., 2007, p. 11). In de resultaten is het oordeel van de docenten samengevat per type onderwijs. Per beoordeelde methode zou men dan van ieder type onderwijs een beoordeling verwachten, in totaal dus twaalf beoordelingen (4 x 3). In werkelijkheid werden slechts vijf beoordelingen vermeld (p. 28-29). In totaal zeven beoordelingen werden zonder opgaaf van reden niet vermeld.

Het vellen van een oordeel over een methode zonder dat men dat baseerde op een expliciet onderwijskundig model bleek bij verder zoeken in de literatuur over het beoordelen van methodes en 'textbook evaluation' geen uitzondering te zijn, maar eerder regel. In tien publicaties op dit gebied (Al Fraidan, 2012; Ansary & Babaii, 2002; Driessen, Westhoff, Haenen & Brekelmans, 2008; Garinger, 2002; Miekley, 2005; Mukundan, Hajimohammadi & Nimehchisalem, 2011; Raseks, Esmae'li, Ghavamnia & Rajabi, 2010; Sheldon, 1988; Wang, 2006; Williams, 1983) werd slechts eenmaal, alleen door Driessen et al. (2008), een expliciet onderwijskundig model vermeld als uitgangspunt voor het beoordelingsschema.

Het vermelde model bestond uit vijf onderwijskundige principes die werden benoemd als de 'SLA penta-pie' (Driessen et al., 2008, p. 809). Strikt genomen vormen deze vijf principes zoals geformuleerd niet een algemeen onderwijskundig model, maar alleen een model voor onderwijs op het gebied van SLA (Second Language Acquisition). Interessant aan het model is dat hoewel het een totaal andere achtergrond heeft dan het ABC-model, namelijk een cognitief-psychologische, van de principes één het belang van de input benadrukt (de A in het ABC-model) en dat maar liefst drie van de vijf principes betrekking hebben op het gedrag van de student (de B in het ABC-model). Het eerste van deze drie principes stelt dat de student iets met de input moet doen, het tweede dat de vorm van de output belangrijk is en het derde pleit voor 'pushed output' (p. 808): studenten moeten ervaring opdoen met het produceren van output. Het belang van feedback is niet opgenomen in het penta-pie-model, in plaats daarvan benadrukte het vijfde principe het belang van strategieën. De beoordelingsschema's van beide modellen focussen verder volledig op de leertaak van de student.

In een van de andere publicaties stelden de auteurs eerst, via een citaat, dat opstellers van beoordelingsschema's op de hoogte moeten zijn van de relevante theorieën. Daarna stelden ze dat hun evaluatieschema gebaseerd was op een review van soortgelijke instrumenten om

de 'construct validiteit' te verzekeren (Mukundan, et al., 2011, p. 23). Enerzijds is deze invulling van het begrip 'constructvaliditeit' nogal verwarrend, anderzijds levert het baseren van een evaluatieschema op voorgaande schema's geen garantie op voor criteriumvaliditeit. Het criterium is dan de grootte van de gerealiseerde leerwinst bij toepassing van de methode in de praktijk.

Ansary & Babaii (2002) stelden aan het begin van hun artikel op zoek te zijn naar 'theory-neutral' criteria. De veronderstelling is kennelijk dat uitgaan van een onderwijskundige theorie docenten kan afschrikken en dat naarmate meer docenten de gebruikte criteria onderschrijven, de criteria ook meer valide zullen zijn. Dat veel mensen een bepaald criterium belangrijk vinden, wil echter nog niet zeggen dat het een goede voorspeller zal zijn van de gerealiseerde leerwinst.

Een tweetal publicaties bevatte niet het gebruikte beoordelingsschema (Wang, 2006; Fraidan, 2012). Slechts de helft van de tien publicaties werkte met gekwantificeerde beoordelingen (Garinger, 2002; Miekley, 2005; Mukundan et al., 2011; Sheldon, 1988; Williams, 1983). Het probleem met niet-gekwantificeerde beoordelingen is dat het samennemen van een aantal beoordelingen voor een eindoordeel uiterst moeilijk wordt en een grote mate van subjectiviteit in het eindoordeel kan introduceren.

Williams (1983) gebruikte in zijn beoordelingsschema 28 beoordelingspunten, maar biedt de beoordelaar vervolgens de ruimte die naar eigen inzicht te wegen, waardoor in feite voor de beoordelaar de optie ontstaat alleen die punten uit het beoordelingsschema te gebruiken die hij zelf wenselijk vindt.

Het aantal te beoordelen punten in de beoordelingsschema's varieerde van 42 tot 15. Veel beoordelingsschema's waren niet algemeen, doordat ze specifiek aandacht vroegen voor de dekking van specifieke vaardigheden en onderwerpen. Zo gebruikte Williams (1983) ondanks de algemene strekking van de titel van zijn publicatie, 'Developing criteria for textbook evaluation', als hoofdcategorieën: General, Speech, Grammar, Vocabulary, Reading, Writing, Technical (p. 253). Garinger (2002) hanteerde een korte en relatief duidelijke checklist, waarin ook praktische vragen voorkomen: 'Is the textbook available?', 'Can the textbook be obtained in a timely manner?' en 'Is the textbook cost-effective?' De te beoordelen punten in de beoordelingsschema's kunnen soms vrij vaag zijn. Zo is bij Williams (1983) een te beoordelen punt: 'takes into account currently accepted methods of ESL/EFL teaching' (p. 255).

Van de 10 publicaties waren er slechts 2 die daadwerkelijk een analyse van een of meer methodes ondernamen. In het ene geval werd één hoofdstuk van een studieboek beoordeeld zonder daarbij overigens een expliciet beoordelingsschema te hanteren (Wang, 2006). In het andere geval werden 4 methodes beoordeeld op 22 punten waarbij een expliciet schema werd gebruikt, maar waarbij de antwoorden niet gekwantificeerd werden (Raseks et al., 2010). Bij Wang ging het om een enkele beoordelaar die de methode beoordeelde. Bij Raseks et al. bestaat ook de mogelijkheid dat alle auteurs samen steeds alle methodes beoordeeld hebben of dat iedere methode beoordeeld is door één auteur; dit wordt niet vermeld.

Bij de beoordeling van de papieren en digitale methodes in dit onderzoek is er voor gekozen alle beoordelingen te structureren via een beoordelingsschema (Van Es, 1985) dat expliciet gebaseerd was op een onderwijskundig model: het ABC-leermodel. Verder ging dit beoordelingsschema uit van gekwantificeerde oordelen, zodat de oordelen per hoofdcategorie en voor het totaal eenvoudig gemiddeld konden worden. Alle methodes werden beoordeeld door dezelfde beoordelaar. Idealiter zouden meerdere beoordelaars alle methodes hebben beoordeeld, zodat de betrouwbaarheid van de beoordeling empirisch viel vast te stellen. In verband met de tijd en de kosten is hiervan afgezien. De verwachte meeropbrengst was bovendien gering, met het oog op het doel van deze deelstudie. In plaats daarvan zijn twee methodes achteraf nogmaals beoordeeld door de mede-auteur. Op het ABC-model is in paragraaf 3.3 dieper ingegaan. In 5.2 wordt uitgebreider ingegaan op het beoordelingsschema.

Een lastige vraag is, hoe betrouwbaar de beoordeling is uitgaande van het beoordelingsschema. In de tien onderzochte publicaties werd niet één keer getracht op dit punt informatie te verzamelen. Om enig zicht op de betrouwbaarheid van de beoordeling te krijgen, is -- zoals hiervoor reeds werd opgemerkt -- mijn mede-auteur (door omstandigheden geruime tijd nadat de eerste beoordeling afgerond was) gevraagd de twee best beoordeelde methodes (voor zover de methodes nog verkrijgbaar waren) opnieuw te beoordelen (zie 5.4).

Voorafgaand aan de beoordeling waren er bepaalde verwachtingen over de uitkomsten. Deze verwachtingen hadden niet zozeer een theoretische basis, maar waren gebaseerd op eigen inzichten en ervaring met verschillende soorten methodes. Zo was de verwachting dat papieren methodes goed zouden scoren qua leerstof en als naslagwerk, terwijl gedacht werd dat digitale methodes beter zouden scoren op het punt van oefenen en feedback. Juist op dit punt leken digitale methodes meer mogelijkheden te bieden dan papieren methodes.

5.2 Methode

Beoordelingsschema studieteksten

De verschillende methodes werden steeds door dezelfde beoordelaar (de onderzoeker in dit geval) beoordeeld op dezelfde criteria. Hiervoor werd uitgegaan van een beoordelingsschema voor studieteksten (Van Es, 1985) dat gebaseerd was op het ABC-model voor effectief leren.

Studieteksten bevatten informatie die de student moet weten, moet begrijpen en moet kunnen toepassen. Uitgaande van het ABC-model moeten vaardigheden ingeoefend worden. Dit betekent dat over de leerstof vragen moeten worden beantwoord. Verder moet de student zijn antwoorden kunnen checken en gemotiveerd worden om door te gaan. Een effectieve studietekst bevat daardoor idealiter drie verschillende delen: de leerstof (de informatiebasis), de vragen die de student daarover moet kunnen beantwoorden (het oefenboek) en feedback (het feedbackmiddel) waarmee de student de eigen antwoorden kan checken.

Het beoordelingsschema gaf voor ieder van deze drie componenten een aantal criteria. Deze criteria komen ook voor bij sturende onderwijsmodellen, zoals het belang van een duidelijke doelstelling, heldere uitleg, veel en relevant oefenmateriaal dat geordend is op grond van de moeilijkheidsgraad en snelle feedback. Het schema is weergegeven in Tabel 5.1. Hoewel een oordeel altijd subjectieve elementen bevat, werd via het schema getracht de beoordeling te systematiseren.

Tijdens de beoordeling bleek het schema op punten onvolledig en het is daarom aangepast (zie Tabel 5.2). In de eerste plaats werden twee criteria aan het schema toegevoegd onder feedbackmiddel: uitleg en voortgangsinformatie. Deze punten ontbraken in het oorspronkelijke schema en leken van belang bij het beoordelen van de feedbackmogelijkheden van een programma. Verder is de omschrijving van sommige criteria uitgebreid; met name de vele mogelijkheden van digitale programma's maakten dit nodig.

Tabel 5.1 Oorspronkelijk schema voor beoordelen van studieteksten

---

Als informatiebasis

- Juistheid: met vakkennis te beoordelen
- Volledigheid: voldoende informatie met het oog op de opdrachten
- Duidelijkheid: geen ingewikkelde formuleringen in de informatie
- Relevantie: geen overbodige informatie met het oog op de doelstelling
- Toegankelijkheid: snelheid waarmee de informatie te vinden is

Als oefenboek

- Veelheid: veel of weinig oefeningen
- Gemakkelijkheid: veel of weinig tijd nodig voor het maken van de oefening
- Relevantie: oefeningen die oefenen wat nodig is, met het oog op de doelstelling
- Volledigheid: oefeningen die oefenen met alle dingen die nodig zijn, gelet op de doelstelling
- Geordendheid: oefeningen die geordend zijn op grond van moeilijkheidsgraad

Als feedbackmiddel

- Veelheid: feedback bij iedere opdracht
- Betrouwbaarheid: feedback die klopt
- Duidelijkheid: feedback die begrijpelijk is voor de student en duidelijk maakt of de opdracht goed gemaakt is
- Snelheid: feedback die snel laat zien of opdracht goed gemaakt is
- Afhankelijkheid: feedback die pas wordt gegeven, nadat het antwoord gegeven is

---

Selectie taalmethodes

Het was met het oog op de tijd niet mogelijk om alle bestaande methodes die gericht zijn op het verbeteren van basale taalvaardigheden bij beginnende hbo-studenten te onderzoeken.

Voor de selectie van de methodes is uitgegaan van:
- de methodes die genoemd werden tijdens hanzebreed overleg op de Hanzehogeschool Groningen in 2008 met collega's van andere Schools om de taalvaardigheid van studenten te vergroten;
- de methodes die ter sprake kwamen bij het Nederlands/Vlaams Platform Taalbeleid Hoger Onderwijs in 2008;
- de methodes die de Taalwinkel van de Universiteit en de Hogeschool van Amsterdam heeft aanbevolen voor studenten met taalproblemen;
- de methodes die in maart 2008 aanwezig waren in de mediatheek van de Hanzehogeschool Groningen;
- de methodes gericht op het hbo die door onderwijsuitgevers in het begin van 2008 werden aangeboden.

In de uiteindelijke selectie zijn niet opgenomen:
- methodes die dateerden van voor 1990 en daarna niet meer opnieuw waren verschenen;
- methodes die voor het eerst verschenen na 1 juli 2008.

Het aantal methodes is verder beperkt door naslagwerken niet in de selectie te betrekken. Zo viel de *Schrijfwijzer* van Renkema (2005) buiten de beoordeling, evenals *Vraagbaak Nederlands* (Tiggeler, 2007), omdat die niet bedoeld zijn om als taalmethode aan studenten voor te leggen om beter te leren schrijven. Deze boeken zijn uitsluitend bedoeld om vragen te beantwoorden over taalkwesties die zich bij het schrijven voordoen.

Ten slotte is als eis gesteld dat de methode zich in ieder geval moest richten op vormfouten. Het gaat in dit onderzoek immers om de aanpak van basale schrijfvaardigheid. Om die reden zijn niet alleen spellingmethodes beoordeeld, maar ook methodes die over stijlaspecten gaan waarin elementaire vormfouten behandeld werden. Aan de andere kant zijn daardoor methodes afgevallen waarin rapporteren behandeld werd of het schrijven van brieven.

Tabel 5.2 Aangepast schema voor beoordelen van studieteksten

Informatiebasis

- Juistheid: met vakkennis te beoordelen
- Volledigheid: voldoende informatie, met het oog op de opdrachten. Volledige dekking stof
- Duidelijkheid: geen ingewikkelde formuleringen in de informatie
- Relevantie: geen overbodige informatie, met het oog op de doelstelling of eventueel uitleg.Geen uitleg die al weer vergeten kan zijn tijdens het maken van de oefeningen. Bij digitale programma's: specifieke uitleg die geraadpleegd kan worden tijdens het oefenen. Uitleg die niet genegeerd kan worden
- Toegankelijkheid: snelheid waarmee de informatie te vinden is door overzichtelijke indeling. Bij digitale programma's: snelheid waarmee specifieke informatie te vinden is tijdens het oefenen

Oefenboek

- Veelheid: veel of weinig oefeningen. Alleen maar simpele of alleen maar moeilijke oefeningen, dan te weinig oefeningen
- Gemakkelijkheid: veel of weinig tijd nodig voor het maken van de oefening Uitsluitend moeilijke opdrachten kosten te veel tijd. Bij digitale programma's: rustig scherm met steeds 1 item per keer en geen animaties is gemakkelijker. Keuze moet ook gemakkelijk te veranderen zijn
- Relevantie: oefeningen die oefenen wat nodig is, met het oog op de doelstelling of eventueel uitleg
- Volledigheid: oefeningen die oefenen met alle dingen die nodig zijn, gelet op de doelstelling of eventueel uitleg
- Geordendheid: oefeningen die geordend zijn op grond van moeilijkheidsgraad

Feedbackmiddel

- Veelheid: feedback bij iedere opdracht
- Betrouwbaarheid: feedback die klopt
- Duidelijkheid: feedback die zonder meer begrijpelijk is voor de student en in een oogopslag duidelijk maakt of het item goed gemaakt is
- Snelheid: feedback die snel komt, na ieder gemaakt item
- Afhankelijkheid: feedback die pas wordt gegeven, nadat het antwoord gegeven is
- Uitleg: feedback die uitleg geeft bij fout antwoord
- Voortgangsinformatie: feedback die aangeeft wanneer de student de stof voldoende beheerst en wat hij moet ondernemen als dat niet het geval is. Bij digitale programma's: feedback die aangeeft of de student weet hoeveel hij nog moet doen van een bepaalde opdracht, hoeveel fouten gemaakt worden in de opdracht en welke opdrachten nog volgen. Feedback over de resultaten van de student bij alle gemaakte opdrachten

Methodes die zich wel richten op elementaire taalfouten, maar geen systematische oefening bieden, zijn evenmin beoordeeld. Een voorbeeld hiervan is het digitale programma TiO (Taalonderwijs in Ontwikkeling). De student die met TiO werkt, krijgt bij zijn eigen tekst algemene opmerkingen, maar het programma bevat zelf geen oefenmateriaal om de basis-taalvaardigheden op een bepaald niveau te brengen.

Deze selectie heeft geresulteerd in de zeventien papieren methodes en negen digitale methodes die in Tabel 5.3 vermeld zijn.

Waardetoekenning aan criteria

Aan de hand van de criteria uit het schema zijn de methodes beoordeeld door de onderzoeker. In eerste instantie heeft dit geresulteerd in een beschrijving. Vervolgens is aan ieder criterium een score toegekend van minimaal 1 ('zeer slecht') en maximaal 5 ('uitstekend'). De waarde 0 werd gegeven als een bepaald deel, bijvoorbeeld de uitleg of feedback, volledig ontbrak. De waarde 0 werd beschouwd als een valide waarde. Het volledig ontbreken van een bepaald onderdeel was een negatief punt dat meetelde in de beoordeling.

Een bepaalde score werd toegekend op basis van een bepaalde beschrijving. Een 1 of een 5 werd alleen aan een criterium gegeven als uit de beschrijving bleek dat de methode op dit punt opvallend slecht of goed was. Een 2 werd toegekend bij een beschrijving waaruit bleek dat de methode op dit punt slecht was. Een 3 werd gegeven bij een beschrijving die duidelijk maakte dat de methode op dit punt matig of redelijk was. Een 4 werd toegekend bij een beschrijving waaruit bleek dat de methode op dat punt goed was.

Bij de papieren methodes kon op grond van de bovenstaande redenering een bepaalde score aan een criterium gegeven worden. Dezelfde weg moest in beginsel ook gevolgd worden bij de digitale programma's. Deze zaten wat structuur en gebruiksmogelijkheden betrof echter anders in elkaar en dat had gevolgen voor de invulling van bepaalde criteria, zoals ook te zien is in Tabel 5.2. Bij informatiebasis zijn de criteria volledigheid, relevantie en toegankelijkheid uitgebreid. Bij oefenprogramma was dat het geval bij het criterium gemakkelijkheid en bij feedbackmiddel zijn drie criteria uitgebreid, namelijk duidelijkheid, snelheid en voortgangsinformatie. Aan de uitbreiding van deze punten zijn bepaalde scores gehecht.

Tabel 5.3 Overzicht beoordeelde taalmethodes

Papieren methodes

1. Bertina, M. (2006). *Gewoon goed Nederlands.* Amsterdam: Boom Onderwijs
2. Bout, M. & Bruijn, H. de (2007). *Basisvaardigheden Spelling voor de pabo*. Groningen: Wolters-Noordhoff
3. Braas, C. & Pas, L. van der (2006). *Taaltopics Spelling* (4e druk). Groningen: Wolters-Noordhoff
4. Braas, C. & Krijgsman, J. (2005). *Taaltopics Formuleren* (2e druk). Groningen: Wolters-Noordhoff
5. Daniëls, W. (2006). *Wolters' Nederlands in je pocket* (herziene druk). Groningen: Wolters-Noordhoff
6. Dijkstra, B.A. & Delden, J. van (1996). *Repetitieboekje Nederlands* (5e druk). Groningen: Wolters-Noordhoff
7. Hogen, R. van (1997). *Praktische cursus Formuleren* (2e druk). Groningen: Wolters-Noordhoff
8. Hogen, R. van & Rietstap, E. (2007). *Basisvaardigheden Taal*. Groningen: Wolters-Noordhoff
9. Hogeweg, R. (2003). *Dat d/t gedoe.* Groningen: Wolters-Noordhoff
10. Kas, W. (1996). *Spelbewust* (2e druk). Zutphen: Thieme
11. Klein, M. & Visscher, M. (2006). *Praktische cursus spelling* (5e druk). Groningen: Wolters-Noordhoff
12. Mante, J. (2006). *Een Goede Spelling*. Utrecht: ThiemeMeulenhoff
13. Moons, A., Bovenhoff, M. & Latjes, G. (2008). *Basisboek Spelling*. Groningen: Wolters-Noordhoff
14. Onrust, M., Verhagen, A. & Doeve, R. (1999). *Formuleren*. Houten: Bohn Stafleu Van Loghum
15. Pak, D. (2007). *Vlekkeloos Nederlands. Spelling en stijl compleet* (2e druk). Den Haag: Dick Pak
16. Schilder, J. (2008). *Van verslag tot rapport.* Amsterdam: Boom Onderwijs
17. Westen, W. van der (2005). *Welgespeld.* Bussum: Coutinho

Digitale methodes

1. *Cambiumned* (oktober 2008). cambiumned.nl
2. *dtkompas* (september 2008). dtkompas.nl
3. *Hogeschooltaal* (april 2008). Deventer: Kluwer
4. *Juf Melis* (oktober 2008). jufmelis.nl
5. *Muiswerk* (september 2008). Uithoorn: Muiswerk Educatief
6. *Nedercom* (augustus 2008). Roden: Nedercom Eduware
7. *Project X 2002* (oktober 2008). projectx2002.org
8. *Studiemeter* (september 2008). Amersfoort: Deviant
9. *TaalONLINE* (september 2008). Den Haag: Jager & Neyndorff

Onder informatiebasis moest bij het criterium volledigheid zonder meer een 2 gegeven worden als een onderwerp ontbrak dat wel binnen het bereik van het programma viel. Bij het criterium relevantie was het bij digitale programma's doorslaggevend of de uitleg tijdens het oefenen opgeroepen kon worden. Wanneer dat niet het geval was, werd een 1 gegeven. Als de uitleg gescheiden van de oefeningen gepresenteerd werd en die kon niet opgeroepen worden, dan is de waarde ervan heel beperkt. Verder was van belang of de uitleg specifiek was. Iemand die bij het oefenen specifiek iets wil weten, heeft geen belang bij een aantal schermen informatie waar hij niets aan heeft. Daarom moest een 3 toegekend worden als uitleg die geraadpleegd werd, niet specifiek was. Een 4 werd toegekend als wel specifieke uitleg geraadpleegd kon worden, maar een 5 was alleen bestemd voor gevallen waarin de uitleg een essentiële rol speelde bij het oefenen en niet genegeerd kon worden. Bij het criterium toegankelijkheid moest een 3 gegeven worden als uitleg op zich snel te vinden was, maar naar specifieke uitleg gezocht moest worden. Een 5 werd alleen toegekend bij specifieke uitleg die snel te vinden was.

Onder oefenprogramma is het criterium gemakkelijkheid uitgebreid. Oefeningen waarbij alleen iets aangeklikt hoefde te worden, gingen heel snel. Maar andere dingen konden die snelheid weer beperken, zoals een onrustig scherm door alle oefenzinnen in één keer te presenteren of het gebruik van animaties. Ook als het niet mogelijk was om een eenmaal gemaakte keuze te veranderen, beïnvloedde dat de snelheid. Als van een of meer van deze beperkingen sprake was, werd een 3 toegekend.

Onder feedbackmiddel werd bij duidelijkheid alleen een 5 toegekend als de feedback in een oogopslag duidelijk maakte hoe het item gemaakt was. Een 3 werd gegeven als gekozen kon worden tussen feedback bij ieder gemaakt item of bij een volledig gemaakte opdracht en een 2 werd toegekend als alleen de mogelijkheid bestond om feedback te krijgen bij een volledige opdracht. Deze waarden werden gegeven op grond van de redenering dat het niet mogelijk was om meteen te zien hoe een bepaald item gemaakt was, als alle items op het scherm bij langsgegaan moesten worden. Duidelijk in een oogopslag slaat dus op degene die leert, de student. Bij het criterium snelheid, slaat snel op de computer en ging het erom hoe snel het scherm met feedback kwam. Bij snelheid werd alleen een 5 toegekend als de feedback meteen kwam na ieder item. Een 1 werd gegeven bij feedback na een volledige opdracht. Als het mogelijk was om te kiezen tussen feedback na ieder item of na een hele opdracht, moest een 3 toegekend worden. Bij voortgangsinformatie werd maximaal een 2 gegeven als toetsen ontbraken of als het bij opdrachten niet duidelijk was hoe die gemaakt

waren. Van een 5 kon alleen sprake zijn als de student echt door het programma gestuurd werd, totdat hij de stof meester was.

Op grond van deze toekenning van scores kan het lijken alsof digitale programma's strenger beoordeeld werden dan papieren methodes. Maar dit verschil werd met name veroorzaakt, doordat digitale programma's anders werken dan papieren methodes.

5.3 Resultaten

Eerst zijn de papieren methodes beoordeeld en vervolgens de digitale methodes.

5.3.1 Beoordeling papieren methodes

De papieren methodes zijn in de onderstaande tekst en in de bijlage alfabetisch gerangschikt op de achternaam van de auteur. In nagenoeg alle methodes staat aangegeven dat ze geschikt zijn voor zelfstudie. Op grond van de beoordeling als informatiebasis, oefenboek en feedbackmiddel is daarom bij iedere methode aangegeven in hoeverre dat het geval was. De beoordeling van alle papieren methodes aan de hand van het beoordelingsschema voor studieteksten staat in bijlage 5. Op basis daarvan is het onderstaande oordeel over de papieren methodes tot stand gekomen.

Gewoon goed Nederlands

Gewoon goed Nederlands gaat over de basisprincipes van het geschreven Nederlands, met als onderwerpen interpunctie, werkwoorden, eenvoudige en samengestelde zinnen en genre en stijl**.** Deze methode is bedoeld voor hbo-studenten en universitaire studenten. Over het niveau van de doelgroep wordt verder niets gezegd. Het doel is dat studenten in staat zijn om teksten te schrijven in helder en correct Nederlands. Het boek is ook bedoeld voor zelfstudie. Bij dit onderzoek gaat het om elementaire taalvaardigheden en daarom bleven buiten beschouwing de enkelvoudige en samengestelde zinnen en genre en stijl, waarbij gekeken wordt naar het verschil tussen fictie en nonfictie. De onderdelen werkwoorden en interpunctie werden beoordeeld.

Als informatiebasis is *Gewoon goed Nederlands* te gebruiken door studenten met een behoorlijke kennis van de Nederlandse grammatica en enige schrijfervaring. Voor deze groep wordt redelijk wat uitleg gegeven over met name het effect van werkwoorden en interpunctie op de lezer. Correct Nederlands speelt op de achtergrond wel steeds een rol. Maar voor de gemiddelde hbo-student die aarzelt bij het correct schrijven van werkwoorden, is deze informatie te hoog gegrepen.

Als oefenboek is het niet geschikt. In deze methode zijn in de eerste plaats niet veel oefeningen opgenomen. Aan het einde van ieder hoofdstuk staan wel een paar globale oefeningen die verband houden met de uitleg, maar die oefenen niet consequent met alles wat in de theorie behandeld is. Het kost ook relatief veel tijd om de oefeningen te maken, omdat de oefeningen vrij lastig zijn en er niets voorgestructureerd is. Er is verder geen sprake van een opbouw in de moeilijkheidsgraad. Voor de opdrachten is meteen een bepaald niveau vereist.

Als feedbackmiddel is de methode ongeschikt. Er wordt geen feedback gegeven, zodat de student na het maken van de opdrachten niet kan nagaan hoe hij de opdrachten gemaakt heeft. Hij is daarvoor afhankelijk van de docent, die zelf ook eerst de opdrachten zal moeten maken om feedback te kunnen geven. Er zijn geen antwoorden beschikbaar.

Dit boek kan mogelijk zelfstandig door de student gebruikt worden als een soort naslagwerk, maar voor zelfstudie om de schriftelijke taalvaardigheid te verbeteren is deze methode ongeschikt.

Basisvaardigheden Spelling voor de pabo

Basisvaardigheden spelling voor de pabo gaat over spelling van werkwoorden, spelling in het algemeen van Nederlandse woorden en interpunctie. Deze methode is bestemd voor studenten in het hoger beroepsonderwijs, met name pabo-studenten. Over het beginniveau wordt niets gezegd. Het doel is dat pabo-studenten zich kunnen voorbereiden op de officiële taaltoets die ze in hun eerste jaar moeten halen. De stof kan zelfstandig worden doorgenomen, maar is ook te gebruiken als lesmateriaal. Bij dit boek hoort een cd-rom met oefeningen en toetsen. Op de bijbehorende website kunnen studenten met behulp van een instaptoets vaststellen hoe hun niveau is. Deze instaptoets is overigens identiek aan één van de instaptoetsen op cd-rom. Hoewel deze methode zich richt op pabo-studenten, kan die in

beginsel evengoed gebruikt worden door andere eerstejaars hbo-studenten om de basistaalvaardigheden op niveau te krijgen. Hier werd gekeken naar de spelling van werkwoorden en interpunctie.

Als informatiebasis is *Basisvaardigheden Spelling voor de pabo* goed bruikbaar, maar de student heeft wel enige kennis nodig van de Nederlandse spelling en van grammaticale begrippen. De uitleg is relevant voor de opdrachten. In het boek staan de oefeningen meteen bij ieder onderdeeltje dat behandeld wordt, zodat er een duidelijk verband is tussen theorie en praktijk.

Als oefenboek is het behoorlijk geschikt. In totaal zijn redelijk veel oefeningen beschikbaar. De oefeningen zijn snel te maken. Maar met interpunctie en het bijvoeglijk naamwoord dat van het werkwoord is afgeleid, wordt zowel in het boek als op cd-rom betrekkelijk weinig geoefend. Bovendien zijn er te weinig oefeningen bij de werkwoordsvormen in die zin, dat er alleen simpele oefeningen beschikbaar zijn. Het niveau dat met name bij de werkwoordsvormen van de studenten gevraagd wordt, is laag op twee manieren. In de eerste plaats is het steeds duidelijk om welke werkwoordsvorm het gaat, bijvoorbeeld oefeningen met uitsluitend de persoonsvorm tegenwoordige tijd of uitsluitend het voltooid deelwoord. Verder zijn de zinnen waarin de werkwoordsvormen voorkomen simpel en kort. Van een opbouw van eenvoudig naar complex is bij de oefeningen nauwelijks sprake. Alleen bij de afsluitende opdracht op cd-rom komen alle behandelde onderdelen terug, maar nog steeds in simpele zinnen.

Als feedbackmiddel is de methode redelijk geschikt. Bij iedere opdracht wordt weliswaar feedback gegeven en die is betrouwbaar, maar op cd-rom is de feedback niet in een oogopslag duidelijk voor de student. Als hij wil weten waar zijn fouten zitten, moet hij op cd-rom namelijk alle items bij langslopen en zeker voor zwakke studenten kan het analyseren van fouten problematisch zijn. De feedback is ook niet snel. Zowel bij het boek als de cd-rom moet eerst een opdracht volledig gemaakt worden, voordat bekeken kan worden hoe de items gemaakt zijn. Als de student in het boek per item checkt, ziet hij meteen de volgende antwoorden. Afgezien van het controleren van de antwoorden, bestaat op cd-rom de mogelijkheid voor het tonen van de goede antwoorden. Bij deze optie gaat het altijd mis. De goede antwoorden kunnen namelijk zichtbaar worden, voordat de student een antwoord gegeven heeft. Als hij al wel de opdracht gemaakt heeft, is hij bij deze optie zijn eigen antwoorden in één keer kwijt. Uitleg ontbreekt bij de opdrachten in het boek. Van enige uitleg is al-

leen sprake bij de toetsen op cd-rom in de vorm van een verwijzing naar de desbetreffende theorie. Voortgangsinformatie ontbreekt in het boek. Op cd-rom is die wel beschikbaar in de vorm van twee begin- en eindtoetsen. Een toets is onvoldoende gemaakt bij meer dan vier fouten. Na afloop van de toets wordt de score vermeld die bestaat uit het aantal goede en foute antwoorden. De student moet zelf vaststellen of hij een voldoende of onvoldoende heeft, want daarover wordt niets aangegeven. Bij een onvoldoende resultaat is het de bedoeling dat de student de theorie of een deel daarvan (opnieuw) bestudeert en oefeningen maakt. Hij kan daarbij vrij gericht te werk gaan door de verwijzing naar een bepaald theoriegedeelte bij een onjuist antwoord, maar dan moet hij wel alle antwoorden bij langslopen. Bij de oefeningen wordt ook voortgangsinformatie gegeven, maar die is nietszeggend. Er wordt vermeld dat wie te veel fouten maakt de stof nog een keer moet bestuderen, zonder dat aangegeven wordt om hoeveel fouten het dan gaat. Verder worden de resultaten niet bijgehouden, zodat het onduidelijk is wat de student gedaan heeft en wat zijn scores waren.

Ondanks de bovengenoemde nadelen, leent deze methode zich vrij goed voor zelfstandig gebruik door een student die enige grammaticale basiskennis heeft.

Taaltopics Spelling

Taaltopics Spelling gaat over spelling van de werkwoordsvormen, algemene spellingkwesties, zoals meervouden, samenstellingen en hoofdletters en interpunctie. Deze methode is bestemd voor studenten in het hoger onderwijs. Over het beginniveau wordt niets gezegd. Het doel is dat studenten zich zelfstandig de spellingregels eigen maken. Bij deze methode is een website beschikbaar die alleen materiaal bevat bij één diagnostische toets uit het boek. Het boek is bedoeld voor zelfstudie. De spelling van de werkwoordsvormen, algemene spellingkwesties over in het oog lopende fouten bij het persoonlijk en bezittelijk voornaamwoord en interpunctie werden beoordeeld.

Als informatiebasis is *Taaltopics Spelling* alleen goed bruikbaar voor de student die grammaticale kennis heeft. Voor studenten met minder basis zal de uitleg soms moeilijk te volgen zijn. Niet alleen de gebruikte terminologie veronderstelt voorkennis, maar soms gaat de uitleg ook wat ver. De bijlage met grammatica moet studenten bijspijkeren, maar het zal niet eenvoudig zijn om dat in kort bestek via zelfstudie te doen. De uitleg over de spelling- en interpunctieregels is alleen redelijk relevant voor de opdrachten, omdat theorie en oefeningen van elkaar losgekoppeld zijn.

Als oefenboek is het in een bepaald opzicht zeker geschikt. Met de werkwoordsvormen kan veel geoefend worden en de meeste oefeningen zijn snel te maken. Bij interpunctie is het aantal oefeningen al beperkter en sommige opdrachten zijn tijdrovend. Maar de ene opdracht die beschikbaar is voor de spelling van het persoonlijk en bezittelijk voornaamwoord is onder de maat. De opdrachten passen wel bij alle behandelde onderdelen uit de theorie. Ook is er bij de werkwoordsvormen en interpunctie sprake van een bepaalde opbouw van gemakkelijk naar moeilijk, hoewel het met de interpunctie oefeningen in het begin misgaat. Daar staat de moeilijke oefening voorop. Het niveau van de zinnen is vrij hoog. Alleen in beperkte mate is wat dat betreft sprake van opbouw in moeilijkheidsgraad. De zinnen in de eerste deeloefeningen zijn weliswaar iets simpeler, maar ze worden al snel complexer.

Als feedbackmiddel is de methode redelijk geschikt. De student kan weliswaar checken hoe hij de oefeningen gemaakt heeft en bij iedere opdracht zijn de antwoorden beschikbaar die ook zonder meer duidelijk maken of het antwoord goed of fout is. Maar van snelle feedback is geen sprake. De antwoorden moeten achterin het boek worden opgezocht en bovendien is het niet mogelijk om per gemaakt item na te kijken, want dan is het volgende antwoord meteen te zien. Uitleg bij een fout antwoord ontbreekt. Voortgangsinformatie lijkt aanwezig te zijn, want meerdere diagnostische toetsen zijn beschikbaar. Maar bij deze toetsen moet de student zelf vaststellen wanneer hij de stof voldoende beheerst. Hij kan namelijk wel vaststellen hoeveel fouten hij heeft gemaakt, maar door het ontbreken van een norm moet hij daar zelf een waarde aan toekennen.

Voor zelfstudie is deze methode alleen bruikbaar voor de student die grammaticale basiskennis heeft.

Taaltopics Formuleren

Taaltopics Formuleren gaat over alinea-indeling, stijlkwesties en grammaticale fouten. Deze methode is bestemd voor studenten in het hoger onderwijs en voor iedereen die zakelijke teksten moet schrijven. Het niveau kan dus verschillend zijn. Het doel is om teksten te produceren die helder, begrijpelijk en correct geformuleerd zijn. Het boek is bedoeld voor zelfstudie. Hier werd gekeken naar foutieve verwijzing.

Als informatiebasis is *Taaltopics Formuleren* alleen bruikbaar voor de student die al wat grammaticale kennis heeft. Het is de vraag of een hbo-student wat dat betreft voldoende weet om deze uitleg te kunnen volgen. Het komt erop neer dat de gegeven informatie niet goed aansluit bij het elementaire karakter van deze fouten. Verder moet de student zelf de link leggen tussen theorie en praktijk, omdat eerst alle theorie gegeven wordt en daarna zijn de opdrachten opgenomen.

Als oefenboek is het niet geschikt. Tussen alle oefenzinnen met grammaticale fouten komen af en toe verwijsfouten voor, maar van een systematische oefening met deze onderdelen is geen sprake. De oefeningen zijn ook tijdrovend.

Als feedbackmiddel is de methode redelijk geschikt. Het is mogelijk om te checken hoe de oefeningen gemaakt zijn, want van iedere opdracht zijn de uitwerkingen aanwezig. Bij die uitwerkingen staat ook uitleg die gericht is op dat specifieke item. Maar van snelle feedback is geen sprake. De antwoorden staan achterin het boek en het is niet mogelijk om per gemaakt item naar het resultaat te kijken, want dan is het volgende antwoord ook meteen zichtbaar. Voortgangsinformatie ontbreekt en daardoor weet de student niet wanneer hij de stof voldoende beheerst.

Voor zelfstudie is deze methode met het oog op verwijswoorden alleen in beperkte mate geschikt.

Wolters' Nederlands in je pocket

Wolters' Nederlands in je pocket gaat over de spelling van Nederlandse woorden, de spelling van werkwoorden, interpunctie, grammatica, stijl, tekststructuur, argumentatie en tekstsoorten. Dit boek is bedoeld voor iedereen die snel iets wil opzoeken over verschillende aspecten van het Nederlands. Het niveau van de doelgroep kan dus divers zijn. Het doel is de taalgebruiker te behoeden voor bepaalde schrijffouten. De stof kan zelfstandig worden doorgenomen. In het kader van dit onderzoek wordt alleen gekeken naar de spelling van werkwoorden, leestekens en verwijswoorden. Met het oog op de criteria die in het begin gesteld zijn aan de methodes die beoordeeld worden, valt dit boek eigenlijk buiten beschouwing. Het is namelijk in de eerste plaats een naslagwerk en geen methode om iets te leren. Toch werd het hier geanalyseerd, omdat een paar studenten van het Instituut voor Marke-

ting Management van de Hanzehogeschool Groningen hebben aangegeven, hoeveel ze aan dit boek hadden om correct te leren spellen.

Als informatiebasis is *Wolters'Nederlands in je pocket* geschikt voor de student die zijn kennis alleen maar hoeft op te halen en al wat weet op het gebied van werkwoordspelling, leestekens en verwijswoorden. Dat laatste past ook bij het karakter van een naslagwerk.

Als oefenboek is het ongeschikt, want oefenmateriaal ontbreekt. De uitleg is daardoor niet toe te passen.

Als feedbackmiddel is het eveneens ongeschikt. Uiteraard ontbreekt de feedback als de oefeningen ontbreken. Het gevolg is dat de student die met dit boek werkt, niet kan vaststellen wanneer hij de stof voldoende beheerst.

Deze methode is dan ook niet geschikt voor zelfstudie, maar dat neemt niet weg dat het wel als naslagwerk gebruikt kan worden. Toch waren een paar studenten die zwak waren in de spelling van werkwoorden positief over dit boek. In dezelfde periode volgden deze studenten lessen in het correct leren spellen van werkwoordsvormen. Na afloop haalden zij een voldoende op de taaltoets, maar wat daarbij de doorslag heeft gegeven is niet duidelijk. Volgens de hier gehanteerde opvatting over wat effectief leren is - uitleg, oefeningen en feedback horen bij elkaar - zou het dit boek niet mogen zijn.

Repetitieboekje Nederlands

Repetitieboekje Nederlands gaat over de spelling van werkwoorden en algemene spellingkwesties, stijlonderdelen en handelstermen. Deze methode is bestemd voor het economisch-administratief en commercieel onderwijs. Het niveau van de doelgroep is niet duidelijk. Het doel wordt voornamelijk geformuleerd vanuit het perspectief van de docent, die met deze methode bepaalde stof vlug kan herhalen en controle kan uitoefenen op de prestaties van de student. Hier werd naar de spelling van werkwoordsvormen gekeken.

Als informatiebasis is *Repetitieboekje Nederlands* alleen bruikbaar voor studenten die al wat grammaticale kennis hebben en die daardoor genoeg hebben aan summiere uitleg. Voor veel hbo-studenten zal dit niet opgaan. Van een directe link tussen de oefeningen en de be-

handelde theorie is ook geen sprake, omdat ze los van elkaar gepresenteerd worden. Verder is de uitleg door de beperkte en globale inhoudsopgave niet snel te vinden.

Als oefenboek is het goed geschikt. Er zijn vrij veel oefeningen beschikbaar die bij de werkwoordsvormen ook opklimmen in moeilijkheidsgraad. De oefeningen oefenen met alle dingen die nodig zijn. De meeste oefeningen kosten niet veel tijd.

Als feedbackmiddel is deze methode volledig ongeschikt, omdat iedere vorm van feedback ontbreekt. De student kan niet checken hoe hij de oefeningen maakt en hoe zijn vorderingen zijn. De student is bij deze methode afhankelijk van de feedback die de docent geeft.

Voor zelfstudie is deze methode daarom niet geschikt.

Praktische cursus Formuleren

Formuleren gaat over stijlkwesties, zoals het gebruik van bepaalde woorden (moeilijk of gemakkelijk, formeel of informeel), lange en korte zinnen (gebrek aan verband binnen de zin), het gebruik van beelden om iets begrijpelijk te maken en middelen als herhaling en opsomming om speciale aandacht te vragen. De doelgroep van *Formuleren* is breed, want het is namelijk bestemd voor iedereen die zijn stijl wil verbeteren bij het schrijven van zakelijke teksten. Het niveau van de doelgroep kan dus verschillend zijn. Het boek is ook bedoeld voor zelfstudie. Hier werd gekeken naar het gebrek aan verband binnen de zin en dan specifiek naar verwijswoorden.

Als informatiebasis is *Praktische cursus Formuleren* bruikbaar voor de taalgebruiker die al kennis heeft op dit gebied. De uitleg is relevant voor het maken van de opdrachten. Bij de behandelde onderdelen worden oefeningen gegeven, die ook meteen bij het desbetreffende theoriegedeelte staan. Maar de student die verwijswoorden problematisch vindt, zal weinig hebben aan de summiere uitleg.

Als oefenboek is het ongeschikt. Per onderdeel zijn maar een paar oefeningen beschikbaar. De oefeningen zijn ook niet snel te maken, omdat ze lastig zijn en niets voorgestructureerd is. Van opbouw in de oefeningen van gemakkelijk naar moeilijk is geen sprake.

Als feedbackmiddel is de methode eveneens ongeschikt. Voor de student zijn alleen uitwerkingen beschikbaar bij de slotopdracht van een hoofdstuk. Dat betekent dat hij bij de andere opdrachten afhankelijk is van de feedback van de docent. De feedback maakt niet altijd duidelijk of de oefening goed gemaakt is, omdat soms meerdere uitwerkingen mogelijk zijn. Verder is de feedback niet snel, want de antwoorden moeten achterin het boek worden opgezocht. De uitwerkingen van de zinnen die nog gemaakt moeten worden, zijn dan meteen zichtbaar. Uitleg ontbreekt, evenals voortgangsinformatie. De student kan daardoor niet weten in hoeverre hij de stof voldoende beheerst.

De bovengenoemde nadelen maken deze methode alleen in beperkte mate geschikt voor zelfstudie.

Basisvaardigheden Taal

Basisvaardigheden taal gaat over fouten op het gebied van woordkeus, zinsbouw, alinea-indeling, spelling van woorden in het algemeen, spelling van werkwoorden en interpunctie. Deze methode richt zich op hbo-studenten. Over het niveau van de doelgroep wordt niets gezegd. Het doel is het verbeteren van de taalvaardigheid van studenten. Bij *Basisvaardigheden taal* hoort een website met een begintoets, een eindtoets en oefeningen. Het boek is bedoeld voor zelfstudie. Hier is gekeken naar de spelling van werkwoorden, interpunctie en voornaamwoorden.

Als informatiebasis is *Basisvaardigheden Taal* alleen bruikbaar voor de student die al kennis heeft op het gebied van spelling van werkwoorden, voornaamwoorden en interpunctie. De uitleg is namelijk summier en zelfs onvoldoende bij het bijvoeglijk naamwoord dat van het werkwoord is afgeleid. Bij de uitleg wordt niet alleen kennis verondersteld, maar ook worden grammaticale begrippen gebruikt waarmee niet iedere hbo-student vertrouwd zal zijn. Bovendien is de uitleg in het boek gescheiden van de oefeningen die op de website staan.

Als oefenboek is het op zich ongeschikt, omdat er geen oefeningen in staan, maar de bijbehorende website met oefenmateriaal hoort bij dit boek. Dat maakt het behoorlijk geschikt

als oefenboek. Bij de behandelde onderdelen in het boek staan oefeningen op de website en die zijn zo voorgestructureerd dat ze snel te maken zijn. Meestal zijn 10 oefenzinnen per onderdeel beschikbaar en daarna kunnen weer nieuwe opgevraagd worden. Tussen die nieuwe items zitten alleen weer veel bekende en dus al gemaakte items. Per onderdeel valt het aantal oefeningen daardoor tegen. Bij het bijvoeglijk naamwoord dat van het werkwoord is afgeleid, zijn relatief weinig oefeningen beschikbaar, terwijl daarmee in de praktijk wel veel fouten worden gemaakt. Van een bepaalde opbouw in moeilijkheidsgraad is geen sprake. Het niveau is vooral bij de werkwoordsvormen vrij laag. De zinnen blijven kort en simpel.

Als feedbackmiddel is deze methode ook vrij geschikt, maar dan alleen met het oog op de website. In het boek staat uiteraard geen feedback, want er staan ook geen oefeningen in. Bij iedere opdracht op de website wordt feedback gegeven en die komt pas, nadat het antwoord gegeven is. Er wordt dus niet zomaar iets prijsgegeven. Maar de feedback is niet zonder meer duidelijk en helemaal niet snel. Verder is de uitleg bij een fout antwoord niet op het item toegespitst, maar standaard en bestaat uit een verwijzing naar een paragraaf uit het boek. Voortgangsinformatie krijgt de student op grond van de opdrachten en toetsen. Bij foute antwoorden krijgt hij het advies om bepaalde paragrafen uit het boek nog eens te bestuderen. Het aantal goede en foute antwoorden met literatuurverwijzing staat boven de gemaakte toets, maar de norm ontbreekt. Het is daardoor niet duidelijk hoeveel fouten gemaakt mogen worden en wanneer de student de stof voldoende beheerst. De begin- en eindtoets hebben vreemd genoeg niet evenveel items. De begintoets heeft 58 items, maar de beide eindtoetsen hebben ieder circa de halve hoeveelheid. Het ligt voor de hand dat studenten alleen al vanwege dat gegeven de eindtoets in het algemeen beter zullen maken. Het is verder opvallend dat zowel de begin-als de eindtoets in totaal weinig items bevatten. Per onderdeel kan op die manier niet veel getoetst worden. De resultaten worden niet bijgehouden.

Vanwege de genoemde nadelen kan alleen een wat gevorderde student deze methode gebruiken voor zelfstudie.

Dat d/t gedoe

Dat d/t gedoe gaat uitsluitend over de spelling van de werkwoordsvormen. Deze methode is bestemd voor iedereen die over de juiste spelling twijfelt of vaak fouten maakt. Het niveau van de doelgroep kan dus heel verschillend zijn. Het doel van deze methode is dat de taalgebruiker de spelling van werkwoordsvormen zo te leert te beheersen, dat hij niet langer onzeker is over d's en t's. Het boek is ook bedoeld voor zelfstudie.

Als informatiebasis is *Dat d/t gedoe* aan een kant geschikt. Er is veel uitleg beschikbaar in begrijpelijk Nederlands over de spelling van werkwoordsvormen en die informatie is relevant voor het verbeteren van de fouten in de voorbeelden. Maar het accent bij de uitleg ligt soms sterk op alles wat mis kan gaan en dat maakt het lastig om de grote lijn te volgen. De voorbeelden die bij de behandeling van een bepaalde werkwoordsvorm staan, werken rondduit verwarrend. Ze staan zonder uitleg tussen de tekst. Iemand die niet zorgvuldig in het begin van het boek de aanwijzingen voor het gebruik heeft gelezen, kan ze zelfs opvatten als voorbeelden van goed taalgebruik.

Als oefenboek is het grotendeels ongeschikt. Bij nagenoeg ieder behandeld onderdeel staan wel voorbeelden die verbeterd kunnen worden, maar die leveren uiteindelijk weinig oefenmateriaal op. In deze methode wordt er vanuit gegaan dat werkelijk te schrijven teksten als praktijkoefening bij de uitleg gebruikt kunnen worden. Maar hoe deze aanname kan resulteren in concreet oefenmateriaal wordt er niet bij vermeld. Verder kan het opsporen van de fouten in de voorbeelden tijdrovend zijn, omdat niets voorgestructureerd is. De voorbeelden zijn ook niet geordend op grond van een bepaalde moeilijkheidsgraad.

Als feedbackmiddel is het boek beperkt geschikt. De fouten uit de voorbeelden worden in het laatste hoofdstuk herhaald en daar voorzien van correctie met uitleg. Die uitleg wordt niet alleen in algemene termen gegeven, maar is meestal toegespitst op de fout zelf. Verder is de feedback weliswaar duidelijk, maar niet snel, want iedere verbetering moet achterin het boek worden opgezocht. Ook kan de feedback gemakkelijk op het verkeerde moment komen, omdat een aantal antwoorden tegelijk zichtbaar is. Voortgangsinformatie ontbreekt en daardoor weet de student niet wanneer hij de stof voldoende beheerst.

Deze nadelen maken de methode alleen in beperkte mate geschikt voor zelfstudie.

Spelbewust

Spelbewust gaat over spelling in het algemeen van Nederlandse woorden, spelling van werkwoorden en interpunctie. Deze methode heeft als doelgroep studenten van het mbo. Er wordt vanuit gegaan dat het niveau van de doelgroep gevarieerd is. Het doel is om studenten zo goed mogelijk voor te bereiden op de vereiste taalvaardigheden tijdens de opleiding en de latere beroepspraktijk. Het boek is ook bedoeld voor zelfstudie. Deze methode over spelling richt zich weliswaar op mbo-studenten, maar zou evengoed geschikt kunnen zijn voor hbo-studenten. Dat is de reden waarom *Spelbewust* hier behandeld is. De spelling van werkwoorden en interpunctie werden beoordeeld.

Als informatiebasis is *Spelbewust* bruikbaar voor de student die al wat kennis heeft op het gebied van spelling en interpunctie. De uitleg is redelijk relevant voor de oefeningen. Er is geen directe link tussen opdrachten en uitleg, omdat ze uit elkaar geplaatst zijn. Het is ook de vraag of niet te veel kennis verondersteld wordt bij zowel studenten van het mbo als het hbo.

Als oefenboek is het voor een deel zeker geschikt. Er zijn vrij veel oefeningen per onderdeel beschikbaar en die zijn zo voorgestructureerd dat ze snel te maken zijn. Verder zijn de oefeningen bij de werkwoordsvormen enigszins geordend op grond van hun moeilijkheidsgraad, maar de oefenzinnen blijven simpel. Het is onduidelijk op welke manier de oefeningen tegelijk als toetsen kunnen dienen.

Als feedbackmiddel is de methode ongeschikt, omdat iedere vorm van feedback ontbreekt. De student kan niet checken hoe hij de oefeningen gemaakt heeft en hij weet dus niet hoe zijn vorderingen zijn. Hij is bij deze methode afhankelijk van de feedback die de docent geeft.

Het is niet duidelijk hoe dit boek geschikt zou moeten zijn voor zelfstudie.

Praktische cursus spelling

Praktische cursus spelling gaat over spelling in het algemeen van Nederlandse woorden, spelling van werkwoorden en interpunctie. De doelgroep van deze methode is breed: middelbare scholieren, hbo-studenten, universitaire studenten en verder iedereen die onzeker is over zijn spelling. Het niveau van de doelgroep is dan ook niet vast te stellen. Het doel is om correct te leren schrijven in het Nederlands. Er wordt vanuit gegaan dat het boek ook geschikt is voor zelfstudie.

Als informatiebasis is *Praktische cursus spelling* bruikbaar voor een student die al bepaalde kennis heeft op het gebied van spelling en interpunctie. De uitleg is zonder meer relevant voor het maken van de opdrachten. Bij de werkwoordsvormen en interpunctie wordt meteen geoefend na een stukje uitleg. Maar de terminologie veronderstelt voorkennis en de uitleg is vrij beknopt en soms onduidelijk bij de voorbeelden. Binnen een hoofdstuk wordt de paragraafindeling bij de uitleg doorkruist door een andere paragraafindeling van de oefeningen. Dat maakt het geheel niet overzichtelijker.

Als oefenboek is het tot op zekere hoogte geschikt. De meeste oefeningen zijn snel te maken. Maar de hoeveelheid opdrachten valt tegen, ook als de toets bij een hoofdstuk als oefenmateriaal wordt beschouwd. De oefeningen zijn beperkt geordend op grond van hun moeilijkheidsgraad, in die zin dat eerst een bepaald element aan de orde komt, terwijl op het laatst alle behandelde onderdelen uit het hoofdstuk terugkomen in een oefening. Maar het niveau van de oefenzinnen is in het begin laag en dat blijft ook zo.

Als feedbackmiddel is de methode alleen voor een deel geschikt. Bij iedere opdracht zijn de antwoorden beschikbaar en die maken ook zonder meer duidelijk of het antwoord goed of fout is. Maar er is geen sprake van snelle feedback, omdat de antwoorden achterin het boek staan. Bovendien moet een opdracht eerst volledig gemaakt worden, omdat anders een volgend antwoord meteen zichtbaar is. Uitleg bij de antwoorden ontbreekt. Voortgangsinformatie is evenmin beschikbaar. Bij de toetsen staat niet aangegeven wanneer de student de stof voldoende beheerst.

Door de genoemde nadelen is deze methode alleen tot op zekere hoogte geschikt voor zelfstudie.

Een Goede Spelling

Een Goede Spelling heeft als onderwerpen algemene spelling van woorden, werkwoordspelling en interpunctie**.** Deze methode is bestemd voor hbo-studenten en richt zich op studenten met zowel een Nederlandse als een anderstalige achtergrond. Het beginniveau kan daarom verschillend zijn. Het doel is dat hbo'ers in staat zijn een correcte tekst te schrijven. Het boek is ook bedoeld voor zelfstudie.

Als informatiebasis is *Een Goede Spelling* bruikbaar voor hbo-studenten die al wat weten op dit gebied. De uitleg is niet meer dan redelijk relevant voor het maken van de opdrachten. Afgezien van een paar kleine opdrachten bij de uitleg zelf, staan de oefeningen namelijk losgekoppeld van de theorie. Voor de beginnende speller is de informatie ongeschikt, want de uitleg is summier en er wordt kennis verondersteld van grammaticale begrippen.

Als oefenboek is het voor een deel geschikt. Het maken van de oefeningen kost meestal niet veel tijd en de oefeningen passen bij de behandelde theorie. Maar er zijn per onderdeel maar weinig oefeningen beschikbaar en met het bijvoeglijk naamwoord wordt zelfs nauwelijks geoefend. De oefeningen zijn tot op zekere hoogte geordend op grond van hun moeilijkheidsgraad. Het niveau van de oefenzinnen is redelijk, maar een bepaalde opbouw valt daarin niet te ontdekken.

Als feedbackmiddel is de methode redelijk geschikt. Bij iedere opdracht zijn de antwoorden beschikbaar, die ook zonder meer duidelijk maken of het gegeven antwoord goed of fout is. Maar de feedback wordt niet snel gegeven, omdat het antwoord achterin het boek opgezocht moet worden. Verder kunnen de antwoorden pas bekeken worden, nadat de hele opdracht gemaakt is, omdat de feedback anders op het verkeerde moment gegeven wordt. Uitleg bij de antwoorden is niet beschikbaar. De voortgangsinformatie bestaat bij de diagnostische toets alleen uit een verwijzing naar een bepaald hoofdstuk als er meerdere fouten zijn gemaakt. Het is niet duidelijk in welk stadium de student dat advies moet opvolgen, doordat de norm bij de toets ontbreekt. Verder is geen voortgangsinformatie beschikbaar. De twee reflectievragen die aan het einde van iedere oefening staan, zijn door hun algemene en vage karakter niet als zodanig te beschouwen.

Voor zelfstudie is dit boek alleen tot op zekere hoogte geschikt voor de al wat gevorderde taalgebruiker op dit gebied.

Basisboek Spelling

Basisboek Spelling gaat over de spelling van woorden, de spelling van werkwoorden**,** interpunctie en een paar stijlkwesties. Deze methode is bestemd voor studenten in het hoger onderwijs en voor iedereen die moeite heeft met de spelling van het Nederlands. Het beginniveau kan dan ook verschillend zijn. Het doel is het voorkomen van storende taalfouten. Het boek is ook bedoeld voor zelfstudie. Bij deze methode is een website beschikbaar met dictees om het beginniveau te bepalen. Hier werden de volgende onderdelen beoordeeld: werkwoordspelling, interpunctie en stijlkwesties over voornaamwoorden ('dat' en 'wat', 'u' en 'uw').

Als informatiebasis is *Basisboek Spelling* goed bruikbaar voor iemand met wat kennis op het gebied van grammaticale begrippen. De uitleg is ook vrij relevant voor het maken van de opdrachten. Maar met het oog op de brede doelgroep, is het de vraag of in het algemeen de basis aanwezig is om met de uitleg uit de voeten te kunnen.

Als oefenboek is het behoorlijk geschikt. Vooral bij de werkwoordsvormen zijn veel oefeningen beschikbaar, maar het oefenmateriaal bij interpunctie en zeker bij de elementaire stijlkwesties is ontoereikend. Het maken van de meeste oefeningen kost niet veel tijd, omdat ze voorgestructureerd zijn. De opdrachten zijn ook redelijk geordend van gemakkelijk naar moeilijk, in die zin dat aan het einde van ieder hoofdstuk eindopdrachten zijn opgenomen, waarin alle behandelde zaken uit dat hoofdstuk aan bod komen. Maar het niveau van de oefenzinnen varieert vrij willekeurig.

Als feedbackmiddel is de methode redelijk geschikt. De antwoorden zijn bij iedere opdracht beschikbaar en die maken duidelijk of het item goed of fout gemaakt is. Maar de feedback is niet snel, want de antwoorden moeten achterin het boek worden opgezocht en dat kan bovendien pas als de opdracht volledig gemaakt is. Per gemaakt item is het antwoord namelijk niet te checken, omdat dan het volgende antwoord ook zichtbaar is. Uitleg bij de antwoorden wordt niet gegeven en voortgangsinformatie is niet beschikbaar. Na de gemaakte dictees van de website en ook na de oefeningen uit het boek, weet de student niet in hoeverre hij de stof beheerst.

Voor zelfstudie is deze methode vrij goed bruikbaar voor iemand die basiskennis heeft van grammaticale begrippen in het Nederlands.

Formuleren

Formuleren gaat over bepaalde stijlverschijnselen. Deze methode is bedoeld voor studenten die het vak taalbeheersing volgen en voor iedereen die een vakmatige belangstelling heeft voor geschreven Nederlands. Het niveau van de beoogde gebruikers van het boek kan dus verschillend zijn. Het doel is te beschrijven en te verklaren wat het effect is van een bepaalde manier van uitdrukken. Het gaat daarom bij de behandelde onderdelen niet zozeer om correctheid, maar om het effect op de inhoud. Het boek is ook bedoeld voor zelfstudie. *Formuleren* geeft adviezen voor helder taalgebruik en dat maakt deze methode de moeite waard om hier te analyseren. Alleen die onderdelen zullen aan bod komen die over elementaire kwesties gaan, zoals tijden en vormen van werkwoorden, verwijswoorden en interpunctie.

Als informatiebasis is *Formuleren* mogelijk interessant voor een student taalbeheersing die geïnteresseerd is in de analyse van taalverschijnselen, maar als informatiebasis voor schrijfadviezen is het niet bruikbaar. Er wordt weliswaar uitleg gegeven met voorbeelden, maar het ontbreekt wel eens aan een toegankelijke gedachtegang en toegankelijk Nederlands. Niet alleen wordt er veel voorkennis verondersteld, maar er is ook geen duidelijk verband tussen de gepresenteerde stof en de opdrachten.

Als oefenboek is het niet geschikt. Aan het einde van ieder hoofdstuk worden wel een paar opdrachten opgenomen, maar het maken van de oefeningen kost relatief veel tijd, omdat ze lastig zijn en beperkt voorgestructureerd. De oefeningen zijn bovendien gedeeltelijk relevant, vanwege hun globale karakter. Verder wordt er niet voldoende geoefend met alles wat behandeld is. Van een opbouw in de oefeningen qua moeilijkheidsgraad is geen sprake.

Als feedbackmiddel is de methode eveneens ongeschikt, omdat er geen feedback gegeven wordt. De student die de doorgaans lijvige opdrachten gemaakt heeft, weet niet hoe hij dat gedaan heeft en of hij de stof beheerst. Hij is met dit studieboek afhankelijk van de feedback van de docent.

Dit boek bevat geen zelfstudiemateriaal voor wie helder wil leren schrijven.

Vlekkeloos Nederlands

Vlekkeloos Nederlands gaat over spelling van woorden in het algemeen, de spelling van werkwoorden, stijlkwesties, interpunctie en grammatica. Dit boek richt zich op een brede doelgroep, namelijk op leerlingen van havo, vwo, mbo en op studenten van hbo en universiteit. Het niveau van de doelgroep is dus divers. Het doel is dat leerlingen en studenten hun kennis van Nederlands op de bovengenoemde gebieden vergroten. Deze methode is ook bedoeld voor zelfstudie. Hier werd gekeken naar de spelling van werkwoorden, elementaire stijlkwesties over voornaamwoorden en interpunctie.

Als informatiebasis is *Vlekkeloos Nederlands* tot op zekere hoogte bruikbaar voor iemand met kennis van grammaticale begrippen en basiskennis op het gebied van spelling, stijl en interpunctie. De uitleg is redelijk relevant voor het maken van de opdrachten. Maar bij het bijvoeglijk naamwoord is de uitleg te beperkt en die kan zelfs aanleiding geven tot misverstanden.

Als oefenboek is het voor een deel behoorlijk geschikt. Relatief veel oefenmateriaal is aanwezig bij de spelling van werkwoorden en de meeste oefeningen zijn snel te maken. De oefeningen zijn ook geordend op grond van hun moeilijkheidsgraad, alleen blijven de zinnen wel vrij simpel. Wat dat betreft, is het oefenmateriaal onvoldoende. Oefeningen in meer complexe zinnen ontbreken. Een nadeel is verder dat weinig geoefend kan worden met het bijvoeglijk naamwoord dat van het werkwoord is afgeleid. Studenten maken daarmee in de praktijk juist wel veel fouten. Stijloefeningen gericht op de voornaamwoorden zijn minder beschikbaar en met het onderdeel interpunctie kan uitgesproken weinig geoefend worden.

Als feedbackmiddel is de methode redelijk geschikt. Bij alle opdrachten zijn de goede antwoorden beschikbaar en die maken duidelijk hoe een opdracht gemaakt is. Maar de feedback is niet snel. Eerst moet een opdracht volledig gemaakt worden, want anders zijn de goede antwoorden op de volgende items al zichtbaar. Uitleg bij de antwoorden ontbreekt, evenals voortgangsinformatie. Na het maken van de opdrachten moet de student dus zelf vaststellen of hij de stof voldoende beheerst.

Deze nadelen beperken de geschiktheid van de methode voor zelfstudie.

Van verslag tot rapport

Van verslag tot rapport gaat over het schrijven van rapporten, van elementaire tot de meer verslagtechnische zaken. Deze methode richt zich met name op studenten in het hbo. Het niveau kan variëren van een beginnende student die projectwerk moet maken tot iemand die al ervaring heeft met het schrijven van rapporten. Het doel van dit boek is dat studenten leren om een professioneel rapport te schrijven. Het boek is vooral bedoeld voor zelfstudie. Bij het boek hoort een website met oefenmateriaal. De analyse werd hier beperkt tot elementaire taalvaardigheidsaspecten, zoals spelling van werkwoorden, verwijswoorden en interpunctie.

Als informatiebasis is *Van verslag tot rapport* niet bruikbaar. De uitleg is summier en wordt niet in toegankelijk Nederlands gegeven voor de beginnende student en vermoedelijk ook niet voor hbo-studenten met meer schrijfervaring. Te veel kennis wordt als bekend verondersteld. De uitleg in deze methode is aan de ene kant ook niet bedoeld om deze elementaire vaardigheden onder de knie te krijgen, maar aan de andere kant is wel aangegeven hoe storend fouten op dit gebied zijn. De oefeningen op de website gaan ook over deze fouten, maar de kennis om ze te vermijden moet de student elders opdoen.

In het boek zelf staan geen oefeningen, maar die staan wel op de bijbehorende website. Als oefenboek is het alleen redelijk geschikt voor de gevorderde student. Het niveau van het oefenmateriaal op de website is namelijk te hoog voor een beginner. Ook voor de student met schrijfervaring is weinig oefenmateriaal voorhanden, want in totaal zijn slechts twee teksten beschikbaar. De oefeningen zijn tijdrovend, omdat een fout steeds aangeklikt moet worden in een lijst waarin alle fouten voorkomen. Dat levert veel zoekwerk op. Het tweede type oefening waarbij fouten in een tekst niet staan aangegeven, maar zelf moeten worden opgezocht, kost nog meer tijd. Van een bepaalde opbouw in de oefeningen is nauwelijks sprake. In het eerste teksttype staat al aangegeven waar de fout zit en in het tweede type moet de student die zelf opzoeken. Maar een ordening van gemakkelijk naar moeilijk ontbreekt bij de oefeningen, omdat die meteen vrij lastig zijn.

Als feedbackmiddel is de methode tot op zekere hoogte geschikt. De feedback is correct en komt meteen na ieder gemaakt item. Maar duidelijk is de feedback niet zonder meer. Als de student het item goed maakt, wordt het goede antwoord zichtbaar, maar hij krijgt niet expliciet een goed-melding. Bij een fout antwoord krijgt de student te lezen dat het foute ant-

woord gekozen is en de student moet het opnieuw proberen. Na drie mislukte pogingen krijgt hij het goede antwoord wel te zien. Dit kan gemakzucht in de hand werken, want bij drie keer klikken wordt het goede antwoord standaard weggegeven. Opvallend is dat het scherm bij een goed- en foutmelding identiek is. In beide gevallen staat 'ok' aangegeven. Bij oppervlakkig kijken kan dat in beide gevallen als een goedmelding worden opgevat. Uitleg bij een fout antwoord is niet beschikbaar, er is ook geen verwijzing naar het desbetreffende onderdeel uit de theorie. Voortgangsinformatie ontbreekt, zodat de student niet weet wanneer hij de stof voldoende beheerst.

Voor zelfstudie is dit boek alleen gedeeltelijk geschikt voor een gevorderde student die het geduld heeft om de oefeningen te maken.

Welgespeld

Welgespeld gaat uitsluitend over de spelling van werkwoorden, met uitzondering van twee slothoofdstukken waarin veel gemaakte fouten bij werkwoorden en zelfstandige naamwoorden vergeleken worden. Deze methode is bestemd voor studenten uit het hoger onderwijs. Het niveau van de doelgroep is van beginnende tot gevorderde speller. De beginnende speller kan ook een student zijn die het Nederlands niet als moedertaal heeft. Het doel is dat studenten zichtbaar beter werkwoorden kunnen spellen in zakelijke teksten die zij tijdens de opleiding of later in de beroepssituatie moeten schrijven. Bij het boek hoort een website met extra materiaal, dat bestaat uit een lijst met veel voorkomende sterke en onregelmatige werkwoorden. Het boek is ook bedoeld voor zelfstudie.

Welgespeld is als informatiebasis tot op zekere hoogte bruikbaar voor de gevorderde student op dit gebied. De uitleg is relevant voor het maken van de oefeningen. Maar het is de vraag of er onder hbo-studenten veel gevorderde spellers zitten. *Welgespeld* richt zich weliswaar op het hele hoger onderwijs, maar het is de vraag of het taalvaardigheidsniveau van studenten aan de universiteiten zoveel hoger ligt. Voor de beginnende speller, onder wie ook de meertalige student die het Nederlands niet als moedertaal heeft, ligt het niveau van de informatie in deze methode te hoog op meerdere terreinen. Er wordt weinig uitleg gegeven, bij de uitleg wordt grammaticale kennis als bekend verondersteld en en de materie die wordt behandeld is soms te hoog gegrepen voor deze doelgroep.

Als oefenboek is het behoorlijk geschikt voor de meer gevorderde speller. De meeste oefeningen zijn weliswaar redelijk snel te maken en bij alle onderdelen worden relevante oefeningen gegeven. Ook is sprake van een bepaalde opbouw in de oefeningen van gemakkelijk naar moeilijk. Maar voor de beginnende speller is per onderdeel onvoldoende oefenmateriaal aanwezig, want het niveau ligt ook bij de relatief gemakkelijke oefeningen meteen vrij hoog.

Als feedbackmiddel is de methode voor een deel zeker geschikt. Voor de student zijn de antwoorden beschikbaar van de meeste opdrachten en de antwoorden zijn correct en maken meteen duidelijk hoe een oefening gedaan is. Bij meerdere antwoorden wordt uitleg gegeven, specifiek gericht op het antwoord of met een verwijzing naar een bepaald hoofdstuk. Bij de antwoorden van de toetsen staat consequent bij ieder antwoord uitleg specifiek gericht op dat antwoord en een hoofdstukverwijzing. Maar de feedback is niet snel. Bij de opdrachten moet ieder antwoord achterin het boek worden opgezocht en dat betekent dat de student zijn werk pas kan nakijken als de hele opdracht gemaakt is, anders kost het te veel tijd. Wie per gemaakt item nakijkt, ziet bovendien meteen het volgende goede antwoord. Voortgangsinformatie is beschikbaar, maar daarin ontbreekt een essentieel element. Bij fouten wordt wel verwezen naar bepaalde hoofdstukken die de student opnieuw moet bestuderen. Verder staan in deze methode meerdere toetsen, maar de norm daarbij ontbreekt. Daardoor kan de student niet weten wanneer hij de stof voldoende beheerst. Alleen voorin het boek staat daarover iets vermeld. Uit die informatie kan de student halen dat de oefeningen en de toetsen foutloos gemaakt moeten worden. Op de eindtoets mogen maximaal drie fouten gemaakt worden. Maar de kans is groot dat studenten over deze informatie heen lezen, omdat die niet bij de oefeningen en toetsen zelf is opgenomen. De website die bij het boek hoort, is overbodig, want de lijst met sterke en onregelmatige werkwoorden staat ook als bijlage achterin het boek.

Voor zelfstudie is dit boek behoorlijk geschikt voor de gevorderde student die zijn kennis wil ophalen.

Eindbeoordeling papieren methodes

In totaal zijn zeventien papieren methodes geanalyseerd. Een paar criteria zijn noodgedwongen anders ingevuld dan de bedoeling was. Bij de criteria relevantie onder informatiebasis en relevantie en volledigheid onder oefenboek had eigenlijk gekeken moeten worden naar de doelstelling van een methode. Dit was niet haalbaar. De doelstellingen waren bij alle methodes vaag geformuleerd of afwezig. Daarom is bij relevante uitleg gekeken naar hoe relevant de informatie was voor het maken van de oefeningen, vanuit de gedachtegang dat die oefeningen operationaliseren wat de student moet kunnen. Bij de relevantie en volledigheid van het oefenmateriaal is gekeken naar hoe relevant en volledig de oefeningen waren met het oog op de uitleg, vanuit het idee dat het oefenmateriaal moest passen bij de gegeven instructie.

In Tabel 5.4 zijn de toegekende scores voor de verschillende aspecten van de methodes vermeld. In Tabel 5.5 wordt per methode het gemiddelde voor informatiebasis, oefenboek en feebackmiddel vermeld, alsmede het totale gemiddelde.

Tabel 5.4 Beoordeling papieren methodes per aspect

	Methode	Informatiebasis					Oefenboek					Feedbackmiddel					
		juist	volledig	duidelijk	relevantie	toegank	veel	gemak	relevantie	volledig	geordend	veel	betrouwb	duidelijk	snel	afhank	voortg
1	*Gewoon goed Nederlands*	5	3	3	3	5	2	2	5	2	1	0	0	0	0	0	0
2	*Basisvaardigheden Spelling*	5	3	4	4	5	3	5	5	5	2	5	5	4	1	1	3
3	*Taaltopics Spelling*	4	3	3	3	5	3	3	5	5	3	5	5	5	1	2	1
4	*Taaltopics Formuleren*	5	3	3	3	5	1	1	5	2	1	5	5	5	1	2	1
5	*Nederlands in je pocket*	5	3	3	1	5	0	0	0	0	0	0	0	0	0	0	0
6	*Repetitieboekje Nederlands*	5	2	3	3	2	4	4	5	5	4	0	0	0	0	0	0
7	*Praktische cursus Formul.*	5	2	3	5	5	1	2	5	5	1	1	5	3	1	2	1
8	*Basisvaardigheden Taal*	5	2	2	3	5	3	5	5	5	1	5	5	4	1	5	3
9	*Dat d/t gedoe*	5	3	3	5	5	1	2	5	4	1	5	5	5	1	2	1
10	*Spelbewust*	5	3	3	3	5	3	5	5	5	2	0	0	0	0	0	0
11	*Praktische cursus spelling*	5	2	2	5	4	2	4	5	5	2	5	5	5	1	2	1
12	*Een goede spelling*	5	3	3	3	5	2	4	5	5	3	5	5	5	1	2	2
13	*Basisboek Spelling*	5	3	3	4	5	3	4	5	5	3	5	5	5	1	2	1
14	*Formuleren*	4	2	2	2	5	1	1	2	2	1	0	0	0	0	0	0
15	*Vlekkeloos Nederlands*	5	2	3	3	5	3	4	5	5	3	5	5	5	2	2	1
16	*Van verslag tot rapport*	4	2	2	1	5	2	2	5	5	2	5	5	1	5	5	1
17	*Welgespeld*	5	2	2	4	5	3	4	5	5	3	5	5	5	1	2	3
	Gemiddelde	**4,8**	**2,5**	**2,8**	**3,2**	**4,8**	**2,2**	**3,1**	**4,5**	**4,1**	**1,9**	**3,2**	**3,5**	**3,0**	**1,0**	**1,7**	**1,1**

Bij feedbackmiddel ontbreekt in Tabel 5.4 het criterium uitleg. Bij de analyse van de methodes is wel aangegeven in hoeverre de student uitleg krijgt bij dingen die hij niet begrijpt en dus fout doet. Dit criterium bleek echter op twee manieren problematisch te zijn. In de eerste plaats was het bij een fout antwoord moeilijk om gerichte uitleg te geven, als er meerdere manieren zijn om iets fout te beantwoorden. Maar ook bij voorgestructureerde antwoorden, waarbij gerichte uitleg wel tot de mogelijkheden behoort, deed zich een probleem voor. Bij uitleg start de hele cyclus namelijk opnieuw, want oefeningen en feedback zijn nodig om te weten of de uitleg begrepen is. Vanwege deze problemen is het criterium uitleg uiteindelijk niet in de tabel opgenomen.

Wat opvalt in Tabel 5.4, is dat onder informatiebasis de hoogste waarden staan bij de criteria juistheid en toegankelijkheid. Dat betekent dat de uitleg doorgaans correct was en dat de informatie snel te vinden was. Relevantie scoorde daarna hoog, wat hier wil zeggen dat de uitleg relevant was, met het oog op de opdrachten.

De criteria volledigheid en duidelijkheid scoorden het laagst, doorgaans kregen ze niet meer dan matig. Bij volledigheid kregen acht methodes een slechte beoordeling en bij duidelijkheid waren dat er vijf. Dit betekent dat de uitleg in die gevallen ontoereikend was en niet in toegankelijk Nederlands werd gegeven, gelet op de doelgroep. Die doelgroep bestond bij de meeste methodes uit hbo-studenten, maar soms was de doelgroep breder of zelfs onbepaald. Toch werd er bij de uitleg vanuit gegaan dat de doelgroep een bepaalde basis had op het gebied van met name grammatica en dat had tot gevolg dat de uitleg vaak niet volledig en duidelijk genoeg was voor studenten die deze basis ontbeerden. Zelfs wanneer een methode zich expliciet ook richtte op studenten die het Nederlands niet als moedertaal hadden, was daarvan bij de uitleg weinig te merken. De impliciete aanname bleef dat deze voorkennis aanwezig was.

Onder oefenboek scoorde het criterium relevantie het hoogst en daarna volledigheid. De oefeningen hadden dus relatief vaak betrekking op de behandelde theorie. De gemakkelijkheid varieerde: het kon soms veel tijd kosten om een oefening te maken.

Veelheid en geordendheid scoorden het laagst. Soms leek in totaal veel oefenmateriaal beschikbaar te zijn, maar kon dat per onderdeel tegenvallen. Bij geordendheid kregen zes methodes zelfs een zeer slechte beoordeling. Soms waren de oefeningen wel enigszins geordend op grond van hun moeilijkheidsgraad, maar het was opvallend, dat weinig opbouw

viel te ontdekken in het niveau van de zinnen. Dat was laag en bleef dat dan ook of het was meteen vrij hoog. Gradatie zou juist goed mogelijk zijn door niet alleen het aantal in te vullen vormen op te voeren, maar het niveau van de zinnen zou ook kunnen oplopen in moeilijkheidsgraad. Hierbij valt te denken aan oefeningen in simpele, enkelvoudige zinnen in het begin en daarna geleidelijk in meer complexe zinnen. In samengestelde zinnen die studenten zelf schrijven, worden juist veel fouten gemaakt.

Het feedbackmiddel ontbrak bij vijf methodes volledig. Bij de methodes die wel feedback gaven, scoorden veelheid, betrouwbaarheid en duidelijkheid het hoogst. Dit betekent dat als er feedback gegeven werd, die vaak bij alle opdrachten voorkwam, de feedback klopte dan ook en maakte doorgaans duidelijk of de opdracht goed gemaakt was.

Snelheid, afhankelijkheid en voortgangsinformatie waren problematisch. Antwoorden moesten achterin het boek worden opgezocht en wie per gemaakt item het antwoord wou weten, zag meteen ook de volgende antwoorden. Nergens waren de antwoorden zo vermeld dat het volgende antwoord niet onmiddellijk zichtbaar was, terwijl dat ook bij papieren methodes wel mogelijk is. De beoordeling 'zeer slecht' kregen tien methodes bij snelheid en acht methodes bij voortgangsinformatie. Voortgangsinformatie ontbrak in veel gevallen of was beperkt aanwezig. In sommige methodes waren wel toetsen opgenomen, maar dan ontbrak daarbij de norm. De student kon op die manier niet vaststellen wanneer hij de stof voldoende beheerste.

In Tabel 5.5 zijn de gemiddelden per methode vermeld voor de criteria informatiebasis, oefenboek en feedbackmiddel die ontleend zijn aan het beoordelingsschema voor studieteksten, uitgaande van het ABC-model voor effectief leren (Van Es, 1985). De hoogste gemiddelden komen voor bij informatiebasis. Hoewel de uitleg in het algemeen niet volledig en duidelijk was, bleken papieren methodes toch het meest geschikt om uitleg te geven. Vervolgens is het gemiddelde van deze drie hoofdpunten vastgesteld. Dit heeft een rangordening opgeleverd van beste naar slechtste methode.

Tabel 5.5 Papieren methodes gerangordend op basis van totaal gemiddelde

	Methode	Informatie	Oefenboek	Feedback	Totaal
1	*Basisvaardigheden Spelling*	4,4	4,0	3,5	4,0
2	*Basisboek Spelling*	4,2	4,2	3,2	3,9
3	*Taaltopics Spelling*	4,0	3,8	3,2	3,7
4	*Basisvaardigheden Taal*	3,6	3,6	4,0	3,7
5	*Vlekkeloos Nederlands*	4,0	3,8	3,3	3,7
6	*Welgespeld*	3,8	4,0	3,3	3,7
7	*Praktische cursus spelling*	3,8	3,8	3,2	3,6
8	*Een goede Spelling*	4,2	3,4	3,3	3,6
9	*Dat d/t gedoe*	3,8	2,8	3,2	3,3
10	*Van verslag tot rapport*	3,4	2,8	3,8	3,3
11	*Praktische cursus Formuleren*	4,0	2,8	2,2	3,0
12	*Taaltopics Formuleren*	4,2	1,4	3,2	2,9
13	*Spelbewust*	4,2	3,8	0,0	2,7
14	*Repetitieboekje Nederlands*	3,4	4,0	0,0	2,5
15	*Gewoon goed Nederlands*	4,0	2,0	0,0	2,0
16	*Formuleren*	3,0	1,4	0,0	1,5
17	*Nederlands in je pocket*	3,4	0,0	0,0	1,1
	Gemiddelde	**3,6**	**3,2**	**2,3**	**3,0**

Methodes waarbij het feedbackmiddel volledig ontbrak, staan onderaan in de lijst. Zo eindigde Het *Repetitieboekje Nederlands* laag, maar het had een van de hoogste gemiddelden voor oefenboek. Afgezien van het ontbreken van feedback, was het bij een paar van deze methodes te verwachten dat ze laag zouden eindigen. Ze richtten zich op een afwijkende doelgroep, ze hadden een andere doelstelling of de methode had toch meer het karakter van een naslagwerk. Een voorbeeld van het laatste is *Nederlands in je pocket* dat helemaal onderaan staat met een gemiddelde van 1.1. Wat informatiebasis betreft, scoorde deze methode helemaal niet slecht, behalve op het criterium relevantie. De uitleg kon met het oog op de oefeningen ook niet relevant zijn, want het oefenmateriaal ontbrak volledig. Toch kan dit boek waarde hebben als naslagwerk en zo is het ook bedoeld. De reden om het hier te analyseren is bij de beschrijving van de methode zelf opgenomen. *Formuleren* staat met een gemiddelde van 1.5 op de een na laatste plaats. Deze methode zou onder andere adviezen geven voor helder taalgebruik, waarbij het gaat om het effect op de inhoud en om die reden

was dit boek beoordeeld. Maar tijdens de beoordeling bleek dat het vooral geschikt is voor studenten taalbeheersing die geïnteresseerd zijn in de analyse van taalverschijnselen vanuit een bepaald theoretisch kader. Hoewel deze methode niet op de laatste plaats eindigde, omdat er oefenmateriaal in staat, is *Nederlands in je pocket* geschikter om uitleg te geven over elementaire vaardigheden.

Basisvaardigheden Spelling voor de pabo staat bovenaan, met een gemiddelde over het totaal van een 3.8. Volgens de beoordeling zou deze methode het meest geschikt zijn om aan studenten voor te leggen bij het leren van elementaire vaardigheden. De andere methodes hebben een lager gemiddelde, hoewel het verschil met de drie methodes die de tweede plaats delen minimaal is. Toch bleek bij de analyse van het boek *Basisvaardigheden Spelling voor de pabo* dat de student basiskennis moest hebben om de uitleg te kunnen volgen, dat de hoeveelheid oefeningen matig was en dat die oefeningen bovendien nauwelijks geordend zijn op grond van hun moeilijkheidsgraad. Verder was van snelle feedback geen sprake en de feedback kon gemakkelijk op het verkeerde moment komen. Voortgangsinformatie was slechts voor een deel aanwezig. Ondanks deze nadelen is *Basisvaardigheden Spelling voor de pabo* in vergelijking met de andere methodes de beste methode en vrij geschikt voor zelfstandig gebruik.

5.3.2 Beoordeling digitale taalmethodes

De digitale methodes zijn in de onderstaande tekst en in de bijlage alfabetisch gerangschikt op naam van de methode. Een ordening op basis van de auteur was niet goed mogelijk, omdat de ontwikkelaar van digitale methodes in meerdere gevallen niet expliciet genoemd wordt. De beoordeling van alle digitale methodes aan de hand van het beoordelingsschema voor studieteksten staat in bijlage 6. Op basis daarvan is het onderstaande oordeel over de digitale methodes tot stand gekomen.

Bij de digitale methodes zijn in bijlage 6 ook de maand en het jaar vermeld waarin de beoordeling plaatsvond. Digitale programma's kunnen namelijk snel bijgesteld worden, zodat het mogelijk is dat in korte tijd verschillende versies beschikbaar zijn.

Van de digitale programma's is alleen het standaardprogramma beoordeeld. Digitale methodes bevatten soms mogelijkheden voor de docent om bepaalde knoppen uit of aan te zetten,

wat weer gevolgen heeft voor de werking van het programma. Deze mogelijkheden zijn buiten beschouwing gebleven.

Cambiumned

Cambiumned is een gratis oefenwebsite voor het vak Nederlands, onder andere over literatuur, poëzie, taalspelletjes en over basisvaardigheden. De website is gemaakt voor scholieren op havo en vwo. Het doel is niet verwoord. Deze site wordt beoordeeld, omdat de Taalwinkel van de Universiteit en Hogeschool van Amsterdam studenten met taalproblemen ernaar verwijst. Hier werd gekeken naar de spelling van werkwoorden, het gebruik van voornaamwoorden en interpunctie, die vallen onder de noemer *Alle oefeningen op Cambiumned*.

Als informatiebasis voldoen *Alle oefeningen op Cambiumned* tot op zekere hoogte. De uitleg over de spelling van werkwoorden, voornaamwoorden en interpunctie is correct en de informatie wordt ook in toegankelijk Nederlands gegeven voor iemand die op de hoogte is van grammaticale begrippen. Maar de summiere informatie over het bijvoeglijk naamwoord dat van het voltooid deelwoord is afgeleid, kan ontoereikend zijn. Eigenaardig is dat de student de mogelijkheid heeft om alle uitleg te negeren. De informatie heeft hier kennelijk weinig relevantie. Standaard wordt namelijk geen uitleg gegeven, maar alleen als daar bij het maken van een oefening om gevraagd wordt. Op zo'n moment kan de uitleg wel relevant zijn, maar hier is dat slechts gedeeltelijk het geval. De student krijgt namelijk meteen alle informatie over het onderwerp en daaruit moet hij de uitleg toegespitst op dat ene item zelf zoeken.

Als oefenboek zijn *Alle oefeningen op Cambiumned* behoorlijk geschikt. Bij de werkwoordsvormen zijn relatief veel opdrachten beschikbaar. Opvallend in positieve zin is, dat die ook zo gekozen kunnen worden dat ze oplopen in moeilijkheidsgraad. Bij de voornaamwoorden is het oefenmateriaal beperkt en bij interpunctie is dat helemaal het geval. Maar ook bij de voornaamwoorden is sprake van enige ordening van de opdrachten op grond van hun niveau. De meeste oefeningen zijn vrij snel te maken. Wat tijd kan kosten, is het doorgaans onrustige oefenscherm.

Als feedbackmiddel is *Cambiumned* met betrekking tot de beoordeelde oefeningen voor een

deel zeker geschikt. Bij ieder item wordt feedback gegeven en het is ook mogelijk om die snel te krijgen. Positief is, dat alleen feedback gegeven wordt, nadat het goede antwoord gegeven is. De student kan net zolang doorgaan met oefenen, totdat hij alles goed gemaakt heeft. Het goede antwoord wordt niet eerder prijsgegeven. Nadelig is dat de feedback bij de voornaamwoorden niet altijd klopt en ook maakt de feedback in het algemeen niet zonder meer duidelijk of een item goed gemaakt is. Voortgangsinformatie wordt op een aantal manieren gegeven. Bij de werkwoordsvormen is het duidelijk welke oefeningen bij hetzelfde onderwerp nog volgen, maar bij de voornaamwoorden en interpunctie is dit niet in een oogopslag te zien. De oefeningen staan daarbij wat verstopt tussen opdrachten over andere onderwerpen. Oefeningen van verschillend niveau moet de student ook zelf opzoeken op de website. Tijdens het maken van een oefening is het totale aantal items of zinnen zichtbaar en het is duidelijk waar de student is. De score wordt eveneens gegeven, overigens zonder dat de norm bekend is, na afloop van de opdracht of meteen na ieder item als om 'nakijken' gevraagd wordt. Maar de behaalde resultaten worden niet bewaard. Bij de twee toetsen die beschikbaar zijn over de spelling van werkwoorden, wordt op dezelfde manier de score gegeven. Vreemd is dat die toetsen slechts voor een deel gaan over werkwoordsvormen, want ook algemene spellingzaken worden getoetst. Bij een slechte score is onduidelijk wat de student moet doen. Het programma stuurt dus op een bepaalde manier bij het maken van een oefening of toets, omdat net zolang doorgeoefend kan worden totdat alles goed gemaakt is. Verder is van sturing geen sprake.

Over het geheel genomen is dit programma redelijk geschikt voor zelfstudie.

dtkompas

De naam van de gratis oefenwebsite *dtkompas* maakt meteen duidelijk dat het om de spelling van werkwoorden gaat. Over de doelgroep wordt op de website niets aangegeven. Het doel van *dtkompas* is om in korte tijd grammaticaregels te veroveren en veilig te reizen door de zee van *d* en *t*. De reden om dit programma hier te beoordelen, is dat de Taalwinkel van de Universiteit en Hogeschool van Amsterdam het geschikt vindt voor studenten met spellingproblemen.

Als informatiebasis is *dtkompas* alleen beperkt bruikbaar voor iemand met enige gramma-

ticale kennis. De uitleg is correct en onnodig moeilijke omschrijvingen worden vermeden. Maar de term 'dt-werkwoord' kan aanleiding geven tot misverstanden. Ook ontbreekt uitleg bij het bijvoeglijk naamwoord dat van het werkwoord is afgeleid. Het is jammer dat de spelling van werkwoorden hier niet iets ruimer wordt opgevat, omdat de juiste schrijfwijze van dit bijvoeglijk naamwoord in de praktijk problematisch blijkt te zijn. Verder is de uitleg niet relevant, omdat uitleg en oefeningen van elkaar gescheiden zijn. Het is op die manier mogelijk om alle uitleg te negeren. Maar ook als dit niet gebeurt, kan de informatie tijdens het oefenen alweer vergeten zijn en die is vervolgens niet snel te raadplegen. Het is ook overbodig dat twee keer uitleg over hetzelfde wordt gegeven in andere bewoordingen. Eén keer iets duidelijk uitleggen is minder gecompliceerd en is daarom te verkiezen.

Als oefenboek is het eveneens beperkt geschikt. De oefeningen in de vorm van tests zijn weliswaar redelijk snel te maken, maar door het gebruik van animaties wordt het oefenscherm onnodig onrustig en dat kan tijd kosten. Verder is het oefenmateriaal ontoereikend, wat hoeveelheid betreft en ook wat het niveau aangaat. Met het bijvoeglijk naamwoord wordt helemaal niet geoefend, maar het programma is wel consequent, want daarbij ontbreekt eveneens de uitleg. Alleen simpele oefeningen zijn beschikbaar. Een ordening in moeilijkheidsgraad valt nauwelijks te ontdekken en dan ook alleen in omgekeerde volgorde. De moeilijkste test waarin alle werkwoordsvormen voorkomen, staat bovenaan en daarna volgen de gemakkelijke. Maar ook bij de eerste test is het niveau van de oefenzinnen heel simpel. Het is onwaarschijnlijk dat dit materiaal toereikend zou kunnen zijn voor studenten in het hoger onderwijs.

Als feedbackmiddel voldoet *dtkompas* matig. Als de student kiest voor een test met alleen de score, dan krijgt hij helemaal geen feedback bij zijn antwoorden en ontbreekt de feedback die digitale programma's normaal wel geven. Hij kan ook kiezen voor een test met antwoorden en toelichting en krijgt in dat geval snel feedback met uitleg toegespitst op het item. Maar door de omslachtigheid van de reactie is die feedback niet in een oogopslag duidelijk. Het goede antwoord wordt ook veel te gemakkelijk weggegeven. Het wordt al zichtbaar als de student niets invult en vervolgens op 'verder' klikt. Voortgangsinformatie is tot op zekere hoogte beschikbaar. Tijdens het oefenen wordt bijgehouden bij welk item de student is en het is duidelijk hoeveel items nog volgen. Het is ook gemakkelijk te zien hoeveel opdrachten beschikbaar zijn. Na afloop van een oefening wordt wel de score gegeven, maar de norm is niet bekend. Resultaten worden niet bewaard. Bij veel foute antwoorden komt het advies om nog maar eens wat te oefenen. Wat de student precies moet ondernemen om

een voldoende te halen, is niet duidelijk. Het programma stuurt de student nergens. Het is ook niet duidelijk wanneer hij de stof voldoende beheerst.

Door de genoemde nadelen is het programma niet geschikt voor zelfstudie.

Hogeschooltaal

Hogeschooltaal wordt gepresenteerd als een individueel online-taalvaardigheidsprogramma, dat bedoeld is voor studenten in het hoger onderwijs. Over het niveau van de doelgroep wordt niets vermeld. Het doel van het programma is niet expliciet verwoord. Deze analyse heeft betrekking op het deel van de *Basismodule Nederlands* voor het hbo dat gaat over de spelling van werkwoorden, interpunctie en het gebruik van voornaamwoorden.

Als informatiebasis is *Basismodule Nederlands* van *Hogeschooltaal* voor een deel bruikbaar. Het is een voordeel dat een doorgaans klein stukje uitleg meteen toegepast kan worden bij de oefeningen. Maar om de summiere en een enkele keer onnodig moeilijke uitleg te kunnen volgen, heeft de student kennis nodig van grammaticale begrippen. Het is de vraag of die aanwezig is. De uitleg over het bijvoeglijk naamwoord is niet relevant, want daarbij ontbreken de opdrachten. Bij de voornaamwoorden ontbreekt de uitleg van het betrekkelijk voornaamwoord, terwijl je die in een basisprogramma wel zou verwachten. In de praktijk worden er veel fouten mee gemaakt. Vrij specifieke uitleg is in een aantal gevallen snel op te roepen tijdens het oefenen, maar die mogelijkheid ontbreekt ook vaak. De uitleg die standaard wordt gegeven is soms gekoppeld aan opdrachten, maar soms ook niet en daardoor is de uitleg niet altijd snel te vinden.

Als oefenboek voldoet deze module in beperkte mate. Bij de spelling van werkwoorden is behoorlijk wat oefenmateriaal beschikbaar, maar met de voornaamwoorden en leestekens kan onvoldoende geoefend worden. Ook op een andere manier zijn er te weinig oefeningen, omdat er alleen maar simpele oefeningen zijn. De oefeningen kunnen wel redelijk snel gemaakt worden, maar het is jammer dat ze niet op vaste plaatsen te vinden zijn. De student moet wat dat betreft steeds actie ondernemen, door terug te gaan naar een beginmenu en te zoeken naar opdrachten. Een voordeel is, dat er een duidelijke link is tussen theorie en praktijk, want de oefeningen oefenen niet alleen met wat in de theorie behandeld is, maar ze volgen vaak al na een klein stukje uitleg. Die link ontbreekt echter bij het bijvoeglijk naam-

woord dat van het werkwoord is afgeleid. Bij het bijvoeglijk naamwoord zijn geen oefeningen beschikbaar, terwijl het in de praktijk vaak fout geschreven wordt. Datzelfde geldt voor het betrekkelijk voornaamwoord en dat komt noch in theorie, noch in de praktijk aan de orde. Van een opbouw van eenvoudig naar complex is bij de oefeningen in beperkte mate sprake. Bij de spelling van werkwoorden en interpunctie komen bij de laatste opdracht wel meerdere vormen terug, maar het niveau is nog steeds laag. De zinnen waarin de werkwoordsvormen voorkomen, zijn vrijwel steeds kort en simpel. Een hbo-student gebruikt in zijn eigen zakelijke teksten complexere zinnen en daarbij gaat het spellen van de werkwoordsvormen juist vaak fout.

Als feedbackmiddel heeft deze module van *Hogeschooltaal* voordelen, maar ook duidelijke nadelen. Bij ieder item wordt feedback gegeven in de vorm van het goede antwoord, maar die maakt niet in een oogopslag duidelijk hoe het item gemaakt is. Echt snel is de feedback ook niet, want zelfs als na één gemaakt item om feedback wordt gevraagd, laat de reactie even op zich wachten. Wanneer het goede antwoord verschijnt, gaat het ook gemakkelijk mis. Niet alleen wordt het goede antwoord zichtbaar van het gemaakte item, maar van alle items uit die zin. Bovendien kan de student die goede antwoorden al zien, als hij een willekeurige letter intypt en dan op het icoon voor opslaan klikt. Dit maakt het voor de student verleidelijk om naar het goede antwoord te kijken, voordat het eigen antwoord gegeven is. Uitleg kan bij een item gevraagd worden, maar die is meer een herhaling van een stukje theorie. Bij voorgestructureerde antwoorden zou uitleg toegespitst op het gemaakte item zelf, wel tot de mogelijkheden behoren. Voortgangsinformatie krijgt de student op een paar manieren. Hij kan zien welke onderwerpen behandeld worden, maar het is niet goed te zien hoeveel opdrachten er precies zijn. Tijdens het maken van een oefening kan hij te weten komen hoeveel hij nog moet doen en na afloop hoeveel fouten hij gemaakt heeft, maar dat gaat niet automatisch. De student moet deze dingen allemaal zelf uitzoeken. Op een persoonlijke pagina wordt bijgehouden welke opdrachten gemaakt zijn en wat de resultaten zijn, maar daarbij is de informatie beperkt. Alleen bij de taaltoets wordt een duidelijke score vermeld, want verder staat uitsluitend aangegeven dat sprake is van zelfcorrectie. Standaard is één taaltoets beschikbaar. Maar als de student geen voldoende haalt, wordt niet aangegeven welke actie hij dan moet ondernemen. Het programma stuurt de student niet.

Deze basismodule is alles in aanmerking genomen alleen matig geschikt om zelfstandig gebruikt te worden.

Juf Melis

De gratis website van *Juf Melis* bevat oefenmateriaal voor spelling en grammatica. De doelgroep bestaat uit iedereen die wil oefenen met de Nederlandse taal. Het doel is spelling en grammatica gemakkelijker te maken. Hier werd *Werkwoordspelling* beoordeeld. Het onderdeel *Spelling* kwam in beeld bij het bijvoeglijk naamwoord dat van het voltooid deelwoord is afgeleid en dat bij deze oefensite niet bij de werkwoordsvormen is behandeld.

Over de informatiebasis van *Juf Melis* valt niets te zeggen, want bij de spelling van werkwoorden ontbreekt de uitleg.

Als oefenboek voldoen *Werkwoordspelling en Spelling* alleen in beperkte mate. De oefeningen zijn vrij snel te maken. Met de persoonsvorm kan redelijk geoefend worden en dat gaat ook nog op voor het voltooid deelwoord. Maar bij het bijvoeglijk naamwoord dat van het voltooid deelwoord is afgeleid, zijn amper oefeningen beschikbaar. Het oefenmateriaal is ook op een andere manier ontoereikend, omdat uitsluitend simpele oefeningen gemaakt kunnen worden. Van een bepaalde ordening van de oefeningen op grond van hun moeilijkheidsgraad is nauwelijks sprake. Bij de latere opdrachten rond een specifieke werkwoordsvorm zijn de oefenzinnen wel iets complexer, maar het verschil is niet substantieel. Bovendien kan bij iedere opdracht met tien zinnen maar met één vorm tegelijk geoefend worden en dan krijgt het antwoord iets voorspelbaars.

Als feedbackmiddel voldoet *Juf Melis* matig. Bij ieder item wordt weliswaar betrouwbare feedback gegeven, maar die maakt niet direct duidelijk hoe het item gemaakt is. Snel is de feedback alleen als per item wordt nagekeken, maar doordat een opdracht helemaal op het scherm staat, zal de neiging bestaan om die eerst volledig te maken. De feedback komt in beginsel pas nadat het antwoord gegeven is, maar dit gaat niet meer op als de optie 'zeg voor' gebruikt wordt. Dan komen de antwoorden van alle items van een opdracht in een keer in beeld. Deze optie is kennelijk toegevoegd voor iemand die vastloopt, maar nu wordt alles prijsgegeven. Voortgangsinformatie wordt gegeven over hoeveel items nog gedaan moeten worden van een oefening, hoe de score is en welke opdrachten nog volgen. Resultaten worden echter niet bewaard. Toetsen zijn afwezig. Wat de student moet doen bij slechte resultaten is niet duidelijk.

Hoewel *Juf Melis* nog matig fungeert als feedbackmiddel is het over het geheel genomen slecht te gebruiken voor zelfstudie.

Muiswerk

Muiswerk brengt standalone programma's op de markt voor het basis tot en met het hoger onderwijs. Het niveau van de doelgroep is divers. Het doel wordt niet genoemd. Hier zijn de programma's *Spelling werkwoorden*, *Leestekens* en *formuleren* voor het hoger onderwijs beoordeeld.

Als informatiebasis voldoen de *Muiswerk*-programma's tot op zekere hoogte. De uitleg wordt gegeven in begrijpelijk Nederlands, waarbij weinig voorkennis vereist is. In het algemeen is de uitleg correct, maar een uitzondering is een voorbeeld met uitleg over het gebruik van de komma in zinnen. Bij de voornaamwoorden is het opvallend, dat niet specifiek iets gezegd wordt over de betrekkelijke voornaamwoorden. Een juist gebruik daarvan is in de praktijk niet vanzelfsprekend. Het is verder de vraag hoe relevant de informatie is, want eerst wordt alle uitleg gegeven en die kan vergeten zijn op het moment dat er geoefend wordt. Maar tijdens het oefenen kan de informatie ook geraadpleegd worden en dan is die niet overbodig. Een nadeel daarbij is dat wie specifiek naar iets zoekt, alle uitleg weer bij langs moet gaan over een bepaald onderwerp en dat gaat niet snel.

Als oefenboek zijn ze gedeeltelijk geschikt. De hoeveelheid oefenmateriaal is redelijk bij de werkwoorden en leestekens. Met de voornaamwoorden kan minder geoefend worden. Een nadeel is ook dat oefeningen met het betrekkelijk voornaamwoord ontbreken, terwijl het gebruik daarvan in de praktijk problemen geeft. Vanuit een andere invalshoek bekeken is het oefenmateriaal over de hele linie ontoereikend, omdat alleen vrij simpel oefenmateriaal beschikbaar is. De oefeningen zijn redelijk snel te maken door het intypen van een woord en soms is sprake van meerkeuzemogelijkheden en dan gaat het echt snel. Verder kan de student werken op een rustig scherm met één item per keer. Ook vanwege hun vrij simpele karakter zijn de oefeningen snel te maken. Het komt voor dat aan het begin van de opdracht al staat aangegeven dat bijvoorbeeld steeds stam + t ingevuld moet worden. Op zich is daar in de beginfase niets op tegen, maar dergelijke oefeningen worden hier niet gevolgd door opdrachten die echt meer vragen van de student. Een bepaalde opbouw in de moeilijkheidsgraad van de oefeningen valt namelijk niet te ontdekken. De oefeningen zijn in het begin vrij eenvoudig en dat blijven ze ook. Tussendoor komt af en toe een oefening voor met iets complexere zinnen, maar de systematiek daarin ontbreekt.

Als feedbackmiddel voldoen deze programma's van *Muiswerk* behoorlijk goed. Na ieder gemaakt item wordt meteen feedback gegeven en die komt ook op het goede moment, nadat het antwoord gegeven is. Maar de feedback klopt niet in alle gevallen, want bij de verwij-

zing met 'het' is de feedback niet helemaal betrouwbaar. De feedback is zonder meer duidelijk, zodat de student weet hoe hij een item gemaakt heeft, maar de wisselende omschrijvingen van of iets goed of fout was, zijn overbodig en leiden af. Na een fout antwoord volgt steeds uitleg, die doorgaans toegespitst is op het item zelf. Voortgangsinformatie wordt op verschillende manieren gegeven met behulp van diagnostische toetsen en met informatie bij de opdrachten zelf, die duidelijk maakt in hoeverre de stof beheerst wordt. Op een persoonlijke pagina worden de resultaten bijgehouden en iedere keer dat de student opnieuw inlogt, krijgt hij ook te zien hoe lang hij al daadwerkelijk met het programma bezig is geweest. Dat laatste kan verhelderend zijn voor iemand die niet goed kan inschatten hoeveel tijd hij aan iets werkt. Na de diagnostische toets krijgt de student op basis van zijn score een aantal oefeningen geselecteerd. Bij een bepaalde opdracht weet de student hoeveel items hij nog moet doen en hij kan in het menu zien hoeveel opdrachten daarna nog volgen. De norm is bij de opdrachten bekend en na iedere opdracht wordt de score gegeven. Opdrachten die voldoende gemaakt zijn, verdwijnen van het scherm. Bij een onvoldoende resultaat volgt het advies om de opdracht nog maar eens te doen. Eindtoetsen ontbreken. Als alle opdrachten voldoende gemaakt zijn, kan opnieuw een diagnostische toets gemaakt worden, waarbij de items in een andere volgorde gepresenteerd worden en deels nieuw zijn. Dan start de cyclus opnieuw. Op grond van de score zal weer een bepaald oefenadvies volgen, maar het is de vraag in hoeverre dan nog geoefend kan worden met onbekende items. Opvallend is dat het programma de student aan de ene kant stuurt, maar aan de andere kant ook vrijblijvend is. Na een onvoldoende opdracht kan de student immers verder gaan met iets anders, want hij krijgt niet automatisch op grond van de onvoldoende opdracht soortgelijke oefenstof aangeboden. Die vrijblijvendheid doet zich ook in de beginfase voor als de student zelf kan kiezen voor een diagnostische toets met daaraan gekoppeld oefeningen of de keuze heeft om meteen oefeningen te gaan maken die niet speciaal voor hem geselecteerd zijn. De student wordt op deze manier voor een deel door het programma geleid en voor een deel kan hij helemaal zelf uitmaken wat hij doet.

Als zelfstudiemateriaal voldoet dit programma alles bij elkaar genomen heel redelijk.

Nedercom

Nedercom bevat diverse online-taalvaardigheidsprogramma's bestemd voor het basisonderwijs tot en met het hoger onderwijs. De programma's *Spelling 3* en *Formuleren 3* zijn beide bedoeld voor het hoger onderwijs. Het niveau van de doelgroep kan verschillend zijn.

Het doel van *Nedercom* is om kennis en beheersing van het Nederlands te laten vergroten. Deze analyse had betrekking op de spelling van werkwoorden, interpunctie en het gebruik van voornaamwoorden.

Als informatiebasis zijn de programma's van *Nedercom* goed bruikbaar voor studenten die al wat kennis hebben van grammaticale begrippen. De uitleg wordt in begrijpelijk Nederlands gegeven en is via de oefeningen gemakkelijk te raadplegen. Maar op een bepaalde manier is de uitleg overbodig, want die kan namelijk volledig genegeerd worden.

Als oefenboek zijn de programma's op een bepaalde manier zeker bruikbaar. Bij de spelling van werkwoorden is behoorlijk veel oefenmateriaal beschikbaar, met uitzondering van het bijvoeglijk naamwoord dat van het voltooid deelwoord is afgeleid, want daarbij staan relatief weinig oefeningen. Met de voornaamwoorden en leestekens kan redelijk geoefend worden. Maar als gelet wordt op het niveau, dan is het oefenmateriaal ontoereikend. Uitsluitend simpele oefenzinnen zijn beschikbaar. Aan het begin van de opdracht is het al duidelijk dat bijvoorbeeld steeds de persoonsvorm ingevuld moet worden. Bij een groot aantal opdrachten met het voltooid deelwoord, wordt het antwoord zelfs weggegeven, omdat al duidelijk is welke letter ingevuld moet worden. Een voordeel is dat de oefeningen in het algemeen vrij snel te maken zijn, hoewel het intypen van zinnen niet echt snel gaat. Het gemakkelijke karakter van de opdrachten werkt ook tijdbesparend, evenals het rustige oefenscherm met steeds één item per keer. Maar een eenmaal gemaakte keuze kan niet snel veranderd worden. Een ordening van de oefeningen op grond van hun moeilijkheidsgraad is alleen in beperkte mate te ontdekken. Uiteindelijk worden wel een paar gemengde opdrachten gegeven waarin naar meerdere vormen gevraagd wordt, maar het niveau van de zinnen blijft simpel.

Als feedbackmiddel zijn de programma's van *Nedercom* voor een belangrijk deel geschikt. Bij ieder gemaakt item wordt onmiddellijk feedback gegeven, die in het algemeen betrouwbaar en begrijpelijk is voor de student. Een uitzondering doet zich voor bij de leestekens in zinnen. Goede antwoorden worden daar soms fout gerekend, als een niet-relevant deel van de zin wordt weggelaten. Het vergelijken van het eigen antwoord met het goed gerekende antwoord is bij de leestekens in zinnen bovendien lastig, want het eigen antwoord valt vaak meteen weg, zodra feedback gegeven wordt. De uitleg die zowel bij een goed als een fout antwoord gegeven wordt, roept soms alleen vragen op en had dan beter achterwege kunnen blijven. In een aantal gevallen is het namelijk niet gelukt om de uitleg te laten passen bij het item zelf. Voortgangsinformatie is aanwezig in de vorm van diagnostische toetsen en zelf-

toetsen. Maar de items voor die toetsen worden geselecteerd uit de opdrachten. Er is dus geen sprake van nieuwe, gelijksoortige items, maar van identieke items. Aan het begin van een bepaald onderdeel kan een diagnostische toets gemaakt worden, maar dat is niet noodzakelijk. Na de toets volgt een oefenadvies, maar het is onduidelijk waarop dat advies gebaseerd is. De norm en de behaalde score ontbreken namelijk. Ook bij de opdrachten wordt voortgangsinformatie gegeven, want het is duidelijk hoeveel items nog gemaakt moeten worden van een bepaalde opdracht en hoeveel opdrachten nog volgen. Ook wordt bijgehouden hoeveel goede en foute antwoorden gegeven zijn. Na afloop van een opdracht wordt het resultaat gegeven in termen van het aantal fouten, met daaraan gekoppeld goed, voldoende of onvoldoende. Maar de norm wordt niet expliciet vermeld. Na afloop van een serie opdrachten over een onderwerp kunnen zelftoetsen gemaakt worden. De student wordt daarin duidelijk gestuurd door het programma, want hij mag zo'n zelftoets pas doen, als hij alle opdrachten bij zo'n onderdeel voldoende heeft gemaakt. Maar de norm wordt weer niet expliciet vermeld. Wie de zelftoetst onvoldoende maakt, krijgt bovendien geen informatie over hoe hij wel op het gewenste niveau kan komen, behalve de mededeling dat hij de toets nogmaals kan maken.

Dit programma is vanwege de meerdere sterke punten behoorlijk geschikt voor zelfstudie.

Project X 2002

Project X 2002 presenteert zichzelf als een digitaal leerplatform. Op dit platform zijn onder andere *Spellingsoefeningen* te vinden. De doelgroep is breed, namelijk leerlingen en iedereen die van taal houdt. Over het doel wordt niets gezegd. Hier werd de spelling van werkwoorden beoordeeld.

Als informatiebasis voldoen de *Spellingsoefeningen* van *Project X 2002* niet. De uitleg is beperkt en om die te kunnen volgen is grammaticale kennis een vereiste. Over het bijvoeglijk naamwoord dat van het voltooid deelwoord is afgeleid wordt nauwelijks informatie gegeven. Verder zijn informatie en oefeningen van elkaar gescheiden. Echt relevant is de uitleg kennelijk niet, want de student heeft de mogelijkheid om de oefeningen te maken zonder ooit naar de uitleg te kijken. Tijdens het oefenen kan de uitleg ook niet opgeroepen worden. De informatie is op zich snel te vinden in het menu, maar wie specifiek naar iets zoekt, verliest tijd door de rommelige indeling bij de uitleg.
Als oefenboek zijn de *Spellingsoefeningen* van *Project X 2002* matig geschikt. Vrij weinig

oefenmateriaal is beschikbaar. Met de persoonsvorm kan nog het meeste geoefend worden, maar bij het voltooid deelwoord is dat al beperkter. Bij het bijvoeglijk naamwoord zijn amper oefeningen te vinden. De oefeningen zijn wel vrij snel te maken, ook doordat een eenmaal gemaakte keuze gemakkelijk gewijzigd kan worden en het niveau van de oefeningen simpel is. Van een ordening van de oefeningen op grond van hun moeilijkheidsgraad is enigszins sprake. Eerst wordt één vorm geoefend en daarna komen verschillende vormen tegelijk terug in een paar opdrachten. In het licht van die ordening is het vreemd, dat de laatste gemengde opdracht gemakkelijker is dan de twee eerdere opdrachten met alle vormen door elkaar.

Als feedbackmiddel voldoet dit programma van *Project X 2002* voor een deel behoorlijk. Bij ieder item wordt betrouwbare feedback gegeven en die komt pas nadat het antwoord gegeven is. Maar de feedback maakt niet in een oogopslag duidelijk of een item goed gemaakt is. Snel is de feedback alleen als per item wordt nagekeken, maar de opdracht staat in z'n geheel op het scherm en doorgaans zullen eerst alle items gemaakt worden. Voortgangsinformatie wordt bij de opdrachten gegeven in die zin, dat de student kan zien hoeveel hij nog moet doen van een bepaalde opdracht en hij weet ook welke oefeningen nog volgen. Bij iedere opdracht staat het behaalde resultaat vermeld, maar de resultaten worden niet bewaard. Toetsen zijn er niet. Bij een opdracht zelf kan hij zijn antwoorden verbeteren, maar door de beperkte opties bij een specifieke werkwoordsvorm is dat meestal niet moeilijk. Op grond daarvan is het niet duidelijk of hij de stof wel voldoende beheerst.

Dit programma is op basis van de genoemde nadelen alleen matig geschikt voor zelfstudie.

Studiemeter

Viastarttaal is een methode voor basale taalvaardigheden die hoort bij de digitale onlineprogramma's van *Studiemeter.* Deze methode is bestemd voor mbo-studenten, maar bij de variant voor het hoogste niveau is weinig verschil te ontdekken met digitale programma's voor hbo-studenten. Over het doel van de methode wordt niet expliciet iets aangegeven. Interpunctie valt in dit programma kennelijk niet onder basale vaardigheden, want aan dit onderdeel wordt geen aandacht besteed. Hier werd gekeken naar de spelling van werkwoorden en stijlkwesties, zoals het gebruik van voornaamwoorden.
Als informatiebasis voldoet *Viastarttaal* van *Studiemeter* niet. Hoewel de uitleg correct is,

moet de student wel grammaticale kennis hebben om die te kunnen volgen. Bij de voornaamwoorden ontbreekt uitleg. Alleen bij de spelling van werkwoorden wordt uitleg gegeven en die is summier en over het bijvoeglijk naamwoord onvolledig. Relevant is de informatie niet te noemen. Uitleg en oefeningen zijn namelijk volledig van elkaar gescheiden, zodat de informatie vergeten kan zijn op het moment van oefenen of zelfs helemaal niet gebruikt wordt. Tijdens het oefenen kan de uitleg niet opgeroepen worden. In het menu is de uitleg bovendien niet snel te vinden.

Als oefenboek is dit programma voor een deel redelijk bruikbaar. Bij de spelling van werkwoordsvormen zijn de meeste opdrachten beschikbaar en het gaat om oefeningen van enig niveau. De oefenzinnen zijn doorgaans samengesteld en verschillende werkwoordsvormen worden door elkaar geoefend. Wat opvalt, is dat simpele oefeningen ontbreken. Bij de voornaamwoorden zijn minder oefeningen beschikbaar en het niveau is vrij eenvoudig. Wat het niveau betreft, is het oefenmateriaal dus over de hele linie ontoereikend. Hoewel steeds een woord ingetypt moet worden, zijn de oefeningen vrij snel te maken. Wat daarbij ook meehelpt, is het rustige scherm, met één oefenzin per keer. Maar wat bepaald niet rustig is en afleidt, zijn de voortdurend bewegende beelden bovenin het scherm. Een bepaalde opbouw in moeilijkheidsgraad valt niet te ontdekken. De oefeningen met werkwoordsvormen zitten meteen op een bepaald niveau, dat in de beginfase mogelijk te hoog ligt voor sommige studenten. De oefeningen met voornaamwoorden zijn in het begin simpel en dat blijven ze ook.

Als feedbackmiddel is *Viastarttaal* van *Studiemeter* alleen in beperkte mate geschikt. Bij ieder item wordt feedback gegeven die klopt, maar begrijpelijk en snel is de feedback niet. Pas na afloop van een gemaakte opdracht met circa 20 items komt de feedback en die is daardoor niet snel. Bovendien komt dan de oefening volledig in beeld en moet de student alle zinnen bij langs lopen om te controleren hoe hij die gemaakt heeft. Het is zo niet in een oogopslag te zien hoe een item gemaakt is. De feedback kan verder gemakkelijk op het verkeerde moment komen. Al na het intypen van één letter, kan de student verder gaan met het volgende item en uiteindelijk het goede antwoord bekijken. Voortgangsinformatie wordt op verschillende manieren gegeven, maar belangrijke informatie ontbreekt. De student krijgt namelijk geen feedback over wat hij moet doen als hij de stof onvoldoende beheerst. Bij de opdrachten bestaat de voortgangsinformatie eruit dat het duidelijk is welke opdrachten al gemaakt zijn en welke nog volgen. Tijdens het maken van een opdracht wordt aangegeven bij welke zin de student is en hoeveel oefenzinnen er in totaal zijn, maar het is niet duidelijk

hoeveel items hij moet doen. De hoeveelheid fouten wordt tijdens het maken van de opdracht ook niet bijgehouden. Bij alle opdrachten is de norm bekend en de score wordt na afloop vermeld. Het is daarbij opvallend dat de norm bij stijloefeningen soepeler is dan bij de oefeningen met werkwoordsvormen. De behaalde resultaten worden bewaard. Bij een onvoldoende resultaat volgt geen advies. De student kan exact dezelfde opdracht opnieuw maken, waarvan hij nu de antwoorden kent, maar hij kan ook aan een volgende opdracht beginnen. Het programma laat de student daarin volledig vrij. Bij de spelling van werkwoorden zijn verschillende soorten toetsen beschikbaar. De instaptoets, oefentoets en eindtoets kunnen gemaakt worden op ieder moment dat de student dit wil. Bij de deeltoets is dat niet het geval, want die kan pas gemaakt worden, nadat een aantal oefeningen met een voldoende is afgesloten. Alle toetsen bestaan uit twintig items, wat een gering aantal is. Nadat een toets gemaakt is, worden de norm en de score van de student vermeld. Evenmin als bij de opdrachten wordt een advies gegeven, daarin stuurt het programma niet.

Studiemeter is over het geheel genomen slechts in beperkte mate geschikt voor zelfstudie.

TaalONLINE

TaalONLINE heeft op de gelijknamige website drie Nederlandse taalvaardigheidsprogramma's. De doelgroep van de programma's is breed, want het is bestemd voor iedereen die in zijn dagelijkse werk met taal te maken heeft. Het doel is om de taalvaardigheden snel, maar volledig op peil te brengen. Hier werden onderdelen uit twee programma's beoordeeld, namelijk de spelling van Nederlandse werkwoorden en leestekens uit *Basisregels van de Spelling* en de spelling van werkwoorden en het gebruik van voornaamwoorden uit het programma *Basisregels Zinsbouw & Grammatica.*

Als informatiebasis zijn de programma's van *TaalONLINE* alleen voor een deel bruikbaar. De uitleg over de spelling van werkwoorden, leestekens en voornaamwoorden is correct, maar er wordt alleen voldoende informatie gegeven voor iemand die bekend is met grammaticale begrippen. Onvoldoende is de uitleg bij het bijvoeglijk naamwoord dat van het werkwoord is afgeleid, want daarbij wordt niets gezegd over de schrijfwijze die zo kort mogelijk moet zijn en die juist in de praktijk problemen geeft. Opvallend is dat bij interpunctie in een basale cursus, zoals *Basisregels van de Spelling*, de punt en de komma in een gewone zin helemaal buiten beschouwing blijven, alsof het gebruik van die leestekens voor zich

spreekt. Hetzelfde geldt bij de voornaamwoorden in *Basisregels Zinsbouw & Grammatica*. Hierbij worden de betrekkelijke voornaamwoorden niet besproken. Relevant is de uitleg niet met het oog op de opdrachten, want eerst wordt alle uitleg gegeven en die kan weer vergeten zijn op het moment dat de oefeningen komen. Tijdens het maken van een oefening kan specifieke uitleg wel snel geraadpleegd worden. Een nadeel is dat de aangeboden onderwerpen in een vaste volgorde doorlopen moeten worden, met het gevolg dat de uitleg bij bijvoorbeeld de spelling van werkwoorden, niet snel te vinden is. Andere spellingonderdelen staan voorop en die moeten eerst gedaan zijn.

Als oefenboek voldoen deze programma's matig. Positief is dat de oefeningen snel te maken zijn, hoewel het oefenscherm niet rustig is. De oefeningen oefenen ook met alle dingen die in de uitleg behandeld zijn. Maar het oefenmateriaal is beperkt tot een klein aantal items en met de leestekens kan zelfs nauwelijks geoefend worden. Verder is het niveau van alle oefeningen vrij simpel. Van een ordening op grond van de moeilijkheidsgraad is geen sprake. Zo worden de werkwoordsvormen aan de ene kant meteen door elkaar geoefend, maar aan de andere kant is het niveau van de zinnen in het begin eenvoudig en dat blijft ook zo. Een bepaalde opbouw valt daarin niet te ontdekken.

Als feedbackmiddel zijn de programma's van *TaalONLINE* in beperkte mate bruikbaar. Na afloop van iedere oefening wordt feedback gegeven en bij zowel goede als foute antwoorden is uitleg beschikbaar, toegespitst op het item zelf. Maar betrouwbaar is de feedback niet altijd en evenmin is de feedback duidelijk en snel. Ook kan het goede antwoord op het verkeerde moment komen, namelijk voordat het eigen antwoord volledig gegeven is. Voortgangsinformatie is op verschillende manieren beschikbaar. Bij het maken van de oefeningen kan de student gemakkelijk zien hoeveel items hij nog moet doen van een oefening en na afloop weet hij hoeveel hij goed gedaan heeft van het totale aantal items. Ook is zichtbaar welke onderdelen nog volgen. Er zijn tussentoetsen en er is een examen. De resultaten worden bewaard. Maar alhoewel de score wel steeds gegeven wordt, ontbreekt de norm bij al het materiaal. Verder valt op dat het programma sturend is, maar vaak niet op een goede manier. Een nadeel is dat de onderwerpen, die op zich verwisselbaar zijn, alleen in een vaste volgorde doorlopen kunnen worden. Als je met een bepaald onderwerp wilt oefenen, ben je daardoor verplicht om eerst de voorgaande onderdelen te doen. Na een slecht gemaakte opdracht volgt automatisch een extra oefening volgt, wat positief is. Maar het vreemde is dat de student daarna, ongeacht het resultaat, automatisch verdergaat met het programma.

Datzelfde geldt bij de tussentoetsen die uit kleine stukjes tekst met fouten bestaan en daarmee overigens anders van vorm zijn dan de oefenstof die uit meerkeuze oefeningen bestaat. Al worden die tussentoets en de herkansing slecht gemaakt, dan nog wordt de student gewoon verder door het programma geleid. Bij het examen worden een aantal onderdelen getoetst en bij een slecht resultaat volgt een herexamen dat weer andere dingen toetst. Sommige onderwerpen komen daarbij niet of nauwelijks aan de orde, zoals de spelling van werkwoorden. Bij een onvoldoende resultaat krijgt de student geen advies over wat hij kan ondernemen om dit te veranderen.

Op grond van dit oordeel over *TaalONLINE* is het niet meer dan matig geschikt om zelfstandig gebruikt te worden.

Eindbeoordeling digitale methodes

In totaal zijn negen digitale programma's beoordeeld. Evenals bij de papieren methodes ontbraken bij de digitale programma's duidelijke doelstellingen. Daarom is bij het criterium relevantie onder informatiebasis gekeken naar hoe relevant de informatie was voor het maken van de oefeningen en bij relevantie en volledigheid onder oefenboek is gekeken naar de uitleg.

Het beoordelingsschema voor studieteksten is oorspronkelijk bedoeld voor studieboeken op papier en niet voor digitale programma's. Bij een digitale methode is het belangrijk hoe een programma qua structuur in elkaar zit: hoe je er doorheen geleid wordt. Doordat een lineaire ordening van de stof vaak ontbreekt, is het toepassen van het schema op digitale programma's daardoor wel eens lastig geweest, maar het was tegelijk verhelderend voor de werking van digitale methodes. In meerdere gevallen werd de uitleg bijvoorbeeld facultatief en gescheiden van de oefeningen aangeboden. Als het programma het dan bij de oefeningen niet mogelijk maakte om die uitleg op te roepen, was de waarde van de uitleg beperkt.

In Tabel 5.6 zijn de programma's gerangordend op basis van de volgorde van bespreking. Per methode wordt per aspect de score vermeld. In Tabel 5.7 is per methode het gemiddelde per hoofdcategorie en voor het totaal van alle aspecten vermeld.
In Tabel 5.6 is te zien dat bij informatiebasis de hoogste waarden staan bij het criterium

juistheid, wat betekent dat de uitleg doorgaans correct was. De andere criteria scoorden veel lager. Bij het criterium volledigheid valt de geringe mate op waarin sommige digitale programma's dekkend waren voor de materie die ze behandelden. Bij bijvoorbeeld de behandeling van voornaamwoorden is het betrekkelijk voornaamwoord vergeten of bij de spelling (van werkwoorden) het bijvoeglijk naamwoord dat van het voltooid deelwoord is afgeleid. Dit heeft bij vijf methodes geleid tot een slechte beoordeling voor volledigheid.

Project X 2002 heeft eveneens een slechte beoordeling gekregen, omdat de uitleg op zich ver onder de maat was. Als gekeken werd hoe duidelijk de uitleg was, dan is op grond van Tabel 5.6 op te maken dat de uitleg bij drie methodes met behulp van te ingewikkelde omschrijvingen gegeven werd. Dit resulteerde in een slechte beoordeling voor duidelijkheid. Op het criterium relevantie is bij drie methodes heel laag gescoord. Als namelijk eerst alle uitleg bij een onderwerp werd gegeven, dan kon die alweer vergeten zijn op het moment dat geoefend werd. Bij deze laag scorende programma's was de uitleg niet op te roepen vanuit de oefenstof. Maar drie andere methodes kregen een goede beoordeling op dit punt. Bij toegankelijkheid hebben drie methodes een zeer slechte score gekregen. Dit betekent dat uitleg en oefeningen gescheiden werden aangeboden en dat tijdens het oefenen de uitleg niet snel te vinden was. Als eerst de oefeningen verlaten moesten worden, voordat de uitleg te raadplegen was, dan kostte dit namelijk tijd.

Bij oefenboek scoorde het criterium relevantie hoog en dat betekent dat de oefeningen oefenden met wat in de uitleg ter sprake kwam. Het criterium geordendheid scoorde opvallend laag. Bij acht van de negen programma's was alleen in beperkte mate sprake van enige ordening van de oefeningen op grond van de moeilijkheidsgraad of die ordening bleef helemaal achterwege. De uitzondering hierop was *Cambiumned*, maar daarbij moest de student wel als het ware zelf een bepaalde ordening aanbrengen, want op de site van dit programma worden de oefeningen van verschillend niveau vrijblijvend aangeboden. Maar doorgaans was het niveau van de oefeningen in het begin simpel en dat bleef ook zo. Ook het criterium veelheid scoorde relatief laag. Niet één programma kwam uit boven de waardering van matig en vijf programma's scoorden op dit punt slecht of zeer slecht. Dit had niet alleen te maken met de hoeveelheid oefenmateriaal, maar ook met het vrij lage niveau van de oefeningen, wat opvallend was bij bijna alle beoordeelde digitale programma's. Als geoefend werd met de werkwoordsvormen gebeurde dat in simpele zinnen. Vanuit die in-

valshoek bekeken was het oefenmateriaal dan ook ontoereikend. Alleen bij *Studiemeter* waren de oefeningen met werkwoordsvormen relatief lastiger. Een voordeel van het lage niveau was, dat de oefeningen daardoor relatief snel te maken waren. Het criterium gemakkelijkheid scoorde daardoor weliswaar hoger dan veelheid, maar toch kwamen zeven programma's niet boven de waardering matig uit. De snelheid veroorzaakt door het lage niveau werd dan namelijk weer tenietgedaan door andere factoren, zoals bewerkelijke items of een druk scherm. Alleen *Muiswerk* en *TaalONLINE* kregen een 4, omdat de oefeningen over de hele linie snel tot zeer snel te maken waren, hoewel bij beide programma's ook sprake was van een snelheidsbeperkende factor. Zo was het bij *Muiswerk* niet mogelijk om een eenmaal gemaakte keuze te veranderen en bij *TaalONLINE* stonden alle oefenzinnen tegelijk op het scherm. Bij volledigheid kregen drie programma's een slechte tot zeer slechte score. Maar daar stonden vier programma's met een goede tot zeer goede beoordeling tegenover die oefenden met alle dingen die nodig waren.

Bij feedbackmiddel kregen alle programma's een uitstekende beoordeling bij het criterium veelheid. Dit betekent dat bij ieder item feedback was. De feedback was in zes van de negen programma's ook betrouwbaar. Zonder meer duidelijk was de feedback bij zeven programma's niet, want die kregen een matige of slechte beoordeling. Echt snel was de feedback niet bij zes programma's, wat blijkt uit een matige of zeer slechte score. Bij het criterium afhankelijkheid valt op dat alleen een uitstekende of zeer slechte beoordeling gegeven is. Bij vijf programma's kon de feedback komen, voordat het eigen antwoord gegeven is en die scoorden zeer slecht. Bij voortgangsinformatie kregen twee van de negen programma's een goede beoordeling. Deze programma's waren voor een deel sturend en bevatten toetsen. Vijf programma's scoorden slecht tot zeer slecht op dit punt. Dit betekende dat de student niet of nauwelijks door het programma geleid werd en dat het niet duidelijk was wanneer hij de stof voldoende beheerste.

Tabel 5.6 Beoordeling digitale methodes per aspect

	Methode	Informatiebasis					Oefenboek					Feedbackmiddel					
		juist	volledig	duidelijk	relevantie	toegank	veel	gemak	relevantie	volledig	geordend	veel	betrouwb	duidelijk	snel	afhank	voortg
1	*Cambiumned*	5	3	3	3	3	3	3	5	5	4	5	3	3	3	5	2
2	*Dtkompas*	5	2	2	1	1	1	3	5	2	1	5	5	3	5	1	1
3	*Hogeschooltaal*	5	2	3	4	3	2	3	5	1	2	5	5	3	3	1	2
4	*Juf Melis*	0	0	0	0	0	2	3	3	3	1	5	5	3	3	1	1
5	*Muiswerk*	4	2	5	3	3	3	4	5	2	1	5	3	5	5	5	4
6	*Nedercom*	5	4	4	4	5	3	3	5	5	2	5	4	4	5	5	4
7	*Project X 2002*	5	2	2	1	1	2	3	5	3	2	5	5	3	3	5	2
8	*Studiemeter*	5	2	2	1	1	3	3	4	4	1	5	5	2	1	1	3
9	*TaalONLINE*	5	2	3	4	3	1	4	5	5	1	5	3	2	1	1	3
	Gemiddelde	**4,3**	**2,1**	**2,7**	**2,3**	**2,2**	**2,2**	**3,2**	**4,7**	**3,3**	**1,7**	**5,0**	**4,2**	**3,1**	**3,2**	**2,8**	**2,4**

Codering
0 = afwezig
1 = zeer slecht
2 = slecht
3 = matig
4 = goed
5 = uitstekend

Methode		*Onderdelen*
Cambiumned	-	Alle oefeningen
Dtkompas	-	-
Hogeschooltaa	-	Basismodule Nederlands
Juf Melis	-	Werkwoordspelling; Spelling
Muiswerk	-	Spelling werkwoorden, Leestekens, Formuleren

Methode		*Onderdelen*
Nedercom	-	Spelling 3, Formuleren 3
Project X 2002	-	Spellingsoefeningen
Studiemeter	-	Viastarttaal
TaalONLINE	-	Basisregels Spelling, Basisregels Zinsbouw & Grammatica

In Tabel 5.7 staat het gemiddelde voor informatiebasis, oefenboek en feedbackmiddel. Vervolgens is het gemiddelde van deze drie hoofdcategorieën bepaald. Dit heeft een rangordening opgeleverd van de beoordeelde digitale programma's.

Onderaan eindigt *Juf Melis* waarbij de uitleg niet opgeroepen kon worden tijdens het oefenen, omdat uitleg volledig ontbrak. Het oefenmateriaal was ontoereikend door hoeveelheid en niveau en de feedback kon op het verkeerde moment komen. Toetsen waren er niet. Het programma *dtkompas* was nauwelijks beter, maar het had wel uitleg. Ook al scoorde dit programma op vier van de vijf criteria onder Informatiebasis uitleg slecht tot zeer slecht, daardoor eindigt het iets hoger.

Tabel 5.7 Digitale methodes gerangordend op basis van totaal gemiddelde

	Methode	Informatie	Oefenboek	Feedback	Totaal
1	*Nedercom*	4,4	3,6	4,5	4,2
2	*Muiswerk*	3,4	3,0	4,5	3,6
3	*Cambiumned*	3,4	4,0	3,5	3,6
4	*Hogeschooltaal*	3,4	2,6	3,2	3,1
5	*TaalONLINE*	3,4	3,2	2,5	3,0
6	*Project X 2002*	2,2	3,0	3,8	3,0
7	*Studiemeter*	2,2	3,0	2,8	2,7
8	*Dtkompas*	2,2	2,4	3,3	2,6
9	*Juf Melis*	0,0	2,4	3,0	1,8
	Gemiddelde	**2,7**	**3,0**	**3,5**	**3,1**

Codering
0 = afwezig
1 = zeer slecht
2 = slecht
3 = matig
4 = goed
5 = uitstekend

Methode	*Onderdelen*
Cambiumned	- Alle oefeningen
Dtkompas	-
Hogeschooltaal	- Basismodule Nederlands
Juf Melis	- Werkwoordspelling; Spelling
Muiswerk	- Spelling werkwoorden, Leestekens, Formuleren
Nedercom	- Spelling 3, Formuleren 3
Project X 2002	- Spellingsoefeningen
Studiemeter	- Viastarttaal
TaalONLINE	- Basisregels Spelling, Basisregels Zinsbouw & Grammatica

Nedercom staat bovenaan met een gemiddelde over het totaal van een 4.2. Dit hoge gemiddelde dankte het programma aan de correcte en vrij duidelijke en specifieke uitleg die tijdens het oefenen snel op te roepen was, aan de hoeveelheid relevante oefeningen die ook redelijk snel te maken waren en aan de duidelijke en snelle feedback die pas kwam, nadat het antwoord gegeven was. Verder was dit programma gedeeltelijk sturend. Zo moest een bepaalde reeks oefeningen eerst voldoende gemaakt zijn, voordat een zogenaamde zelftoets gemaakt kon worden. *Muiswerk* staat op een gedeelde tweede plaats, samen met *Cambiumned*. *Muiswerk* leek in meerdere opzichten op *Nedercom*. Het verschil zat vooral in de uitleg die bij *Muiswerk* niet volledig was en specifieke uitleg was niet snel te raadplegen. Bij het oefenen was een nadeel dat een bepaalde soort opdrachten ontbrak die wel bij de oefenstof hoorde. *Cambiumned* scoorde opvallend goed bij het ordenen van de oefenstof van gemakkelijk naar moeilijk. Alleen brengt het programma die ordening niet echt aan, dat moet de gebruiker van de website zelf doen door oefeningen met een oplopend niveau te selecteren.

Toch had ook *Nedercom* substantiële nadelen. Sturend was dit programma alleen gedeeltelijk. Als uit de zelftoets bleek dat de stof onvoldoende beheerst werd, bleef een duidelijk traject uit hoe dan wel op het vereiste niveau te komen. De diagnostische toetsen die aan het begin van ieder onderdeel gemaakt konden worden, waren sowieso niet sturend, want ze konden genegeerd worden. Ook ontbraken norm en score bij de toetsen en daardoor was het onduidelijk waarop uitspraken over het behaalde niveau gebaseerd waren. Verder was het oefenmateriaal bij *Nedercom* vergeleken met de andere programma's weliswaar omvangrijk op het terrein van de werkwoordsvormen, maar het schoot ook te kort. Met het bijvoeglijk naamwoord dat van het voltooid deelwoord is afgeleid kon betrekkelijk weinig geoefend worden en ook op een andere manier was het oefenmateriaal ontoereikend. Alleen oefeningen van een vrij simpel niveau waren voorhanden. Hoewel sprake was van enige ordening op grond van de moeilijkheidsgraad, waren de meest lastige oefeningen waarbij verschillende werkwoordsvormen tegelijk in een opdracht voorkwamen, nog steeds vrij simpel. Dit werd met name veroorzaakt door het niveau van de zinnen die, al waren ze samengesteld, toch eenvoudig van structuur bleven. Studenten schrijven zelf samengestelde zinnen die complexer zijn en daarin maken ze juist fouten. Maar in vergelijking met de andere programma's is *Nedercom* het beste programma en het meest geschikt voor zelfstudie.

Overeenkomsten en verschillen

Aan de hand van Tabel 5.4 voor papieren methodes en Tabel 5.6 voor digitale methodes is gekeken naar overeenkomsten en verschillen tussen beide soorten methodes. Tabel 5.8 laat het verschil zien tussen de gemiddelde waarden bij papieren en digitale methodes.

In Tabel 5.8 scoort juiste uitleg onder informatiebasis bij zowel papieren als digitale methodes hoog. Bij de criteria relevantie en toegankelijkheid zijn papieren methodes beter dan digitale programma's. Als uitleg en oefeningen van elkaar gescheiden waren, had dat bij digitale programma's namelijk meer consequenties voor de relevantie van de uitleg. Digitale programma's waren ook minder toegankelijk, doordat met name specifieke uitleg vaak niet snel te vinden was. Via de inhoudsopgave van een papieren methode was dit doorgaans geen probleem, maar bij digitale programma's moesten soms meerdere schermen met informatie bekeken worden, voordat een stukje uitleg kwam wat er toe deed.

Onder oefenboek blijkt dat bij digitale programma's soms oefenonderdelen ontbraken die wel binnen de reikwijdte van de stof vielen. Op volledigheid deden papieren methodes het iets beter.

Onder feedbackmiddel scoorden digitale programma's gemiddeld belangrijk hoger bij de criteria veelheid, snelheid, afhankelijkheid en voortgangsinformatie. Alle digitale programma's gaven feedback bij iedere opdracht en die feedback kwam ook beduidend sneller dan bij de papieren methodes. Op het punt van feedback over de voortgang en de beheersing van de stof, waren digitale programma's eveneens duidelijk beter. Ook kwam de feedback vaker op het goede moment.

Tabel 5.8 Vergelijking papieren en digitale methodes per aspect

Methode	Informatiebasis					Oefenboek					Feedbackmiddel					
	juist	volledig	duidelijk	relevantie	toegank	veel	gemak	relevantie	volledig	geordend	veel	betrouwb	duidelijk	snel	afhank	voortg
Papier	4,8	2,5	2,8	3,2	4,8	2,2	3,1	4,5	4,1	1,9	3,3	3,5	3,1	1	1,7	1,1
Digitaal	4,3	2,1	2,7	2,3	2,2	2,2	3,2	4,7	3,3	1,7	5	4,3	3,1	3,2	2,8	2,4
Verschil*	**0,5**	**0,4**	**0,1**	**0,9**	**2,6**	**0,0**	**-0,1**	**-0,2**	**0,8**	**0,2**	**-1,7**	**-0,8**	**0,0**	**-2,2**	**-1,1**	**-1,3**
Beste methode**	**P**	**P**	**P**	**P**	**P**	**-**	**D**	**D**	**P**	**P**	**D**	**D**	**-**	**D**	**D**	**D**

P=papier; D= digitaal

Tabel 5.9 Vergelijking papieren en digitale methodes op hoofdaspecten

Methode	Informatiebasis	Oefenboek	Feedbackmiddel
Papier	3.6	3.2	2.3
Digitaal	2.7	3.0	3.5
Verschil*	0.9	0.1	-1.2
Beste methode	Papier	Papier	Digitaal

** De vermelde verschillen zijn berekend met 2 decimalen en kunnen daardoor afwijken van de verwachte waarden.*

In tabel 5.9 zijn de papieren en de digitale methodes vergeleken op hoofdaspecten. Het beeld dat uit de tabel naar voren komt, is dat papieren methodes beter werken als informatiebasis en digitale programma's beter als feedbackmiddel.

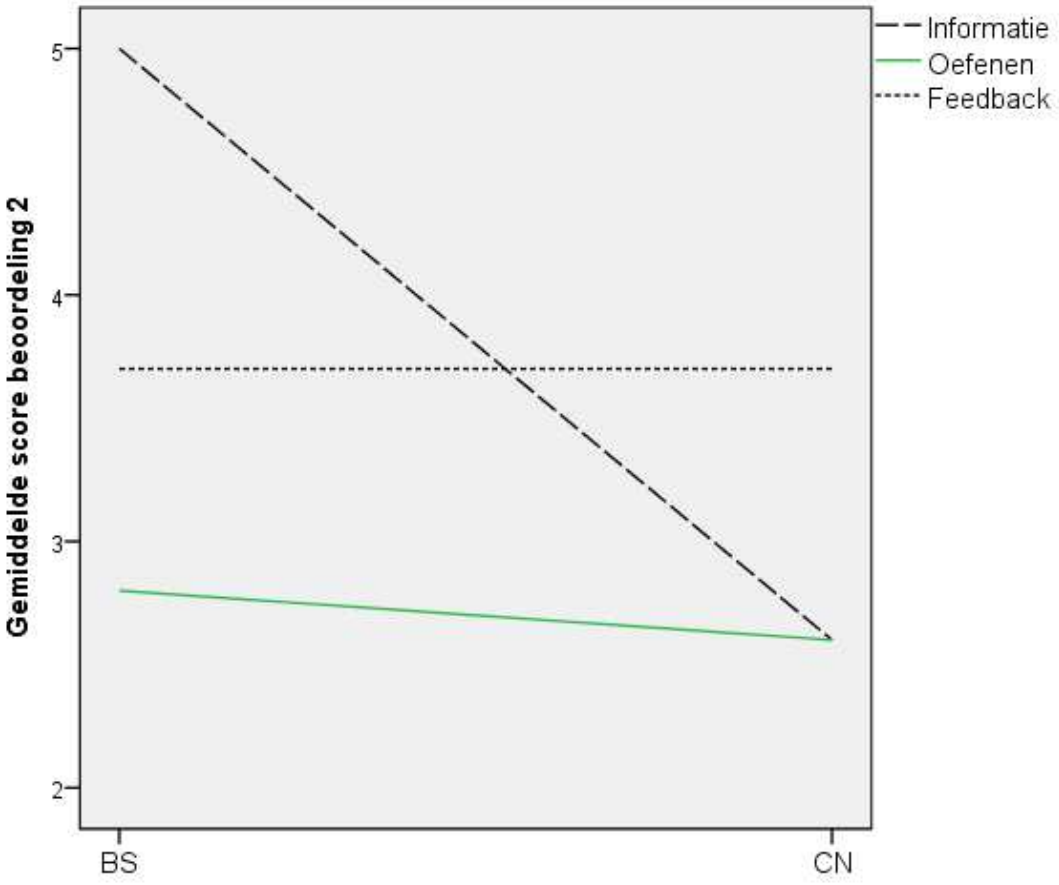
5.4 Betrouwbaarheid van de beoordeling

Om enig idee te krijgen van de betrouwbaarheid van de beoordeling is de mede-auteur gevraagd twee beoordeelde methodes opnieuw te beoordelen. Het verslag van deze beoordeling is opgenomen in bijlage 17. Beoordeeld zijn de papieren methode Basisvaardigheden Spelling en de digitale methode CambiumNed.

De resultaten van de tweede beoordeling zijn weergegeven in Figuur 5.1 en de overeenkomstige resultaten van de eerste beoordeling zijn weergegeven in Figuur 5.2. De figuren laten zien hoe de twee beoordeelde methodes scoren op de hoofdcategorieën van het beoordelingsschema: informatiebasis, oefenboek en feedbackmiddel. Beide methodes werden bij beide beoordelingen als oefenboek en als feedbackmiddel ongeveer gelijk beoordeeld (de desbetreffende lijn loopt horizontaal). Bij beide beoordelingen werd Basisvaardigheden Spelling als informatiebasis beoordeeld als belangrijk beter dan CambiumNed (de desbetreffende lijn loopt schuin naar beneden).

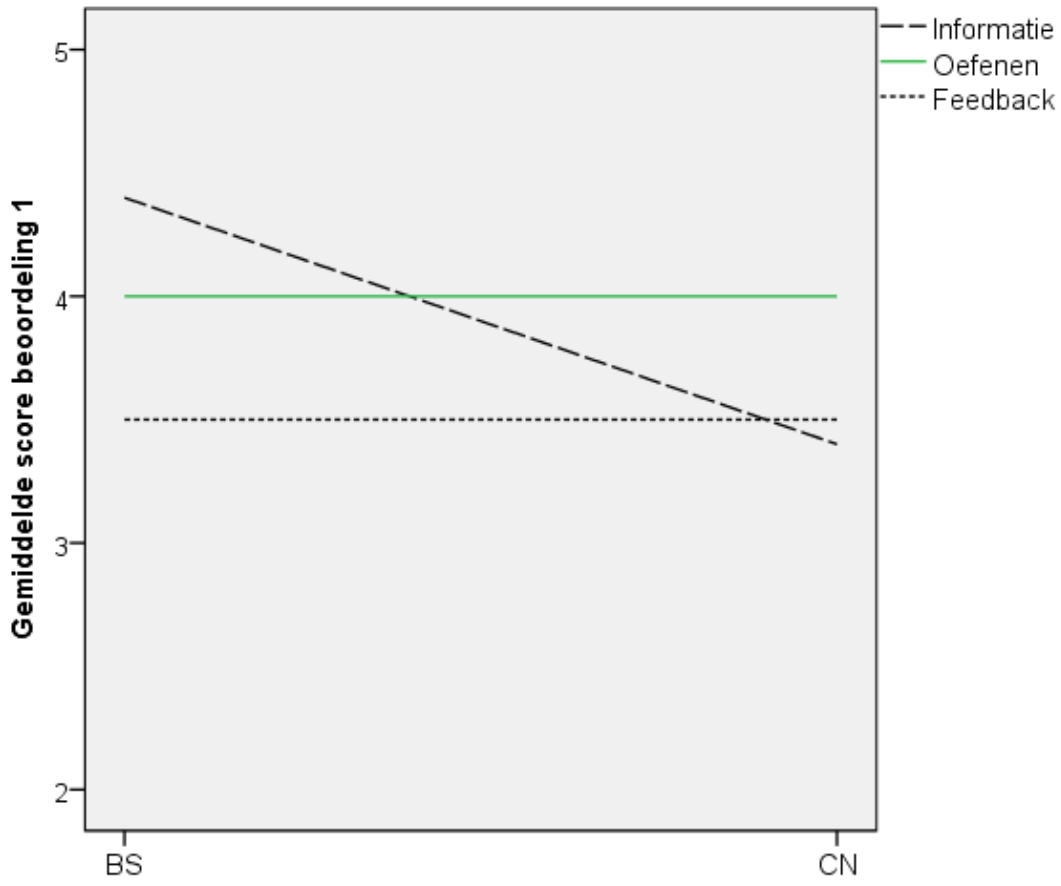
Beide beoordelingen leverden op de hoofdcategorieën daarmee hetzelfde beeld van de twee beoordeelde methodes. Per hoofdcategorie zijn er in beginsel drie mogelijkheden: de eerste methode is beter (A), beide methodes zijn ongeveer gelijk (B), de tweede methode is beter (C). Wanneer Basisvaardigheden Spelling fungeert als eerste methode en CambiumNed als tweede methode, leverde de eerste beoordeling voor de hoofdcategorieën informatiebasis, oefenboek en feedbackmiddel respectievelijk: A, B, B. De tweede beoordeling leverde voor de hoofdcategorieën dezelfde uitkomsten: A, B, B.

Figuur 5.1 De gemiddelde scores van beoordeling 2 voor Basisvaardigheden Spelling (BS) en CambiumNed (CN). Beide methodes scoren vergelijkbaar behalve op Informatie.



Uit de figuren blijken de beoordelingen op één punt wel belangrijk te verschillen. Bij de tweede beoordeling is het gebruik als oefenboek belangrijk lager beoordeeld dan bij de eerste beoordeling. De helling van de lijn is niet veranderd (beide oefenboeklijnen lopen ongeveer horizontaal), beide methodes zijn op dit punt dus ongeveer gelijk beoordeeld. Wat wel veranderd is, is het niveau van de lijn. Bij de tweede beoordeling loopt de oefenboeklijn belangrijk lager dan bij de eerste beoordeling.

Figuur 5.2 De gemiddelde scores van beoordeling 1 voor Basisvaardigheden Spelling (BS) en CambiumNed (CN). Beide methodes scoren vergelijkbaar behalve op Informatie.



Deze discrepantie in absolute waarde tussen beide beoordelingen is ontstaan doordat bij de tweede beoordeling is uitgegaan van de specifieke geschiktheid van de methode om hbo-studenten schrijfvaardiger te maken via het oefenen met het corrigeren van foute zinnen, terwijl bij de eerste beoordeling is uitgegaan van de doelstelling van de methode zelf (zie 5.5). Het beoordelingsschema was op dit punt niet volledig dwingend en liet ruimte aan de beoordelaar om zelf de doelstelling te kiezen waarvan wordt uitgegaan.

De beoordeling van de methodes had twee hoofddoelen: ten eerste de selectie van de beste (beschikbare) methode en ten tweede een inschatting van de optimaliteit van de beste methode. Daarbij ging het om de vraag of de beste methode naar verwachting maximaal effectief is om leerwinst te realiseren bij studenten op schrijfvaardigheidsgebied.

Met betrekking tot de eerste hoofddoelstelling leverden beide beoordelingen hetzelfde eindresultaat. Uitgaande van deze twee methodes is Basisvaardigheden Spelling de beste methode van de twee. Ook wanneer men de beoordelingen zou beperken tot de twee hoofdcategorieën die volgens het ABC-model als het belangrijkst gelden, oefenboek en feedbackmiddel, gaven beide beoordelingen dezelfde uitkomst. In dat geval zouden beiden methodes als (vrijwel) gelijk worden beoordeeld.

Met betrekking tot de tweede hoofddoelstelling lieten beide beoordelingen ook een soortgelijke uitkomst zien, namelijk dat de beoordeelde methodes op het punt van oefeningen en feedback niet optimaal zijn.

De tweede beoordeling leverde daarmee op alle drie onderzochte punten (de vergelijking op de drie hoofdcategorieën van de methodes, de beste methode en de optimaliteit van de beste methode) dezelfde resultaten als de eerste beoordeling. Op basis van dit resultaat is er geen reden om aan te nemen dat de beoordeling volgens het beoordelingsschema onbetrouwbaar was.

Aan dit positieve resultaat dienen echter geen al te sterke conclusies verbonden te worden. Allereerst was de beoordeling niet volledig blind, de tweede beoordelaar was op de hoogte met de hoofdconclusies van de eerste beoordeling en had geruime tijd voor de tweede beoordeling de uitkomsten van de eerste beoordeling gelezen waardoor hij voorkennis kon hebben. Een tweede factor waardoor de beoordelingen meer kunnen overeenstemmen dan anders het geval geweest zou zijn, is dat beide auteurs een bepaalde vertrouwdheid hebben ontwikkeld met het ABC-model. Het lijkt niet bij voorbaat uit te sluiten dat iemand die die vertrouwdheid mist, meer problemen kan hebben met het beoordelen van de diverse punten in het beoordelingsschema.

5.5 Conclusies en discussie

Zeventien papieren en negen digitale programma's over elementaire taalfouten op het gebied van spelling en stijl zijn aan de hand van het beoordelingsschema voor studieteksten beoordeeld op hun verwachte geschiktheid voor eerstejaars hbo-studenten. Dit beoor-

delingsschema veronderstelt dat goede studieteksten behalve informatie ook oefeningen moeten bevatten en een middel waarmee studenten zichzelf feedback kunnen geven. Uiteindelijk zal echter uit empirisch onderzoek moeten blijken of een methode leidt tot bevredigende leerresultaten.

Van de twee criteria die onder feedbackmiddel aan het schema zijn toegevoegd, namelijk uitleg bij feedback na een fout antwoord en voortgangsinformatie, bleek alleen het laatste criterium van belang te zijn. Uitleg bij feedback na een fout antwoord was problematisch en alleen goed te realiseren bij antwoorden die voorgestructureerd waren. Bij de beoordeling van de papieren methodes bleek dat uitleg na een fout antwoord meestal ontbrak of zeer globaal was. Digitale programma's maakten vaak gebruik van voorgestructureerde antwoorden en hadden daardoor wel de mogelijkheid om uitleg te geven na een specifiek fout antwoord. Slechts vier van de negen digitale methodes maakten gebruik van deze mogelijkheid. Bij vier methodes ontbrak uitleg na een fout antwoord helemaal en bij één methode werd alleen globale uitleg gegeven. Maar ook als sprake was van specifieke uitleg, was onduidelijk of de uitleg begrepen werd. Om dat te weten zou de hele cyclus opnieuw moeten starten met oefeningen en feedback. Daarom is dit criterium wel in de beschrijving van elke methode opgenomen, maar weggelaten in de tabel.

Verder konden enkele criteria niet goed beoordeeld worden door ontbrekende of onduidelijke doelstellingen. Alle programma's, zowel papieren als digitale, hadden een vage of ontbrekende doelstelling waardoor de relevantie van de informatie en de relevantie en volledigheid van de oefeningen moeilijk te beoordelen waren.

Door de papieren en de digitale taalmethodes zo veel mogelijk op dezelfde criteria te beoordelen was het in beginsel mogelijk een vergelijking te maken. De verwachtingen vooraf over de papieren en digitale methodes werden slechts gedeeltelijk bevestigd.

De papieren methodes werden als informatiebasis het hoogst beoordeeld (gemiddeld 3.6). De digitale methodes scoorden op dit punt overwegend lager (gemiddeld 2.7). Als informatiebasis werden de papieren methodes daarmee positiever beoordeeld. Dit was volgens de verwachting. De volledigheid en duidelijkheid van de informatie waren bij zowel de papieren als digitale methodes, afgaande op de gemiddelde scores, vaak een probleem. Bij de digitale methodes was ook de toegankelijkheid van de informatie een probleem.

Als oefenboek scoorden de papieren en de digitale methodes gemiddeld vergelijkbaar (respectievelijk 3.2 en 3.0). De verwachting dat digitale methodes op dit punt in het voordeel zouden zijn, werd niet bevestigd. De hoeveelheid en de geordendheid van de oefeningen was bij zowel de papieren als de digitale methodes, afgaande op de gemiddelde beoordelingen, vaak een probleem. De hoeveelheid oefeningen was soms zeer beperkt. Verder viel er vaak weinig ordening in te ontdekken. Het niveau was laag en bleef dat dan ook of het was meteen vrij hoog. Bij de digitale programma's was het niveau vaak voortdurend laag.

Qua geschiktheid als feedbackmiddel werden de papieren methodes gemiddeld beoordeeld met 2.3, terwijl de digitale methodes beoordeeld werden met 3.5. De digitale methodes waren in doorsnee op dit punt dus belangrijk geschikter. Dit was volgens de verwachting. Hoewel de digitale methodes voor feedback als geschikter werden beoordeeld, deden zich hier ook nog problemen voor. Met name de snelheid en de duidelijkheid van de feedback bleek vaak problematisch: respectievelijk bij zeven en zes van de negen digitale methodes. Ook kwam bij vijf methodes de feedback al voordat het antwoord gegeven was. Op het punt van de voortgangsinformatie werden slechts twee van de negen digitale methodes als goed beoordeeld.

Wat was de implicatie van deze beoordeling voor het interventieprogramma? De beste papieren en de beste digitale methode bleken, hoewel beter beoordeeld dan de andere methodes van dezelfde soort, nog steeds duidelijke bezwaren te hebben.

Basisvaardigheden Spelling werd beoordeeld als de beste papieren methode, maar op het gebied van uitleg en hoeveelheid oefenmateriaal scoorde deze methode matig. Verder waren de oefeningen nauwelijks geordend op grond van hun moeilijkheidsgraad. De oefeningen waren in het begin simpel en dat bleven ze ook. Deze methode werd als slecht beoordeeld op het punt van snelheid van de feedback en de afhankelijkheid daarvan. Voortgangsinformatie was slechts gedeeltelijk aanwezig.

Nedercom werd beoordeeld als de beste digitale methode. Op het punt van hoeveelheid oefenmateriaal en de snelheid waarmee de oefeningen gemaakt konden worden, werd het programma echter beoordeeld als matig. De ordening van de oefenstof op grond van de moeilijkheidsgraad werd als slecht beoordeeld. Het niveau van de oefeningen was simpel en van een opklimmende moeilijkheidsgraad was nauwelijks sprake. Wat dat betreft, was er weinig

onderscheid met de meeste andere digitale methodes. Verder was *Nedercom* weliswaar in behoorlijke mate sturend, maar de student werd niet vastgehouden totdat hij de stof meester was.

Bij een keuze tussen deze twee methodes zou die in het voordeel uitvallen van *Nedercom.* Bij *Basisvaardigheden Spelling* waren de oefeningen weliswaar sneller te maken, maar de feedback bij *Nedercom* scoorde op twee punten beduidend hoger, namelijk bij snelheid en afhankelijkheid. Voor het overige waren de verschillen minimaal.

Als de beste papieren en digitale methode slechts een matige hoeveelheid oefenmateriaal hebben, waarbij de oefeningen simpel zijn en amper oplopen in moeilijkheidsgraad, lijkt geen enkel programma te voldoen in het licht van het beoordelingsschema voor studieteksten. Niet een van de onderzochte taalmethodes werd daarom als geschikt beoordeeld voor het effectief wegwerken van taaldeficiënties.

Na de afronding van de beoordeling van de verschillende methodes kwamen eveneens de resultaten van het foutenonderzoek beschikbaar. Hierdoor werd duidelijk dat er bij alle onderzochte taalmethodes ook een probleem was met de inhoud. De verschillende methodes bleken zich vrijwel zonder uitzondering op de verkeerde onderwerpen te richten.

Het foutenonderzoek leverde twintig soorten fouten op die studenten maken als ze zelf schrijven (zie Tabel 4.4). Sommige foutsoorten kwamen heel vaak voor, zoals 'Verkeerd woord', 'Niet-lopende zin', 'Interpunctie', 'Overbodig woord/overbodige zin'. Daarentegen kwam de 'D/t-fout' heel weinig voor. Door deze foutenlijst was het mogelijk om na te gaan of de beoordeelde taalmethodes zich inderdaad op deze fouten richtten.

De spelling van werkwoordsvormen werd vrijwel altijd behandeld, namelijk in alle digitale methodes en in vijftien papieren methodes (van de zeventien). Uit de resultaten van het foutenonderzoek bleek dat de 'D/t-fout' weliswaar een heel zichtbare fout was, in die zin dat deze fouten als enige gesignaleerd werden door steeds alle vier de beoordelaars, maar d/t-fouten kwamen niet vaak voor. De 'D/t-fout' kwam in de dertig teksten 16 keer voor, terwijl 'Verkeerd woord' 156 keer voorkwam.

'Verkeerd woord' was de meest voorkomende fout, maar deze fout werd nauwelijks behandeld. Alleen in twee papieren methodes werd hier iets over opgemerkt. Het is opvallend dat ook met de andere meest voorkomende fouten, zoals 'Niet-lopende zin', 'Overbodig woord/overbodige zin', 'Alinea-indeling', 'Voorzetsel' en 'Ontbrekend woord' niet of nauwelijks geoefend werd in de taalmethodes. Deze fouten horen bij de eerste acht foutcategorieën die samen 75% van alle bevestigde fouten voor hun rekening namen.

Slechts twee fouten die bij de acht meest voorkomende fouten staan, kwamen in redelijk veel taalmethodes voor. 'Interpunctie' werd behandeld in twaalf papieren en vijf digitale methodes en 'Spelfout' (los van werkwoordspelling) in tien papieren en drie digitale methodes.

De uitkomsten samenvattend kan gesteld worden dat de inhoud van de methodes zich niet richtte op de meest voorkomend problemen, terwijl de hoeveelheid oefeningen en de ordening van de oefeningen te wensen overlieten. De digitale methodes voldeden beter op het punt van feedback dan de papieren methodes, maar waren ook op dit punt lang niet optimaal.

6

Deelstudie 3

Constructie en effect TAVAN-programma

6.1 Inleiding

In dit hoofdstuk wordt de vraag beantwoord hoe een nieuw onderwijsprogramma eruit zou moeten zien om basale schrijfvaardigheid bij hbo-studenten te verbeteren (onderzoeksvraag 3). De volgende vraag was hoeveel effect dit nieuwe onderwijsprogramma had op de basale schrijfvaardigheid van eerstejaars hbo-studenten (onderzoeksvraag 4).

6.1.1 Doel en randvoorwaarden

Het doel van het interventieprogramma was: foutloos schrijven. De doelvariable was: het aantal fouten per honderd woorden. Foutloos schrijven, nul fouten per honderd woorden, is een ideaal dat in de praktijk nooit gerealiseerd zal worden. Om richting te geven aan het onderwijs is het echter een duidelijk doel. Bij iedere tekst kan gestreefd worden naar perfectie. Iedere fout wordt gezien als een mogelijke afwijking van het ideaal en is er dan één te veel (zie paragraaf 2.4).

Op grond van de uitgangspunten van goed schrijfonderwijs (zie 2.4) en het ABC-model (zie 3.3) leek het duidelijk dat studenten moesten oefenen met het opsporen en wegwerken van fouten. Hierbij speelden de in het foutenonderzoek gevonden foutcategorieën en hun frequentie een belangrijke rol. Het heeft immers weinig zin te oefenen met fouten die niet gemaakt worden. Verder moest rekening gehouden worden met een aantal randvoorwaarden als docentafhankelijkheid, kosten per student, beschikbare tijd en factoren met betrekking tot de invoering.

Docentonafhankelijk

Docenten in het hbo zijn vermoedelijk geneigd te veronderstellen dat basale schrijfvaardigheid reeds lang verworven is in het voorgaande onderwijs en dat foutloos schrijven niet het doel van het hbo-schrijfonderwijs behoort te zijn. Methodes als *Leren Communiceren* (Steehouder et al., 2006), *Zakelijk Schrijven* (Ter Horst & Molenaar, 2006), *Zakelijke Communicatie deel 1, Zakelijke Communicatie deel 2* (Janssen, Jansen & Kinkhorst, 2007) en *Zakelijke Communicatie - Schriftelijk* (Knispel, 2008) richten zich niet op basale schrijfvaardigheid en veronderstellen impliciet dat studenten over deze vaardigheid beschikken.

In bijvoorbeeld *Leren communiceren* van Steehouder (2006) gaat het over doelgerichte communicatie, het structureren van de tekst met behulp van bouwplannen en over verschillende tekstsoorten. Het nakijken van de eigen tekst door de student op fouten en gebreken wordt gepresenteerd als een vanzelfsprekendheid. Hoe de student dit eventueel zou moeten leren, wordt niet behandeld. In het hoger beroepsonderwijs gaat het in de eerste plaats om schrijven op hogeschoolniveau, met weinig aandacht voor basale taalfouten. Studenten met onvoldoende basale schrijfvaardigheid moeten zichzelf proberen bij te spijkeren op dit gebied.

Verder wordt het werken met schrijfopdrachten vermoedelijk (terecht) gezien als niet erg effectief en arbeidsintensief. In onderzoek naar schrijfonderwijs werd zelf laten schrijven niet als een effectieve optie voorgesteld (zie 2.2). Behalve dat het geven van feedback op schriftelijk werk van studenten problematisch is door het vele nakijkwerk (zie 2.1), komt de feedback te laat om didactisch zinvol te zijn. Op het moment dat de feedback komt, is de student de geformuleerde zinnen al vergeten.

Een volgend probleem is de inhoud van de feedback. Een docent die aangeeft dat het 'goed' was, communiceert dat er geen verdere verbetering nodig is. Als een docent aangeeft dat het 'fout' was, is dat niet bevorderlijk voor het ontwikkelen van zelfvertrouwen bij het schrijven. Volgens Zimmerman en Risemberg (1997) beïnvloedt het vertrouwen in het eigen kunnen de intrinsieke motivatie om te schrijven (geciteerd in Graham, 2006). Onderzoek van Ahmed (2010) liet zien dat emoties van leerlingen bij wiskunde onderwijs van invloed zijn op hun prestaties. Bij nadruk op fouten in de tekst is de student bovendien geneigd die informatie voor kennisgeving aan te nemen, zonder daadwerkelijk de fout weg te werken. Hij leert op die manier alleen dat hij het fout deed. Verder is feedback dat het 'fout' was, vaak onvoldoende informatief om voor de student bruikbaar te zijn. De vraag is immers, hoe het wel moet. Maar volledig herschrijven van de foute passage door de docent is qua beschikbare tijd geen optie, terwijl de kans groot is dat de student een eventueel wel voorgestelde oplossing van de fout amper bekijkt. Zonder feedback kan echter gemakkelijk de indruk worden gewekt dat het niet uitmaakt hoe je formuleert: "Without feedback on minor errors, students may not feel motivated to improve their writing skills" (Bacon & Scott Anderson, 2004, p. 443).

Wanneer binnen het hoger onderwijs wel aandacht was voor de geringe taalvaardigheid van studenten, werd niet gefocust op foutloos schrijven en het trainen van basale schrijfvaardig-

heid. In 2006 is het Nederlands/Vlaams Platform Taalbeleid Hoger Onderwijs opgericht vanwege bezorgdheid over de beheersing van het Nederlands in het hoger onderwijs. Bij een mogelijke aanpak van dit probleem werd op hogeschoolniveau vaak breed en integraal gedacht: taalvaardigheid had op alle aspecten van taal betrekking (luisteren, spreken, lezen, schrijven) en moest bij alle vakken aan de orde komen. Iedere docent was taaldocent (Beijer, Gangaram Panday & Hajer, 2010; Berntsen & Gangaram Panday, 2007; Gangaram Panday, Droop & Rutten, 2008; Van der Westen, 2006). Dit heeft geleid tot verschillende benaderingen, waarbij de uitwerking nog in ontwikkeling is. Het meetbaar maken van het zogenaamde 'taalontwikkelend lesgeven' bleek lastig te zijn (Kuiken, 2010).

Docenten Nederlands die vanuit de brede visie de schrijfvaardigheid van studenten beoordeelden, waren niet specifiek gericht op foutloos schrijven (Beijer & Hajer, 2007; Hajer, 2005). Andere docenten dan neerlandici die de inhoud van de teksten beoordeelden, constateerden wel dat er veel fouten gemaakt werden, maar beschouwden dit meer als een gegeven waaraan niets te doen was (Van den Brink, 2007; Van den Westen, 2003).

Een mogelijk probleem van andere aard is dat docenten bij het implementeren van onderwijsmethodes in de praktijk de methode vaak niet of slechts beperkt volgen. Een docent kan stukken van de stof overslaan, accenten anders leggen, zelf uitleg gaan geven buiten de methode om, oefeningen niet laten uitvoeren, oefeningen wel opgeven, maar nalaten te checken of ze inderdaad worden gemaakt. Door dit soort problemen is het resultaat van een onderwijsmethode vaak sterk afhankelijk van de betrokken docenten.

Wat betekent dit voor een interventieprogramma om basale schrijfvaardigheid bij studenten te verbeteren? Docenten in het hoger onderwijs zullen naar verwachting niet gemotiveerd zijn prioriteit te geven aan vaardigheden die studenten in het voorgaande onderwijs hadden moeten verwerven. Verder zijn de mogelijkheden van docenten om te laten oefenen met schrijven beperkt, doordat hun feedback te traag komt en de benodigde tijd ontbreekt. Docenten zijn verder vermoedelijk meer geneigd om te doceren dan te laten oefenen. Een interventieprogramma dat relatief docentonafhankelijk is en in beginsel geschikt is voor zelfstudie, verdient daarom de voorkeur. Dat wil zeggen dat de methode zo geconstrueerd moet zijn dat deze door de student in beginsel zelfstandig kan worden doorgewerkt. De rol van de docent verandert daardoor belangrijk. In plaats van vol vuur de stof over te dragen, wordt hij iemand die rondloopt om te checken of iedereen wel aan het werk is. Zijn functie wijzigt van bevlogen docent naar manager van een leersysteem.

Dit idee sluit ook aan bij PSI (Personalized System of Instruction) waar zelfstudie de primaire methode van onderwijs vormt en de docent niet langer doceert, maar vooral optreedt als organisator en bewaker van het studiesysteem. Bij PSI levert de organisatie van de deeltoetsen met behulp van tutors echter veel hoofdbrekens op. Om die reden worden de deeltoetsen in het systeem tegenwoordig vaak via de computer afgenomen en laat men het systeem van proctors dan vervallen (Pear & Martin, 2004). Doordat de leerstof via de computer als tekst of als video kan worden aangeboden, ontstaat daarmee de mogelijkheid voor een volledig geautomatiseerd leersysteem, zoals MITx van Massachusetts Institute of Technology (Coughlan, 2012) of Udacity dat online-onderwijs aanbiedt aan studenten van over de hele wereld (DeSantis, 2012). De kosten van het onderwijs dalen door deze automatisering belangrijk, terwijl hetzelfde onderwijs veel meer studenten bereikt.

Overigens zijn zelfstudie-systemen niet echt nieuw en is een geautomatiseerd systeem niet altijd noodzakelijk. Na de uitvinding van de boekdrukkunst werden studieboeken steeds belangrijker. Later verschenen studieboeken in de vorm van geprogrammeerde instructies (GI). Hoewel GI een tijdlang uit de mode is geweest, laat Kuhn (1996) zien dat het mogelijk is een inleidende cursus op universitair niveau aan te bieden als GI en op die manier meer dan 2 miljoen 'studenten' te bereiken. Een nadeel van studieboeken en GI is dat de docent iedere controle over het leerproces verliest, terwijl de situatie voor de student zeer vrijblijvend wordt. Zo is het bij de GI van Kuhn onduidelijk hoeveel van zijn lezers daadwerkelijk de hele GI hebben doorgewerkt en wat ze daarvan precies hebben opgestoken.

Veel geautomatiseerde systemen zijn afgeleid van PSI en hoewel PSI belangrijk effectiever is dan traditioneel onderwijs, is het vermoedelijk nog lang niet optimaal effectief, doordat gewerkt wordt met relatief grote leerstofeenheden. Op grond van het ABC-model is effectief onderwijs, onderwijs dat zeer interactief is. Feedback krijgen na dertig seconden bezig geweest te zijn, werkt sneller en duidelijker dan feedback krijgen na een week bezig geweest te zijn (deeltoetsen in PSI hebben vaak de omvang van 1 week studiestof). Door te werken met veel korte opdrachten kan de effectiviteit ten opzichte van PSI daardoor vermoedelijk nog belangrijk worden vergroot.

Het effect van snelle feedback is vele malen aangetoond en staat bekend staat als 'immediacy of reinforcement' (Cooper, Heron & Heward, 2007, p. 259) en 'delay of reinforcement' (Renner, 1964, p. 341). Hull verklaarde het fenomeen al in 1932 met de 'goal gradient hypothese' (p. 26). Riesen (1940) rapporteerde dat chimpansees door een vertraging van vier

seconden er niet in slaagden een discriminatietaak te leren, ondanks 600 'trials'. Cooper et al. (2007, p. 259) merkten op: "Emphasizing the importance of the immediacy of reinforcement is essential. . . . a response-to-reinforcement delay of 1 second will be less effective than a 0-second delay. This is because behaviors other than the target behavior occur during the delay; the behavior temporally closest to the presentation of the reinforcer will be strengthened by its presentation."

Hoewel docentonafhankelijk onderwijs belangrijke voordelen lijkt te bieden, werkt een te vrijblijvend onderwijsaanbod in de praktijk niet goed. Zo bleek dropout bij PSI een probleem, doordat studenten het halen van een eenheid te lang uitstelden. Fox (2004) merkte op: "Students proceed through the course at their own pace, but strategies to reduce procrastination are recommended" (p. 212). Het inroosteren van vaste en verplichte contacturen kan om die reden van groot belang zijn.

Kosten per student

Een voordeel van een interventieprogramma voor basale schrijfvaardigheid dat overwegend uitgaat van zelfstudie, is dat de kosten wat betreft de begeleidingstijd beperkt kunnen worden. De rol van de docent of begeleider is primair te zorgen dat het programma doorgewerkt wordt. Hierdoor kunnen de kosten voor de instelling beperkt blijven, terwijl de opbrengsten groot kunnen zijn qua verbeterde basale schrijfvaardigheid, lagere dropout, minder studievertraging en een hogere kwaliteit van de afgestudeerden.

Overigens vormen de kosten van onderwijs niet altijd de doorslaggevende overweging bij de invoering van onderwijs. Zo werd in het cursusjaar 2009-2010 op de Hanzehogeschool Groningen de zeer arbeidsintensieve Da Vinci minor aangeboden. In deze minor werd een half jaar lang, vier dagen per week gewerkt met groepen van vijf studenten onder begeleiding van docenten (Dijkma, 2010; Lücker-De Boer, 2010).

Beschikbare tijd

Ook het idee dat er slechts een beperkte tijd beschikbaar zou zijn voor het onderwijs op het gebied van basale schrijfvaardigheid lijkt aanvechtbaar. Zo werd er in het foutenonderzoek

(zie 4.4) op gewezen dat de studenten op dat moment al 13.300 uur onderwijs hadden gekregen. Kennelijk heeft in de onderwijspraktijk het maximaliseren van het aantal onderwijsuren soms prioriteit boven het maximaliseren van de leerwinst.

Uit onderzoek blijkt een positief effect op leerprestaties van het 'spacing effect', het verdelen van de stof over een langere periode (Rohrer & Pashler, 2010). Dit betekent niet dat er meer uren onderwijs moeten worden gegeven, maar het betekent vooral dat de tijd die de student besteedt om de vaardigheid in te oefenen, beter gespreid kan worden over een langere periode dan over een korte periode. Vijf weken lang iedere werkdag een uur oefenen is effectiever dan een week 5 uur per dag oefenen. Dit principe lijkt te pleiten voor een relatief langdurige opzet met lessen van een beperkte omvang.

Factoren met betrekking tot de invoering

Er zijn bekende voorbeelden van methodes die hun populariteit verloren of nooit werden ingevoerd, doordat ze een extra belasting voor de docent betekenden of voor de docent geen toegevoegde waarde hadden. Zo wordt voor PSI aangegeven dat de complexe organisatie, administratie en het extra werk dat deze onderwijsmethode met zich meebrengt, vermoedelijk de belangrijkste reden was waarom deze effectieve methode tegenwoordig minder populair is (Pear & Martin, 2004, p. 223).

Ook Direct Instruction heeft vermoedelijk om een soortgelijke reden nooit op grote schaal ingang gevonden. In deze methode wordt namelijk een groot beroep gedaan op de inzet van de docent (Coombs, 1998). Docenten, onderzoekers en beleidsmakers bleken voortdurend de voorkeur te geven aan andere methodes waarvan niet alleen bekend was dat ze belangrijk slechter presteerden, maar waarvan zelfs bekend was dat ze negatief presteerden (slechter dan het traditionele onderwijs van de controlegroep). Nadler (1998) merkt in dit verband op: "In public education, nothing succeeds like failure" (p. 39).

Aannemelijker is echter te veronderstellen dat docenten geneigd zijn het aantal uren te maximaliseren en de taakbelasting te minimaliseren. Een onderwijsmethode die enkele malen zo effectief is als een bestaande, kan op termijn gemakkelijk tot minder inzet van docenten leiden. Wanneer diezelfde methode ook nog eens belangrijk belastender is voor docenten, is de animo bij docenten voor de nieuwe methode vermoedelijk niet bijster groot.

Om ingevoerd te worden, is het daarom wenselijk dat het interventieprogramma de docent routinewerk uit handen neemt, terwijl hij op een hoger niveau wel nodig blijft als organisator en leider van het leersysteem.

Mogelijk is dit ook de reden waarom GI weinig populair was. De docent die een cursus in GI-vorm probeert te schrijven, bezorgt zichzelf veel meer werk en stelt veel hogere eisen aan het eigen didactisch inzicht dan wanneer hij een boek of syllabus schrijft. Vervolgens blijkt hij zichzelf echter grotendeels overbodig gemaakt te hebben. De studenten kunnen immers zelf de stof doorwerken, zonder dat hij nodig is. Om dezelfde reden zal ook een effectieve, verkrijgbare methode in GI-vorm bij docenten vermoedelijk niet enthousiast ontvangen worden. De GI van Kuhn (1996) werd door de uitgever aangeprezen als 'self-teaching guide'. Kuhn zelf formuleerde voor zijn GI vier doelgroepen (p. xi). Slechts één van die vier doelgroepen ging uit van gebruik in een reguliere cursus.

De opmerking van Nadler (1998) die hiervoor werd aangehaald, kan dus beter anders geformuleerd worden. In het onderwijs kiezen docenten bij voorkeur die methode die hun taak als docent optimaal belonend en minimaal belastend maakt in plaats van de methode die de grootste leerwinst levert. De impliciete verontwaardiging van Nadler hierover is wel begrijpelijk, maar mogelijk niet zinvol. Docenten gedragen zich volgens deze regel immers precies als op basis van het ABC-model verwacht zou moeten worden.

Deze regel zou ook de ineffectiviteit van het bestaande schrijfonderwijs met betrekking tot basale schrijfvaardigheid eenvoudig verklaren. Uitgaande van het ABC-model is er om goed te leren schrijven, veel oefening nodig. Voor de docent betekent dat echter dat hij daardoor tijdens de contacturen grotendeels overbodig lijkt, terwijl hij buiten de vaste contacturen veel saai en moeizaam nakijkwerk te doen heeft, zonder duidelijke positieve consequenties. Iedere vorm van onderwijs waarbij niet geschreven wordt en geen nakijkwerk valt te verrichten, vormt dan een betere keuze. Bij de invoering van een nieuwe methode, is dus vermoedelijk niet zozeer de effectiviteit bepalend, maar de opbrengst van de methode voor de docent.

6.1.2 Herschrijfopdrachten

Hoe kan de schrijftaak voorgestructureerd worden bij een interventieprogramma dat zich richt op leren schrijven zonder zichtbare fouten? In het schrijfonderwijs wordt vaak gewerkt met grote en vage schrijfopdrachten. Werkstukken van vele bladzijden moeten worden geschreven over een onderwerp. Verder wordt schrijven gezien als het verzamelen van informatie, het evalueren en het ordenen daarvan. Het resultaat is dat schrijven gaat samenvallen met analyserend lezen en onderzoek doen en dat de schrijftaak groot, complex en oncontroleerbaar wordt. Het aspect van informatieverzameling moet daarom uit de schrijfopdracht, wat betekent dat de student moet schrijven over wat hij al weet of over informatie die hij bij de schrijfopdracht krijgt meegeleverd. Studenten laten schrijven over wat ze geacht worden te weten, leidt ook weer tot vage schrijfopdrachten. De ene student blijkt namelijk altijd net iets meer te weten dan de andere en die verschillen zullen doorwerken in de geproduceerde tekst.

De enige overblijvende optie is daarom de benodigde informatie in de schrijfopdracht mee te leveren. Een probleem daarbij is dat die informatie door de auteur gewoonlijk zo goed mogelijk wordt geformuleerd. Aan de ene kant wil de student de informatie uit de schrijfopdracht letterlijk overnemen, omdat die goed geformuleerd is en aan de andere kant moet hij het zelf verwoorden, omdat de docent geen genoegen neemt met letterlijk overschrijven. Het resultaat is dat de opdracht wordt een goede tekst te herschrijven in een andere goede tekst: een moeilijke en weinig praktijkgerichte opdracht. De meegeleverde informatie moet daarom niet correct geformuleerd aangeleverd worden, maar juist niet goed geformuleerd. De schrijfopdracht houdt vervolgens in dat de slecht geformuleerde informatie bewerkt wordt tot goed geformuleerde. Het accent in het schrijfonderwijs verschuift dus van de planfase naar de revisiefase. Ook nu blijft scherp lezen noodzakelijk, maar in dit geval is duidelijk wat gelezen wordt en wat daaruit moet worden afgeleid. Om snel en gericht feedback te kunnen geven, werken kleine, korte opdrachten het meest optimaal. Dat betekent voor de wat langere opdrachten hoogstens een A4 tekst en voor korte opdrachten een enkele zin met één of meer fouten die herschreven moet worden.

Op deze wijze was het mogelijk met alle acht uitgangspunten voor beter schrijfonderwijs rekening te houden (zie 2.4). De beoordeling kon geobjectiveerd worden (1), het product (de herschreven zin of zinnen) stond centraal (2), de nadruk lag op lezen (3) en op reviseren (4) in plaats van op plannen, het accent lag op fouten in de tekst (5) en er kon veel geoefend worden (6) met kleine (7) en duidelijke (8) opdrachten.

Eigen tempo

Op de Hanzehogeschool Groningen wordt getracht rekening te houden met verschillen tussen studenten (*Oog voor etnische en culturele diversiteit*, 2009). Studenten hebben niet altijd dezelfde etnische achtergrond en vormen geen homogene groep. De vooropleiding kan uiteenlopen. Nederlands is niet altijd de moedertaal. Ook wanneer Nederlands wel de moedertaal is, komen de studenten uit milieus waarin verschillend met het Nederlands is omgegaan. Soms zijn studenten dyslectisch of hebben ze een functiebeperking. Ondanks deze verschillen was het interventieprogramma voor alle studenten hetzelfde.

De belangrijkste reden hiervoor was praktisch. Het construeren en onderzoeken op effectiviteit van een enkel interventieprogramma is al veel meer dan wat normaal plaatsvindt. Dit is te vergelijken met het schrijven van een studieboek. Een normale auteur is blij wanneer het hem gelukt is een studieboek te schrijven. Testen van het boek op effectiviteit vindt niet plaats. Wanneer hij echter rekening zou moeten houden met de verschillen in zijn lezerspubliek, zou hij minstens twee boeken moeten schrijven en liefst meer. Vervolgens zou ieder boek idealiter onderzocht moeten worden op effectiviteit. In de praktijk is deze benadering niet uitvoerbaar en is het beter te streven naar één goed interventieprogramma en dat te testen op effectiviteit. Binnen zekere grenzen was het echter wel mogelijk om de opgaven in eigen tempo door te werken. Dit sluit aan bij de 'self pacing' van PSI, waardoor juist langzame studenten evenzeer kunnen profiteren van PSI als snelle studenten (Tyree, 1997).

Mastery learning

Van der Werf en Weide (1991) vonden dat effectief onderwijs aan leerlingen die een andere moedertaal hebben dan het Nederlands zich kenmerkte door veel tijd te besteden aan basisvaardigheden en hoge eisen te stellen aan de doelen die bereikt moesten worden. In PSI is 'mastery learning' een belangrijke component. Voor de deeltoetsen wordt een hoge norm gesteld. Een student die de norm niet haalt, moet dezelfde eenheid opnieuw bestuderen en daarna weer de toets proberen. Men zou dit kunnen omschrijven als een blokkadesysteem. Bij een doorloopsysteem mag de student ook na een onvoldoende resultaat doorgaan met het volgende onderdeel. Het voordeel van een blokkadesysteem is dat het leidt tot hogere beheersing, het nadeel is dat het leidt tot extra dropout (Fox, 2004). In het nieuwe interventieprogramma is niet gewerkt met een blokkadesysteem om te voorkomen dat studenten zouden vastlopen in het programma.

6.1.3 TAVAN als studietekst

Uitgangspunt voor het nieuw te construeren programma TAVAN (TAalVAardigheid Nieuw) was het beoordelingsschema voor studieteksten (Tabel 5.2). Volgens dit schema zijn de functies van een studietekst opgesplitst in: informatiebasis, oefenboek en feedbackmiddel.

Voor het TAVAN-programma is besloten om geen theorie over basale taalregels op te nemen, zodat de informatiebasis ontbreekt. Onderzoek naar het effect van uitleg bij het goede antwoord liet volgens Beetsma (2010) zien dat het geven van uitleg soms net zo effectief was als het geven van alleen het goede antwoord en soms minder effectief. Ook bij de zogenaamde Delftse methode voor Nederlands aan buitenlanders werd geen aandacht besteed aan uitleg. Het accent lag op teksten als leermiddel en op het geven van onmiddellijke feedback (Blom 2006; Montens & Sciarone, 1992). Hieraan kan worden toegevoegd dat het effect van het geven van informatie in het algemeen onduidelijk is. Het is mogelijk om uitleg te geven, maar vervolgens is er geen controle op wat met die uitleg gedaan wordt. In sturende leermodellen ligt het accent daarom op wat iemand doet en niet op uitleg van de docent (Parsons & Polson, 2000). Bij het eigen onderwijs aan hbo-studenten was de ervaring dat uitleg bij bestaande digitale Nederlandse taalvaardigheidsprogramma's geen meerwaarde had. Studenten raadpleegden de uitleg niet of als ze dat wel deden, begrepen ze die niet.

Goed schrijven lijkt vooral een vaardigheid te zijn, die geleerd moet worden. Bij het leren van vaardigheden ligt het accent op oefenen in plaats van dat de theorie centraal gesteld wordt. Met het ABC-model als uitgangspunt moet een onderwijsprogramma opgebouwd zijn uit een reeks vragen of opdrachten. Een afzonderlijke informatiebasis was dus niet automatisch nodig. In het interventieprogramma werd wel informatie gegeven in de vorm van feedback. In behavioristische opvattingen zou feedback vooral reïnforcement zijn: de student krijgt te horen dat het antwoord goed was. Na een eventueel fout antwoord zou een student dan nog steeds het goede antwoord niet weten. Aangezien bij papieren methodes de student het eigen antwoord moet checken, was het ook bij behavioristische methodes noodzakelijk het docentantwoord te geven bij de feedback (Holland & Skinner, 1961; Holland, Solomon, Doran & Frezza, 1976; Kuhn, 1996; Mechner, 1965). Na een fout antwoord levert dit docentantwoord de student echter een stukje extra informatie op. De student kan daarna niet alleen weten dat zijn antwoord fout was, maar ook wat het antwoord wel moet zijn.

In een goede geprogrammeerde instructie is de norm dat de student weinig foute antwoorden maakt, waardoor dit aspect in de praktijk vermoedelijk weinig gevolgen heeft. Op het moment dat de vragen vaak te moeilijk uitvallen, gaat deze extra informatie vermoedelijk wel een belangrijke rol spelen. Wanneer informatie in de vorm van een studieboek gegeven wordt, zit de student eerst met het selectieprobleem. Het boek is omvangrijk, om welk stuk informatie gaat het precies? Vervolgens is die informatie niet gebruiksklaar geformuleerd. Uitleg in studieboeken is in algemene bewoordingen gesteld en niet toegespitst op een specifiek geval. In het geval van een testsituatie is de toestand precies omgekeerd. De student denkt mogelijk als volgt. 'Deze vraag kan ook in de toets voorkomen. Dit is het juiste antwoord. Waarom is dit het juiste antwoord? Hoe kan ik dat zo simpel mogelijk onthouden?' Hij is gemotiveerd om een klein stukje informatie optimaal te benutten.

Informatie die in de feedbackfase na een fout antwoord gegeven wordt, blijkt inderdaad een groot leereffect te kunnen genereren. Bij een multiple-choice trivia-test bleken studenten na een fout antwoord dat gevolgd werd door feedback en het goede antwoord vervolgens 78% van deze items goed te beantwoorden in een tweede test met open antwoorden. Zonder die informatie beantwoordden ze 6% goed (Butler et al., 2008). Dergelijke informatie had in deze specifieke situatie dus veel effect. De uitleg werd in TAVAN daarom beperkt tot het presenteren van het docentantwoord in de feedbackfase.

Als oefenboek moet een effectief programma aan een aantal criteria voldoen. Er moet sprake zijn van veel en relevante oefeningen met het oog op de doelstelling en die oefeningen moeten snel te maken zijn. De oefeningen moeten ook geordend zijn op grond van hun moeilijkheidsgraad. Voor TAVAN betekende dit veel oefeningen die binnen een lesuur af te ronden waren. Verder moesten alle fouten van de twintig foutcategorieën voorkomen. Oefenmateriaal uit bestaande methodes viel daarom af. Ongemerkt worden dan bestaande foutindelingen gevolgd die volgens het foutenonderzoek slecht overeenkwamen met fouten die studenten maken als ze zelf schrijven. Verder moesten de oefeningen gemakkelijk beginnen met 1 fout per zin in relatief korte zinnen en geleidelijk oplopen naar meer fouten in complexere zinnen. Bij het herschrijven van teksten waren de tekstgedeeltes eerst kort met weinig fouten per zin en gaandeweg lastiger door de grotere omvang en de hoeveelheid fouten.

De criteria onder feedbackmiddel maken duidelijk wat effectieve feedback inhoudt volgens het beoordelingsschema. Op ieder gemaakt item moet feedback komen die klopt en die niet

alleen duidelijk is, maar ook snel. De feedback moet op het goede moment gegeven worden, namelijk pas nadat het antwoord gegeven is. Verder moet het docentantwoord gepresenteerd worden na een fout antwoord. Ook feedback die aangeeft wat het behaalde resultaat is, kan motiverend werken.

Om als feedbackmiddel effectief te zijn, moet de feedback dus snel komen, meteen na ieder gemaakt item. Het geven van snelle feedback kan alleen met een digitaal programma. Dat bleek ook uit de evaluatie van bestaande methodes, want op dat punt waren digitale programma's duidelijk in het voordeel. De keuze van een digitaal programma voor een deel van TAVAN leek daardoor vanzelfsprekend. Een ander voordeel van een digitaal programma was dat het structurerend werkt, doordat het programma automatisch de opdrachten presenteert. Het voordeel van een online-programma boven een lokaal programma was dat er geen software geïnstalleerd behoeft te worden en dat de de computer de resultaten bijhoudt en administreert.

Verwachtingen

De verwachting was dat de TAVAN-groep een grotere leerwinst (verschil tussen aantal fouten per honderd woorden in begin- en eindtekst) zou vertonen dan de controlegroep. Doordat de TAVAN-training relatief kort was en schrijven gedurende een lange reeks jaren geoefend en gevormd wordt, leek het te verwachten effect van TAVAN niet al te groot, zodat gestreefd werd naar een groepsgrootte voor TAVAN van 30. (De uiteindelijk gerealiseerde groepsgrootte voor TAVAN werd door uitval en dropout ten slotte 18.)

Uit het foutenonderzoek kwam naar voren dat studenten veel slordigheidsfouten leken te maken. Goed schrijven vereist ook een bepaalde houding. De verwachting was dat de houding ten opzichte van schrijven door het volgen van het TAVAN-programma zou verbeteren. Studenten zouden zich er meer bewust van worden dat schrijven tijd en inzet eist en hierdoor een betere schrijfattitude ontwikkelen dan de controlegroep.

Een derde verwachting was dat door het oefenen de TAVAN-studenten zichzelf mogelijk negatiever zouden gaan beoordelen op het gebied van schrijfvaardigheid, doordat ze vaker feedback zouden krijgen dat hun schrijven niet goed was.

6.2 Methode

Onderzoeksopzet

De taaltraining TAVAN werd gegeven in de periode september tot en met december 2010 op de Hanzehogeschool Groningen aan een deel van de eerstejaarsstudenten International Business and Languages (IBL) van het Instituut voor Marketing Management. De trainingsperiode besloeg het eerste onderwijsblok en de helft van het tweede blok.

De onderzoeksopzet bestond uit twee groepen studenten: de experimentele groep (TAVAN-groep) volgde het nieuw geconstrueerde TAVAN-programma, de controlegroep volgde het reguliere taalprogramma TAVA. De eerste onafhankelijke variabele was daarmee het gevolgde programma.

Beide groepen kregen vooraf een tekst te verbeteren, de begintekst, en na afloop een andere tekst als eindtekst. De belangrijkste afhankelijke variabele was het verschil tussen de begin- en eindtekst: de leerwinst. Leerwinst werd overeenkomstig het doel van TAVAN geoperationaliseerd als het aantal fouten PHW (per honderd woorden) dat in de eindtekst minder werd gemaakt dan in de begintekst. De reden om studenten een bestaande tekst met fouten te geven om te herschrijven, was dat TAVAN qua doelstelling zich richt op correct taalgebruik en niet op het bedenken van een inhoud. Verder zouden bij een volledige open opdracht ook onnodig inhoudelijke verschillen ontstaan tussen de teksten die zouden kunnen doorwerken in de beoordeling op aantal fouten PHW.

Voor de begin- en eindtekst waren twee teksten beschikbaar: de Groningen-tekst (G) en de Light-tekst (L). Wanneer de ene tekst steeds als begintekst gebruikt werd en de andere steeds als eindtekst, kon een eventuele vermindering in het aantal fouten PHW het gevolg zijn van een verschil in moeilijkheid tussen de begin- en eindtekst. Het meetinstrument bij de beginmeting was in dit geval immers niet precies gelijk aan het meetinstrument bij de eindmeting. Het zou dan nog wel mogelijk zijn het ene programma te vergelijken met het andere programma op effectiviteit, maar het zou niet langer mogelijk zijn de leerwinst van een programma (de afname van het aantal fouten PHW tussen begin- en eindtekst) in absolute zin vast te stellen. Wanneer bijvoorbeeld gevonden zou worden dat een programma

leidde tot een vermindering van 5 fouten PHW zou deze vermindering het gevolg kunnen zijn van het programma, maar het zou ook kunnen zijn dat dit verschil veroorzaakt was door het verschil tussen begin- en eindtekst. Het zou zelfs kunnen zijn dat een programma een negatief effect zou hebben (bijvoorbeeld -5), dat meer dan gecompenseerd werd door een positief effect van de tekstvolgorde (bijvoorbeeld +8). Welke tekst als begin- of als eindtekst gebruikt werd, kon dus grote invloed hebben op de afname van het aantal fouten PHW.

Om dit volgorde-effect vast te stellen, was het nodig de beide tekstvolgordes (L-G en G-L) als tweede onafhankelijke variabele in het design op te nemen. Een aselecte helft van de studenten in iedere groep kreeg als begintekst G en als eindtekst L (de volgorde G-L); de andere helft kreeg de begintekst L en als eindtekst G (de volgorde L-G). Op deze wijze ontstond een 2x2 design met 4 subgroepen: controlegroep G-L, controlegroep L-G, TAVAN-groep G-L en TAVAN-groep L-G.

Getracht werd in de experimentele en controle condities de beide teksten (L en G) even vaak als begintekst te gebruiken, zodat de volgordes L-G en G-L in de TAVAN-groep en controlegroep even vaak zouden voorkomen. Doordat sommige studenten de opleiding staakten, konden uiteindelijk toch verschillen ontstaan tussen de condities in dit opzicht. Deze verschillen in aantal per subgroep hadden echter geen invloed op de schatting van de effecten tussen de groepen, doordat de variantie-analyse werkte met het gemiddelde per subgroep.

Voor de correlatieberekeningen was het belangrijk het teksteffect (het verschil in moeilijkheid tussen de Groningen-tekst en de Light-tekst) te verwijderen uit de leerwinst en te werken met de 'gestandaardiseerde' leerwinst. Voor de standaardisering werden de aantallen fouten PHW van iedere beoordelaar-tekst combinatie zo gewogen dat iedere beoordelaar-tekst combinatie uitkwam op hetzelfde gemiddelde aantal fouten PHW. Iedere tekst en iedere beoordelaar leverde daardoor gemiddeld evenveel gestandaardiseerde fouten. Het gemiddelde aantal fouten PHW dat in de teksten gevonden werd, veranderde hierdoor niet. Daarna zijn de gewogen aantallen fouten PHW van beide beoordelaars gemiddeld per tekst. De gestandaardiseerde leerwinst werd vervolgens berekend als het verschil tussen begin- en eindtekst (gestandaardiseerde fouten begintekst PHW - gestandaardiseerde fouten eindtekst PHW).

Zowel na de begintekst als na de eindtekst werd dezelfde vragenlijst voorgelegd met vragen over onder meer de schrijfhouding en de eigen schrijfvaardigheid. Aangezien deze vragenlijst ingevuld was door in beginsel alle studenten, werd voor de vragenlijstscores niet de relatief kleine controlegroep gebruikt, maar de grotere groep 'overige' studenten (zie 6.3.3).

De studenten uit de experimentele groep en de controlegroep moesten op dezelfde manier gemotiveerd worden om aan de begin- en eindtekst mee te doen. Het was niet mogelijk om de ene groep een cijfer te geven en de andere groep niet. Daarom is voor de volgende oplossing gekozen. De studenten uit beide groepen moesten verplicht twee keer een schrijfopdracht doen (de begintekst en de eindtekst) als voorwaarde voor het behalen van het cijfer van taalvaardigheid. De condities voor alle studenten waren daarmee hetzelfde: ze moesten allemaal de verplichte schrijfopdracht twee keer doen.

De groepen zijn als volgt samengesteld. De experimentele groep bestond uit één volledige IBL-klas. De controlegroep werd getrokken uit de twee resterende IBL-klassen. Deze klassen volgden het reguliere taalvaardigheidsprogramma (TAVA) en deden mee aan de begin- en eindtekst. Achteraf werd uit deze klassen een aantal studenten aselect getrokken. Alleen van deze studenten werd de begin- en eindtekst beoordeeld. De reden om de controlegroep tot een steekproef te beperken was het tijdrovende beoordelingswerk. Voor de variabelen die gemeten werden via de vragenlijst gold dit argument niet, doordat de antwoorden zonder verdere beoordeling in te voeren waren. Voor deze variabelen is waar mogelijk gebruik gemaakt van alle overige studenten als controlegroep.

De bedoeling was om zowel de experimentele groep als de controlegroep uit circa dertig studenten te laten bestaan. Voor het aantonen van het effect van het interventieprogramma was de grootte van de groepen belangrijk. Bij te kleine groepen was het moeilijk om een verschil te vinden. Aan de andere kant was een belangrijk punt dat het geheel beheersbaar moest blijven. Bij dit aantal studenten in de experimentele en de controlegroep moesten in totaal al 120 teksten beoordeeld worden op fouten. Veel meer leek niet mogelijk (door uitval en dropout bevatte de TAVAN-groep uiteindelijk achttien studenten, voor de controlegroep werden de teksten van twintig studenten gebruikt).

Het doel was de IBL-klas die het TAVAN-programma zou volgen, aselect te vormen. De toewijzing van de studenten aan deze klas werd echter niet door de onderzoeker verricht, maar door de administratie, zodat een strikt willekeurige toewijzing niet kon worden gegarandeerd.

TAVAN

Het doel van het TAVAN-programma was foutloos schrijven. De doelvariabele was het aantal fouten per honderd woorden. De soorten fouten waarmee iedere les geoefend werd, waren afkomstig van de foutcategorieën uit het foutenonderzoek (zie Tabel 4.4).

In totaal besloeg het TAVAN-programma tien weken met twee contacturen per week in een computerlokaal. Alle lessen hadden dezelfde structuur: eerst een uur 100 zinnen verbeteren via het online-programma en daarna een uur een tekst herschrijven met Word. De opbouw van de tien lessen is te vinden in bijlage 7. De lessen werden gegeven door de eerste auteur in het kader van haar promotie-onderzoek.

Voor het oefenen met het verbeteren van zinnen zijn tien lessen gemaakt die via een online-programma aangeboden werden.[*] De zinnen bevatten de eerste acht lessen één fout per zin en de laatste twee lessen twee fouten per zin. Alleen in les 1 waren de fouten onderstreept om de eerste les minder moeilijk te maken. Van deze les is een papieren versie gemaakt, omdat eerstejaarsstudenten zich soms laat inschrijven en dan mogelijk niet meteen bij de eerste les van TAVAN op de computer konden inloggen. Een voorbeeld van de eerste tien zinnen uit deze papieren versie staat in bijlage 8. De zinnen die verbeterd moesten worden, zijn gemaakt op basis van kranten- en tijdschriftartikelen waarin onderwerpen behandeld werden over economie, marketing, politiek, muziek en sport. Ook werden zinnen ontleend aan het foutenonderzoek.

Het nakijken van de verbeterde zinnen door het online-programma was niet volledig feilloos en daarom werd door het programma na een 'goed' antwoord alleen aangegeven dat de door de student verbeterde zin 'matchte' met het docentantwoord. Na een 'fout' antwoord werd deze melding achterwege gelaten en werd het antwoord van de docent gegeven. Het programma gaf verder aan hoeveel procent van de reeds gegeven antwoorden matchte (de TAVAN-score of het matchpercentage). Het behaalde matchpercentage van de les werd per student geregistreerd en telde mee voor het uiteindelijke cijfer dat de student kreeg. Verder registreerde het programma de tijd die de student gebruikte om de les door te werken.

[*]Het online-computerprogramma is ontwikkeld door M. van Es. De lesstof van TAVAN is ontwikkeld door A. van Eerden. Alle rechten voorbehouden.

De teksten werden herschreven op de computer in Word. Een voorbeeld van een tekst die herschreven moest worden, staat in bijlage 9. De teksten die herschreven moesten worden, waren afkomstig uit kranten met onderwerpen die studenten zoveel mogelijk aanspraken. De docent gaf commentaar op de verbeterde versie (zie 6.3.2).

Het cijfer dat de TAVAN-studenten voor het studie-onderdeel kregen, kwam als volgt tot stand. Het gemiddelde werd genomen van de matchpercentages op de gemaakte lessen en het gemiddelde cijfer op het schrijfdossier dat de studenten aangelegd hadden. In dit dossier zaten alle uitwerkingen van de herschrijfopdrachten die ze in het tweede uur kregen. Dit cijfer is verder niet gebruikt bij de data-analyse. De reden om het schrijfdossier niet te gebruiken in het onderzoek was praktisch: de omvang van het onderzoek werd te groot. Verder werd niet verwacht dat deze informatie veel zou toevoegen.

TAVA

De controlegroep kreeg gedurende drie onderwijsperiodes, drie keer per periode één uur het reguliere taalvaardigheidsprogramma TAVA aangeboden, net als voorheen**.** Dit kwam neer op in totaal negen uur TAVA. Dit onderwijs werd verzorgd door collega's van de TAVAN-docent**.**

TAVA bestond uit de onderdelen werkwoordspelling, interpunctie en stijl. In deze volgorde kwam één onderdeel per onderwijsperiode aan bod. Bij werkwoordspelling ging het om de juiste schrijfwijze van de persoonsvorm in de tegenwoordige en verleden tijd, de infinitief, het voltooid deelwoord en het bijvoeglijk gebruikte voltooid deelwoord. Bij interpunctie werden lees- en woordtekens behandeld, het gebruik van hoofdletters en de schrijfwijze van samengestelde woorden. Stijl handelde over grammaticale begrippen, foutieve samentrekking en inversie, foutieve verwijzingen en verbindingswoorden.

TAVA moest de studenten aan de hand van de syllabus *Commercieel correct schrijven* helpen om de kennis over schriftelijk taalgebruik consequent toe te passen in alle zakelijke teksten die ze schreven, met de bedoeling om in het propedeusejaar te laten zien dat ze foutloos konden schrijven (Wubs & Nauta, 2010).

In het eerste uur TAVA van iedere periode werd een diagnostische toets gedaan over een bepaald onderdeel. Daarna moesten de studenten zelfstandig oefeningen maken uit de aange-

boden syllabus die vervolgens besproken werden in de twee resterende lessen van een uur. Het cijfer voor TAVA is vastgesteld op basis van een tentamen per onderdeel en is verder niet in dit onderzoek gebruikt.

Beoordeling begin- en eindtekst

Aan het begin en aan het einde van de trainingsperiode van TAVAN moesten de studenten uit zowel de experimentele groep als de controlegroep in twee uur tijd een korte tekst (300 woorden) gesteld in slecht Nederlands herschrijven. De tekst moest op de computer bewerkt worden (zie bijlage 10). Deze schrijfopdrachten voor de begin- en eindtekst pasten bij wat bereikt moest worden met het interventieprogramma en bij het reguliere taalvaardigheidsprogramma. In beide gevallen was het doel studenten correcte zakelijke teksten te leren schrijven.

Alle teksten zijn door beide onderzoekers beoordeeld. De teksten zijn pas na de eindtekst nagekeken, zodat ze in één keer beoordeeld konden worden om verschil in beoordeling te voorkomen. De teksten werden door de studenten via Word aangemaakt en vervolgens als bijlage via de mail toegestuurd. Deze bijlagen werden gedownload en in een map op de computer opgeslagen. Hierna werden alle te gebruiken teksten in een Word-document geplakt waarbij iedere tekst één pagina besloeg. De namen van de studenten werden vervangen door willekeurige codenummers en genoteerd op een lijst samen met het codenummer. Bij de afname van de begintekst en eindtekst was voor iedere student al genoteerd welke van de twee teksten (Groningen of Light) gemaakt werd bij die gelegenheid. Hierna werden alle teksten tweemaal geprint en werd de volgorde van de teksten gerandomiseerd.

De teksten werden blind beoordeeld: op de teksten stond geen naam, alleen een (gecodeerd) studentnummer en tekstnummer. De gerandomiseerde volgorde per onderzoeker verschilde om schijnovereenstemming door volgorde-effecten te voorkomen. De onderzoekers konden ook aan het soort opdracht niet herkennen of het om de begin- of eindtekst ging, omdat beide schrijfopdrachten voor zowel de begin- als eindtekst gebruikt werden.

De onderzoekers hadden als taak alle fouten in de teksten aan te strepen en vervolgens per tekst te tellen. Verder dienden de teksten voorzien te worden van een volgnummer, zodat achteraf gecheckt kon worden op volgorde-effecten of de beoordeling geleidelijk strenger of minder streng werd.

Het aantal fouten van iedere beoordelaar per tekst is omgerekend naar het aantal fouten PHW. Voor de toetsing van het effect van het programma via de 2x2 variantie-analyse is gewerkt met het gemiddelde aantal ongestandaardiseerde fouten PHW van beide beoordelaars voor iedere tekst.

Constructie van de Vragenlijst Taalvaardigheid en de schalen

De Vragenlijst Taalvaardigheid werd afgenomen bij de eerstejaarsstudenten International Business and Languages (IBL) van de Hanzehogeschool Groningen. Deze vragenlijst was onderdeel van de begintoets Taalvaardigheid in september 2010 en de eindtoets in december 2010. De studenten moesten tijdens deze toetsen een korte tekst herschrijven. De Vragenlijst Taalvaardigheid bestond uit 46 vragen en werd dus tweemaal ingevuld. Bij de eerste afname deden 75 studenten mee en in totaal kwamen 75 ingevulde lijsten retour. Alleen de resultaten van de eerste afname van de Vragenlijst Taalvaardigheid zijn gebruikt voor de constructie van de schalen.

Subjectieve taalfactor 3SW

De schalen Subjectieve Schrijfvaardigheid, Subjectieve Spelvaardigheid en Subjectieve Woordenschat bleken onderling hoog te correleren met een gemiddelde onderlinge correlatie van .52. De 15 items van deze 3 schalen samengenomen, leverden een alfa van .89 met een gemiddelde onderlinge correlatie van .34. Deze schaal is daarmee zeer betrouwbaar. Zie hierna voor de samenstelling van de subschalen.

Subjectieve Schrijfvaardigheid

De vragen uit de schaal Subjectieve Schrijfvaardigheid waren bedoeld de eigen mening van de studenten te peilen over hun schrijfvaardigheid. Het ging om de volgende 9 items.

1. Het maken van een verslag lukt me altijd wel. nee / ? / ja
14. Ik weet dat ik goed kan schrijven. nee / ? / ja
17. Ik ben niet zo goed in schrijven. nee / ? / ja
20. Het schrijven van een mailtje gaat me gemakkelijk af. nee / ? / ja
25. Ik schrijf vaak fouten die ik zelf niet zie. nee / ? / ja
29. Ik weet dat ik soms moeite met schrijven heb. nee / ? / ja
33. Hoe vaak heb je moeite met schrijfopdrachten?
nooit / soms / regelmatig / vaak / altijd
34. Hoe schrijf je in vergelijking met anderen? slechter / even goed / beter
36. Geef een schoolcijfer (1-10) voor je eigen schrijfvaardigheid: .....

De antwoord-alternatieven 'nee', '?' en 'ja' werden respectievelijk gecodeerd als 0, 1, 2. De vragen 17, 25, 29 zijn omgecodeerd, zodat de antwoord-alternatieven respectievelijk de waarden 2, 1, 0 kregen. De alternatieven van vraag 33 'nooit', 'soms', 'regelmatig', 'vaak' en 'altijd' werden respectievelijk gecodeerd als 2; 1,5; 1; 0,5 en 0. Vraag 34 werd weer gecodeerd als 0, 1 en 2. Bij vraag 36 is het vermelde schoolcijfer ingevoerd. Op deze manier kon op iedere vraag maximaal 2 punten worden gehaald met uitzondering van vraag 36. De alfa-betrouwbaarheid van de 9 items van de schaal Subjectieve Schrijfvaardigheid bedroeg . 83 met een gemiddelde onderlinge correlatie van .38. Dit kan worden beschouwd als betrouwbaar en hoog.

Subjectieve Spelvaardigheid

De vragen van de schaal Subjectieve Spelvaardigheid waren bedoeld de mening van de studenten te peilen over hun eigen spelvaardigheid. Deze schaal bestond uit de volgende items.
15. Ik ben vrij goed in spelling. nee / ? / ja
18. Ik weet dat ik nogal wat spelfouten maak. nee / ? / ja
37. De klas krijgt een dictee. Welk cijfer haal je, denk je? .....

De antwoorden op de geprecodeerde vragen werden opnieuw gecodeerd als 0, 1 en 2. Het antwoord op vraag 37 is als getal van maximaal 10 ingevoerd. Vraag 18 is omgecodeerd. Deze 3 vragen hadden een alfa-betrouwbaarheid van .74 en een gemiddelde onderlinge correlatie van .51. Voor onderzoeksdoeleinden is dit een voldoende betrouwbare schaal.

Subjectieve Woordenschat

De schaal Subjectieve Woordenschat bestond ten slotte uit de volgende 3 items.

2. Ik ken veel moeilijke woorden. nee / ? / ja
12. Ik heb een grote woordenschat. nee / ? / ja
27. Ik kom vaak woorden tegen die ik niet ken. nee / ? / ja

De antwoorden werden respectievelijk gecodeerd als 0, 1 en 2. Vraag 27 werd omgecodeerd. De 3 vragen hadden een alfa-betrouwbaarheid van .69 en een gemiddelde onderlinge correlatie van .43. Voor onderzoeksdoeleinden is deze schaal daarmee voldoende betrouwbaar.

Schrijfattitude

Goed schrijven vereist een bepaalde houding. Dit werd gemeten met de volgende items.

5. Het schrijven van een verslag moet snel gaan. nee / ? / ja
6. Correct schrijven is voor mij belangrijk. nee / ? / ja
7. Wat ik schrijf, moet goed zijn. nee / ? / ja
9. Ik vind schrijven wel leuk. nee / ? / ja
11. Mijn mails zijn meestal nogal kort. nee / ? / ja
13. Ik ben bereid veel tijd in het schrijven van een verslag te steken. nee / ? / ja
22. Spelling vind ik eigenlijk niet zo belangrijk. nee / ? / ja
24. Schrijven gaat bij mij zo snel mogelijk. nee / ? / ja
28. Als ik niet oppas, schrijf ik vaak meer dan mag. nee / ? / ja
30. De spellingschecker haalt de spelfouten er wel uit. nee / ? / ja
35. Hoe vaak maak je uittreksels van te bestuderen stof?

nooit / soms / regelmatig / vaak / altijd

De vragen 5, 11, 22, 24 en 30 zijn omgecodeerd. De codering was weer van 0 tot en met 2. Schrijfattitude heeft 11 items, een coëfficiënt alfa van .68 en een gemiddelde onderlinge itemcorrelatie van 0.19 en is daarmee voor onderzoeksdoeleinden voldoende betrouwbaar.

Schrijfhoeveelheid

Of men veel of weinig schreef, werd geprobeerd te meten met de volgende 4 items.

11. Mijn mails zijn meestal nogal kort. nee / ? / ja
23. Ik heb wel eens een dagboek bijgehouden. nee / ? / ja
32. Hoe vaak mail je? nooit / iedere week / iedere dag / meerdere keren per dag
35. Hoe vaak maak je uittreksels van te bestuderen stof?
nooit / soms / regelmatig / vaak / altijd

De antwoorden voor de achtereenvolgende alternatieven zijn weer gecodeerd op een schaal van 0 tot en met 2: voor 'nee / ? / ja' dus respectievelijk 0, 1, 2. Voor vraag 32 werden de waarden respectievelijk 0; 0.67; 1.33; 2 en voor vraag 35: 0; 0.5; 1; 1.5; 2. Vraag 11 werd omgecodeerd. De coëfficiënt alfa bedroeg .41. De gemiddelde onderlinge correlatie tussen de items bedroeg .13. Op basis van deze 4 items was de meting van de schrijfhoeveelheid, gemeten via coëfficiënt alfa, dus niet betrouwbaar.

Leesschaal

Een 9-tal vragen had betrekking op wat men las en of men veel of weinig las.

8. Kranten vormen voor mij de belangrijkste nieuwsbron. nee / ? / ja
19. Gratis kranten lees ik altijd als ik ze tegenkom. nee / ? / ja
31. Hoeveel lees je? weinig / normaal / veel
41. Hoe vaak per week bezoek je een nieuwssite? .....
42. Hoeveel minuten per dag breng je op nieuwssites door? .....
43. Hoeveel minuten lees je per dag een betaalde krant? .....
44. Hoeveel minuten lees je per dag een gratis krant? .....
45. Hoe vaak lees je per week een betaalde krant? .....
46. Hoe vaak per week lees je een gratis krant? .....

Vragen met geprecodeerde antwoorden werden gescoord van 0 tot en met 2. Bij vragen met open antwoorden werd het vermelde getal overgenomen (afgerond op 1 decimaal). Er zijn geen vragen omgecodeerd. De coëfficiënt alfa van deze 9 vragen bedroeg .70 na standaardisatie, zodat de vragen dezelfde standaarddeviatie kregen. De gemiddelde onderlinge corre-

latie bedroeg .20. Wegens lage gecorrigeerde itemtotaal-correlaties en na inspectie van de items en antwoorden, werden de items 31, 43 en 45 verwijderd. Dit leverde opnieuw een alfa van .70, maar nu met een gemiddelde onderlinge correlatie van .28.

TV-kijken

De vragenlijst bevatte 4 items over tv-kijken.

26. De televisie is voor mij de belangrijkste nieuwsbron.	nee / ? / ja
38. Hoe vaak kijk je per week naar actualiteitenprogramma's?	.....
39. Hoe vaak kijk je per week naar het journaal?	.....
40. Hoeveel uur tv kijk je per dag?	.....

De codering van vraag 26 was weer van 0 tot en met 2. De coëfficiënt alfa van deze schaal bedroeg .64. De gemiddelde onderlinge correlatie tussen de vragen bedroeg .32. De betrouwbaarheid van deze schaal is daarmee voor onderzoeksdoeleinden nog toereikend.

Samenvatting

Om het effect van het nieuwe programma op (de verandering in) het eigen oordeel over de schrijfvaardigheid en de schrijfhouding na te gaan, is de vragenlijst Taalvaardigheid ontwikkeld die 46 vragen bevatte naar de eigen inschatting van de schrijfvaardigheid (3SW-schaal) en naar de schrijfattitude (SA-schaal). Dit waren naast het aantal fouten per honderd woorden de andere afhankelijke variabelen. Verder waren er een aantal vragen naar het lezen, het tv-kijken en hoe men het nieuws bijhield.

De vragenlijst is afgenomen bij de begin- en eindtoets. Aan de eerste afname deden 75 studenten mee. Op basis van deze eerste afname zijn de schalen psychometrisch onderzocht en op grond van de itemanalyse zijn soms items uit een schaal verwijderd of toegevoegd aan een andere schaal.

De subjectieve inschatting van de eigen schrijfvaardigheid werd gemeten met 15 items. De vragen naar de inschatting van de eigen schrijfvaardigheid waren onderverdeeld in drie sub-

schalen: spelvaardigheid, woordenschat, schrijfvaardigheid, die echter belangrijk bleken te correleren en daarom zijn samengenomen. Deze schaal (3SW-schaal) leverde een alfa-betrouwbaarheid van .89. Een voorbeeld van een item was: 'Het maken van een verslag lukt me altijd wel. nee / ? / ja'.

Om de houding tegenover schrijven te meten speciaal met betrekking tot de tijd die men wilde investeren, zijn uiteindelijk 11 vragen geselecteerd (de SA-schaal). Dit leverde een alfa-betrouwbaarheid van .68. Een voorbeelditem is: 'Het schrijven van een verslag moet snel gaan. nee / ? / ja'. Een ander voorbeeld: 'Correct schrijven is voor mij belangrijk. nee / ? / ja'. Studenten die hoog scoorden op schrijfattitude, bleken schrijven vaker leuk te vinden (item 9) en aan te geven spelling belangrijk te vinden (item 22).

De SA-schaal bleek niet te correleren met de 3SW-schaal (r=.00), zodat inderdaad iets anders gemeten werd dan de eigen inschatting. De SA-schaal bleek wel te correleren met de vragen die vroegen naar hoeveel men schreef, de schaal Schrijfhoeveelheid (r=.53).

6.3 Resultaten TAVAN

6.3.1 Dropout en uitval

De klas die het TAVAN-programma kreeg, bestond in het begin uit 27 studenten. Bij de laatste les waren er nog 21 over: 6 studenten stopten in de tussenliggende periode met de opleiding (de dropout). De 21 overblijvende studenten hebben ieder zeven of meer TAVAN-lessen gemaakt. Slechts 14 studenten waren bij alle tien lessen aanwezig. Het gemiddelde aantal gevolgde lessen voor de overgebleven groep van 21 studenten bedroeg 9.3 met een SD van 1.1. Van deze studenten werd van 18 een begin- en eindtekst verkregen. In deze TAVAN-groep van 18 studenten die de gegevens voor het onderzoek leverde, was men gemiddeld 9.4 les aanwezig (SD=1.0).

De drie studenten die tot het einde deelnamen aan het TAVAN-programma, maar die niet een begin- en/of eindtekst inleverden en daardoor uit de TAVAN-groep vielen (de uitval), waren gemiddeld 9.0 les aanwezig. De uitval bleek niet gerelateerd aan het aantal lessen (r=.13, p=.57, tweezijdig, N=21). Deze uitval bleek ook niet gerelateerd aan de TAVAN-score (het matchpercentage): het gemiddelde percentage zinnen bij het doorwerken van de TAVAN-lessen dat matchte met het docentantwoord (r=.03, p=.88, tweezijdig, N=21). Ook bleek er geen verband met vooropleiding (r=.29, p=.20, tweezijdig, N=21).

6.3.2 Lesverloop TAVAN

Eerste uur: foute zinnen herschrijven

De eerste keer begonnen de studenten ongeconcentreerd. Ze praatten met elkaar, rommelden in hun tassen en leken op goed geluk toetsen aan te slaan op de computer. Mogelijk maakten ze daardoor veel fouten bij het verbeteren van de zinnen, wat zichtbaar werd in het matchpercentage (de TAVAN-score) op het scherm. Zodra ze dit begrepen, veranderde hun werkwijze. Ze namen meer de tijd, werkten nauwgezetter en waren geconcentreerd bezig met het verbeteren van de oefenzinnen. Na de eerste les werkten de studenten vrijwel altijd geconcentreerd.

Doordat het online-programma per student onder andere het matchpercentage per les bijhield, viel na te gaan hoe de verschillende lessen gemaakt werden. In Tabel 6.1 zijn voor de 18 studenten van de experimentele groep de gemiddelde matchpercentages weergegeven. De eerste kolom vermeldt het gemiddelde matchpercentage van de student voor alle gemaakte lessen. Daarna volgen de matchpercentages per gemaakte les.

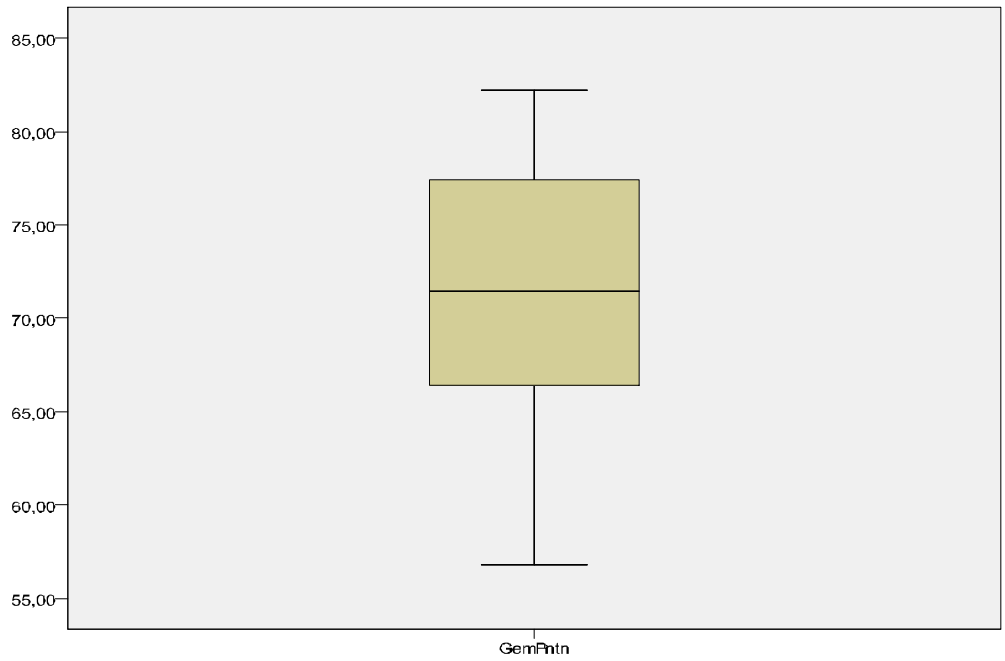
Tabel 6.1 Matchpercentages (MP, percentage goed verbeterde zinnen) per gemaakte TAVAN-les (1-10) voor de 18 studenten uit de experimentele groep met het gemiddelde matchpercentage (Gem) voor alle gemaakte lessen

Gem	MP1	MP2	MP3	MP4	MP5	MP6	MP7	MP8	MP9	MP10
70,7	67	66	78	73	70	76	78	67	67	65
79,2	75	84	85	84	76	78	86	78	72	74
82,2	72	78	85	79	83	85	89	84	84	83
56,8	50	57	73	59	49	68	63	61	42	46
70,7	65	70	82	76	73	71	70	78	63	59
74,6	67	78	75	76	75	79	78	80	71	67
71,4	72	71	85	66	69	77	74	74	55	---
74,1	64	77	86	77	78	79	75	---	57	---
79,6	59	85	82	91	82	80	86	78	73	80
80,5	64	84	88	86	82	83	87	75	77	79
59,1	52	69	78	56	56	64	61	---	---	37
64,0	71	66	74	66	66	63	63	61	58	52
58,0	69	62	72	65	49	54	62	---	47	42
66,4	57	71	83	71	69	67	67	68	57	54
67,2	53	78	69	65	70	72	72	68	63	62
76,0	73	78	86	80	75	78	77	77	70	66
77,4	65	76	81	81	78	83	78	---	---	---
71,4	---	68	71	57	59	66	76	58	59	57

De eerste student maakte tien lessen en haalde een gemiddeld matchpercentage van 71% (afgerond). Van de 1000 door hem bewerkte zinnen, matchte 29% dus niet met de door de docent opgegeven zin(nen). De hoogste score haalde deze student bij les 3 en bij les 7 met 78%. De laagste score haalde hij bij les 10 met 65%.

In Figuur 6.1 is het gemiddelde matchpercentage van de 18 studenten uit de experimentele groep weergegeven via een boxplot. De mediaan van het gemiddelde matchpercentage ligt rond de 71. Het minimum bedraagt 56.8, het maximum 82.2. Het gemiddelde bleek 71.0 te zijn met een SD van 7.7.

Figuur 6.1 Boxplot van het gemiddelde matchpercentage voor de 18 studenten van de experimentele groep



De bedoeling van het TAVAN-programma was dat door het oefenen geleidelijk een grotere beheersing en een hoger niveau zou ontstaan. In dat verband werd gestreefd naar een programma met een hoog percentage goed. Als norm werd gestreefd naar een matchpercentage van 90%. Vergeleken met dit criterium was het TAVAN-programma steeds veel te moeilijk: geen enkele student haalde 90% goed en veel studenten scoorden daar voortdurend ver onder.

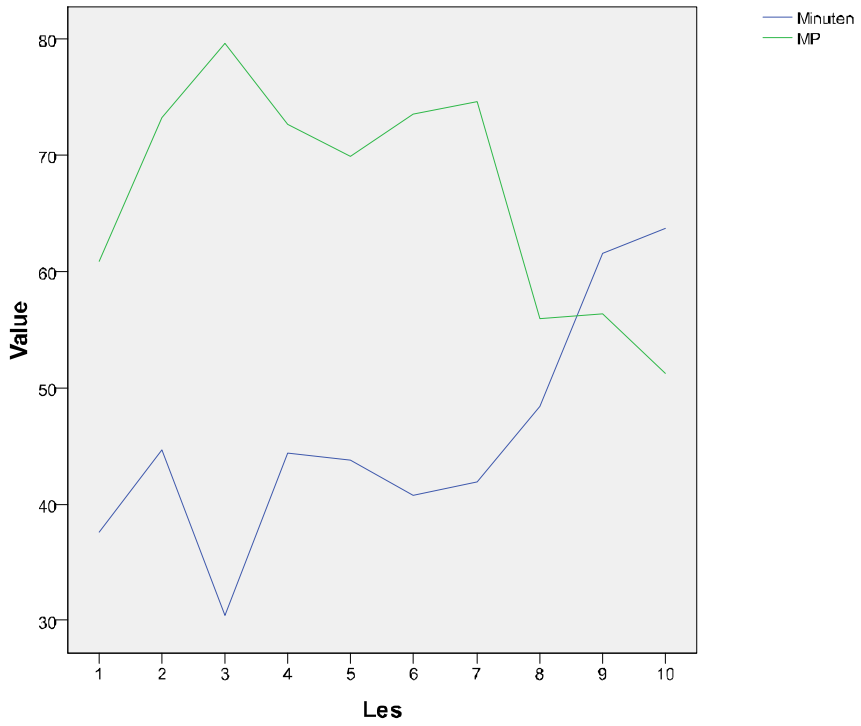
De moeilijkheid van de lessen bleek ook als studenten door afwezigheid een les moesten inhalen. Studenten konden dan zelf via internet inloggen en de online beschikbare les maken. Technisch gezien ging dit probleemloos, maar toch slaagden studenten er doorgaans niet in om voor de afgesproken deadline de honderd zinnen van de les door te werken en te verbeteren. Wanneer studenten bij de TAVAN-les aanwezig waren, leverde het doorwerken van de les niet daadwerkelijk problemen op. De omgeving was dan voldoende dwingend en gestructureerd om de les daadwerkelijk door te werken.

Het online-programma hield ook de tijd bij die de student nodig had om de les door te werken. In Figuur 6.2 zijn weergegeven de gemiddelde tijd per les (in minuten, onderste lijn) die de 18 studenten van de experimentele groep gebruikten en het gemiddelde matchper-

centage van die les. Bij les 1 gebruikten de studenten gemiddeld minder dan 40 minuten, bij les 10 meer dan 60 minuten.

De eerste les bleek relatief slecht gemaakt te worden met een matchpercentage van ongeveer 60%. De lessen 2, 3, 4, 5, 6 en 7 hadden vervolgens een matchpercentage tussen de 70 en 80%. Les 8, 9 en 10 waren daarna veel moeilijker met matchpercentages die terugliepen naar 56 en 51%.

Figuur 6.2 De gemiddelde tijd per les (onderste lijn behalve uiterst rechts) en het gemiddelde matchpercentage per les (bovenste lijn behalve uiterst rechts), (N=18)



Te zien valt dat naarmate studenten gemiddeld meer tijd nodig hadden, een les gemiddeld minder matchende antwoorden leverde (r= -.79, p=.007, tweezijdig, N=10). Bij enkele lessen ging dit echter niet op. Hoewel les 2 meer tijd vergde dan les 1, lag het matchpercentage belangrijk hoger. Studenten ontdekten vermoedelijk in les 1 dat ze te snel antwoord gaven en dat dit niet bijdroeg aan hun matchpercentage. Ook bij les 8 en 9 deed zich een afwijking voor. Afgaande op de bestede tijd was les 9 veel moeilijker, maar ondanks dat de studenten langer nodig hadden, bleef het matchpercentage toch gelijk.

De matchpercentages per les bleken onderling hoog te correleren (gemiddelde correlatie: . 70), zodat het gemiddelde matchpercentage voor de 10 lessen als index voor de basale schrijfvaardigheid van de studenten zeer betrouwbaar was (coëfficiënt alfa: .96, N=14). De score van het online-programma vormde een uiterst betrouwbare meting.

Tweede uur: tekst herschrijven met Word

Bij het tweede uur van iedere TAVAN-les moesten de studenten een stukje tekst met fouten dat op de computer in Word stond, herschrijven tot een foutloze tekst. Ze kregen wekelijks van de docent te horen dat het de bedoeling was om de tekst volledig te verbeteren. Toch lukte dit evenmin als bij het verbeteren van de zinnen. Alleen in de periode dat de tekst verder in de les verbeterd mocht worden tijdens de klassikale bespreking van de fouten (les 1 tot en met 6), lukte het de meeste studenten - overigens pas bij les 6 - om een nagenoeg foutloze tekst in te leveren. Vanaf les 7 werd de herschreven tekst niet meer nagekeken in de klas en mailden de studenten hun herschreven versie meteen naar de docent. Van onmiddellijke feedback was daardoor geen sprake. Bij les 9 moesten de studenten een tekst verbeteren met een omvang van 250 woorden, waarin 82 verbeteringen konden worden aangebracht. De zes beste studenten bij het herschrijven van deze tekst, lieten tussen de 9 en 16 fouten zitten.

Een probleem bij dit uur was dat de taak veel groter was en daardoor minder gestructureerd en dat er niet individueel feedback kon worden gegeven. Ook werden de resultaten niet automatisch geregistreerd.

6.3.3 Aantallen studenten

De klas TAVAN-studenten en de overige twee klassen eerstejaars IBL-studenten kregen een schrijfopdracht bij aanvang van hun studie in september en één vlak voor de kerstvakantie. Van de TAVAN-studenten werd van 18 studenten zowel een begin- als eindtekst verkregen. Als controle werden aselect de teksten van 20 overige studenten getrokken zodanig dat van beide tekstvolgordes, Light-Groningen (L-G) en Groningen-Light (G-L), 10 werden opgenomen.

In Tabel 6.2 zijn de aantallen studenten per subgroep weergegeven. Hoewel de opzet was dat de helft van de TAVAN-groep als begintekst de Groningen-tekst zou krijgen en de andere helft de Light-tekst, bleek dit door dropout en uitval uiteindelijk niet te lukken.

Voor de vragenlijstscores werd uitgegaan van alle 50 studenten van wie de gegevens van de voor- en nameting beschikbaar waren, zodat in dit geval de controlegroep (hierna aan te duiden als: 'overige studenten') groter was en maximaal 32 studenten telde (niet iedere student beantwoordde steeds alle vragen).

De in totaal 76 teksten werden 'blind' beoordeeld door twee onafhankelijk werkende beoordelaars die onderstreepten wat zij als fout zagen en daarna de fouten telden en de teksten voorzagen van een volgnummer om te kunnen checken op volgorde-effecten. Er werden geen significante verbanden (5%, tweezijdig) gevonden tussen het aantal fouten en het volgnummer. Beoordelaars werden dus niet geleidelijk strenger of minder streng.

Tabel 6.2 Aantallen studenten per subgroep. Groningen-Light volgorde (G-L), Light-Groningen volgorde (L-G)

	Programma	**Tekstvolgorde**	**Aantal**
Subgroep 1	Controle	G-L	10
Subgroep 2	TAVAN	G-L	7
Subgroep 3	Controle	L-G	10
Subgroep 4	TAVAN	L-G	11

De overeenstemming (correlatie) tussen de twee beoordelaars voor het aantal fouten in de 76 teksten bedroeg gemiddeld .84. (Groningen-tekst: r=.86; Light-tekst: r=.82.) Dit kwam overeen met een beoordelaars-alfa van gemiddeld .89 (beoordelaarsalfa Groningen-tekst: 0.92; beoordelaarsalfa Light-tekst: 0.85). De beoordelaars stemden daarmee goed overeen.

6.3.4 Validiteit basale schrijfvaardigheid

Vormden de twee teksten een goede maat om basale schrijfvaardigheid vast te stellen? De aantallen gestandaardiseerde fouten PHW bij de begin- en eindmeting correleerden signifi-

cant (r=.51, N=38, p=.001, tweezijdig). De alfa-betrouwbaarheid bedroeg .67. Beide teksten maten met enige betrouwbaarheid dezelfde eigenschap.

De eigen inschatting van de schrijfvaardigheid (3SW, afname 1) bleek significant te correleren met het aantal fouten PHW in begin- en eindtekst. De gemiddelde correlatie bedroeg -.55. Dit punt wordt in 6.3.7 uitgebreider besproken.

Het percentage fouten in de online-lessen bleek hoog te correleren met het gemiddelde aantal gestandaardiseerde fouten PHW in de begin- en eindtekst (r=.79, p=.00)) en zeer betrouwbaar gemeten te kunnen worden (coëfficiënt alfa=.96). Rekening houdend met de geschatte betrouwbaarheid van de begin- en eindtekst (alfa=.67) werd na correctie voor onbetrouwbaarheid een correlatie van .985 gevonden. Deze uitkomst laat zien dat de TAVAN-score en de begin- en eindtekst dezelfde vaardigheid maten.

De TAVAN-score bleek verder een voorspeller van dropout te zijn (r= -.52, p=.01, tweezijdig, N=27). De studenten die laag scoorden op de TAVAN-lessen haakten vaak af en verlieten de opleiding IBL. Bij een TAVAN-score van 70% of hoger bleek geen dropout meer voor te komen. Bij een score onder de 65% bleek bijna de helft van de studenten (44.4%) de opleiding te staken. Deze percentages zijn gebaseerd op een klein aantal (6) dropout-studenten en dus indicatief.

De TAVAN-score per student bleek significant verband te houden met de benodigde tijd (r= -.39, p=.04, tweezijdig, N=27) om de online-lessen te maken. Studenten die lager scoorden, bleken meer tijd nodig te hebben. Na verwijdering van de dropout-studenten bleek dit verband duidelijker te worden (r= -.61, p=.003, tweezijdig, N=21).

De gevolgde vooropleiding (mbo=1, havo/vwo=2) vertoonde een sterk verband met basale schrijfvaardigheid (gemeten via de TAVAN-score). De correlatie bedroeg .60 (p=.001, N=26). De havo/vwo-studenten hadden een grotere basale schrijfvaardigheid.

De gevonden aantallen fouten PHW bleken vergelijkbaar te zijn met de hoge aantallen fouten gevonden in het foutenonderzoek. Voor de controlegroep bleek het gemiddelde aantal fouten op de begin- en eindtekst samen voor alle studenten tussen ongeveer 15 en 25 fouten PHW te liggen met uitzondering van één student die meer dan 30 fouten PHW scoorde. Doordat een van de onderzoekers als beoordelaar fungeerde bij het foutenonderzoek was

een rechtstreekse vergelijking mogelijk tussen de aantallen gesignaleerde fouten PHW in het foutenonderzoek en dit onderzoek. Het gemiddelde voor de hbo-teksten uit het foutenonderzoek was 23.8 PHW (SD=7.9, N=20). Voor de begintekst bedroeg het overeenkomstige gemiddelde 25.4 PHW (SD=4.3, N=38). Een t-toets onafhankelijke steekproeven leverde geen significant verschil op (p=.41, tweezijdig, t[25.1]=-0.84).[1] Het aantal fouten PHW dat voor de herschreven teksten in het TAVAN-onderzoek gevonden werd, week daarmee niet aantoonbaar af van de waarde die gevonden was bij het foutenonderzoek voor door studenten zelf geschreven teksten.

6.3.5 Effect TAVAN op aantal fouten

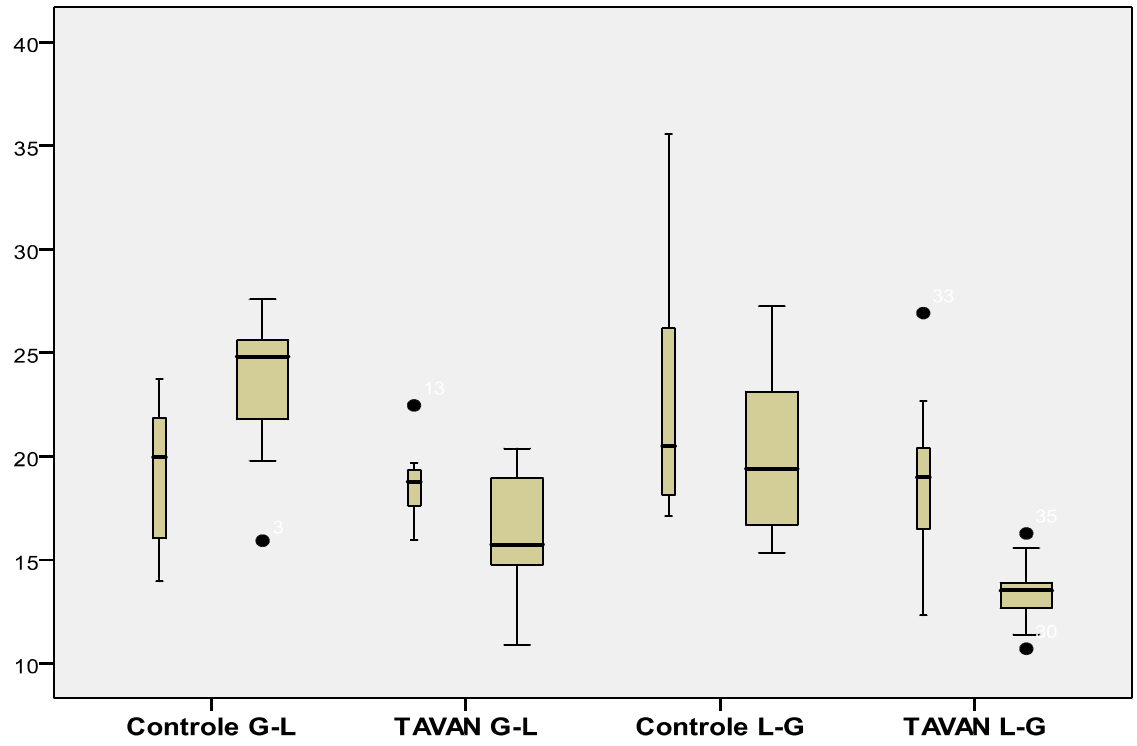
Eerst is gekeken naar een mogelijk verschil van de TAVAN-groep en controlegroep op de begintekst. Een 2x2 variantie-analyse (onderwijsprogramma x begintekst) van het aantal fouten PHW in de begintekst leverde geen significant verschil tussen de TAVAN-groep en de controlegroep. Voor het verschil tussen de controlegroep en de experimentele groep werd p=.108 gevonden (F[1,34]=2.72). Hoewel dit verschil niet significant is, valt niet uit te sluiten dat de groepen iets verschilden doordat de groepen niet door de onderzoekers waren gerandomiseerd, maar bestaande klassen waren, die door de administratie waren gevormd. Door voor de afhankelijke variabele te werken met het verschil tussen eind- en begintekst werden eventuele bestaande verschillen tussen de groepen gecompenseerd.

Voor het verschil tussen de twee teksten werd p=.229 gevonden (F[1,34]=1.50). Ook de interactie tussen deze twee variabelen leverde geen significant verschil: p=.211 (F[1,34]=1.62).

In Figuur 6.3 zijn de aantallen fouten PHW per subgroep voor de begintekst en de eindtekst aangegeven via boxplots. De controlegroep die als begintekst de Groningen-tekst had, vertoonde op de Light-tekst een duidelijke toename in het aantal fouten PHW. De overeenkomstige TAVAN-groep vertoonde een grote afname van het aantal fouten. De controlegroep die als begintekst de Light-tekst had, scoorde op de Groningen-tekst iets minder fouten. De overeenkomstige TAVAN-groep vertoonde echter een zeer grote vermindering van het aantal fouten PHW.

[1] Berekend met *Joosse's Two-sample T-test calculator.* Beschikbaar op: *http://insilico.net/statistics/ttest*

Figuur 6.3 Boxplots van het aantal (ongestandaardiseerde) fouten PHW in de begin- en eindtekst voor de vier subgroepen. De begintekst is weergegeven als smalle box, de eindtekst als brede box.



Een 2x2 variantie-analyse (onderwijsprogramma x tekstvolgorde) van de afname in het aantal fouten PHW tussen begin- en eindtekst leverde een significant effect voor het gevolgde onderwijsprogramma (p=.003, F[1,34]=10.65). De tekstvolgorde maakte ook significant verschil (p=.001, F[1,34]=12.31). Er was geen significante interactie tussen de tekstvolgorde en het onderwijsprogramma (p=.160, F[1,34]=2.06). Aan de assumptie van gelijke varianties werd volgens de Levene test voldaan (p=.482). Verder leken de histogrammen per subgroep redelijk normaal.

In Tabel 6.3 zijn de gemiddelde aantallen fouten PHW en standaarddeviaties van de verschillende subgroepen weergegeven. Verder is het geschatte randgemiddelde van de totale controlegroep en de totale TAVAN-groep vermeld met de gepoolde SD.

Tabel 6.3 Het gemiddelde aantal fouten PHW voor de vier subgroepen in begintekst, eindtekst en het verschil (begintekst - eindtekst). De totalen zijn het ongewogen gemiddelde van beide subgroepen ('geschatte randgemiddelden'). Tussen haakjes: de standaarddeviatie. Bij de totalen is dit de gepoolde standaarddeviatie van beide subgroepen.

Conditie, tekstvolgorde	Begintekst	Eindtekst	Verschil	N
Controlegroep, G-L	19.25 (3.30)	23.44 (3.49)	-4.19 (4.82)	10
TAVAN-groep, G-L	18.73 (2.14)	16.34 (3.31)	2.38 (3.05)	7
Controlegroep, L-G	22.71 (5.86)	19.98 (3.90)	2.73 (3.81)	10
TAVAN-groep, L-G	18.66 (4.08)	13.38 (1.62)	5.28 (4.67)	11
Controlegroep, totaal	**20.98 (4.34)**	**21.71 (3.70)**	**-0.73 (4.35)**	**20**
TAVAN-groep, totaal	**18.69 (3.48)**	**14.86 (2.40)**	**3.83 (4.14)**	**18**

De TAVAN-groep maakte 3.83 fout PHW minder in de eindtekst dan in de begintekst, een vermindering van 20.5%. Omgerekend naar een A4 (500 woorden) kwam dit neer op 19 fouten minder. De controlegroep maakte 0.73 fout PHW meer in de eindtekst dan in de begintekst, een toename van 3.5%. Omgerekend naar een A4-tekst kwam dit neer op 3.5 fouten meer.

De TAVAN-groep verbeterde 4.56 fout PHW meer dan de controlegroep. De grootte van het verschil kwam overeen met 1.05 SD (van de controlegroep). Dit geldt als een groot effect. De verwachting dat de TAVAN-groep meer leerwinst zou realiseren dan de controlegroep werd daarmee bevestigd.

Gestandaardiseerde leerwinst

Bij de hiervoor besproken variantie-analyse van de leerwinst bleek de tekstvolgorde uit te maken (p=.001). De Light-tekst was moeilijker dan de Groningen-tekst en daardoor leken de groepen die de Light-tekst als begintekst hadden, meer vooruit te gaan dan de groepen die de Groningen-tekst als begintekst hadden. Het verschil in gemiddelden tussen de L-G en de G-L groepen bleek 4.92 fout PHW te bedragen. Voor het berekenen van de correlaties

met de (gestandaardiseerde) leerwinst is dit teksteffect verwijderd door de teksten per beoordelaar te standaardiseren zoals vermeld is in 6.2.

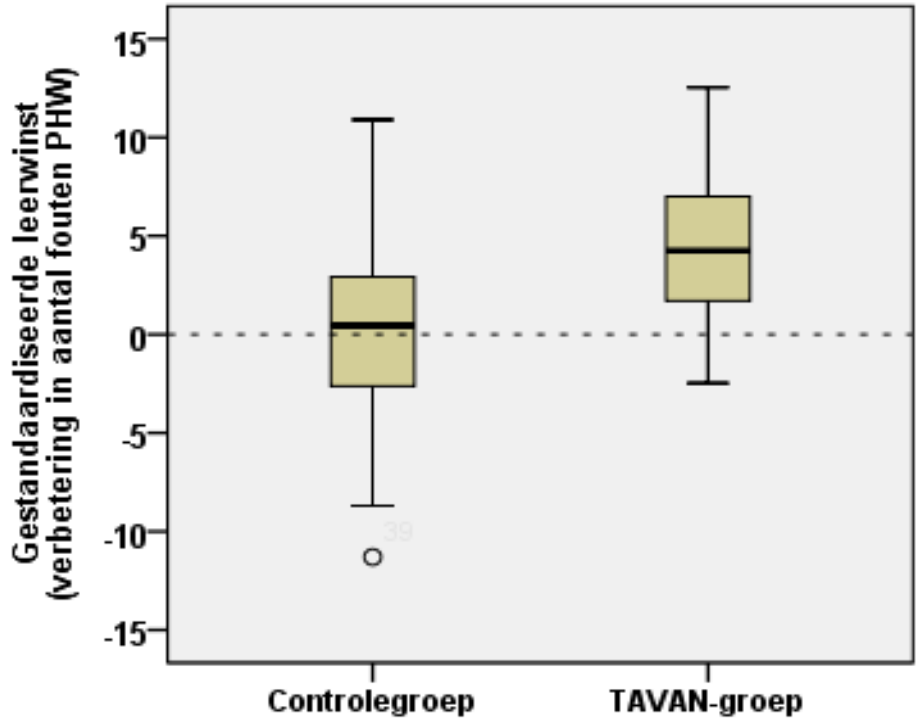
Om de standaardisering te checken is de correlatie berekend tussen tekstvolgorde (gecodeerd als: 1=G-L, 2=L-G) en de gestandaardiseerde leerwinst. Deze was vrijwel nul (-.04). Zonder standaardisering bleek de correlatie .51 te zijn. De standaardisering leidde dus inderdaad tot verwijdering van het volgorde-effect.

In Figuur 6.4 is de gestandaardiseerde leerwinst (de afname in aantal gestandaardiseerde fouten PHW tussen begin- en eindtekst) voor de controlegroep en de TAVAN-groep weergegeven in twee boxplots. De controlegroep bleef vrijwel gelijk, in de TAVAN-groep maakte vrijwel iedereen minder fouten PHW in de eindtekst.

De schrijfvaardigheid gemeten via de TAVAN-score bleek ongerelateerd (r= -.04) te zijn aan de gestandaardiseerde leerwinst. Studenten die beter dan gemiddeld waren in het herschrijven van foute zinnen verbeterden gemiddeld evenveel als studenten die slechter dan gemiddeld waren.

De grootte van de gestandaardiseerde leerwinst bleek in de TAVAN-groep significant samen te hangen met de eigen inschatting van schrijfvaardigheid bij de begintekst (r= -.58, p=.01, tweezijdig, N=18) en tv kijken (r= -.81, p=.000, tweezijdig, N=17), maar niet met andere variabelen uit de vragenlijst. De studenten die hun eigen schrijfvaardigheid lager dan gemiddeld inschatten en minder dan gemiddeld tv keken, bleken meer vooruitgang te boeken. De eigen inschatting van de schrijfvaardigheid en het tv-kijken bleken echter duidelijk samen te hangen (r=.59, p=.01, tweezijdig). Studenten die relatief veel tv kijken waren vaak positiever over de eigen schrijfvaardigheid, maar boekten minder vooruitgang.

Figuur 6.4 Boxplots voor de gestandaardiseerde leerwinst, de afname van het aantal gestandaardiseerde fouten PHW tussen begin- en eindtekst bij de controlegroep en de TAVAN-groep. De controlegroep scoorde rond de nullijn en verbeterde in doorsnee niet. De TAVAN-groep scoorde vrijwel volledig boven de nullijn: vrijwel iedereen verbeterde.



6.3.6 Effect TAVAN op schrijfattitude

De schrijfattitude (of men veel of weinig tijd wilde investeren in het schrijfproces) bleek niet significant te correleren met de begin- en eindtekst, met de TAVAN-score, met de eigen inschatting van de schrijfvaardigheid en met de gestandaardiseerde leerwinst. Studenten die relatief weinig fouten PHW maakten, scoorden dus niet beter qua schrijfattitude dan studenten die veel fouten PHW maakten.

De TAVAN-groep bleek wat betreft verandering in schrijfattitude niet te verschillen van de groep overige studenten (t-toets, p=.74, tweezijdig, t[49]=.34). Ook ten opzichte van zichzelf bleken beide groepen qua schrijfattitude gemiddeld niet veranderd te zijn (TAVAN-groep: p=.89, t[17]=.14; controlegroep: p=.42, t[32]=.82; steeds tweezijdig, t-toets gepaarde

steekproeven). De verwachting dat door deelname aan het TAVAN-programma de schrijfattitude zou toenemen, werd daarmee niet bevestigd.

Wel bleek dat studenten die beter dan gemiddeld scoorden in het online-programma qua schrijfattitude verslechterden, terwijl studenten die slechter scoorden dan gemiddeld wat betreft schrijfattitude juist verbeterden (r=.55, p=.02, tweezijdig, verschilvariabele). De verandering in schrijfattitude ging echter niet samen met de gestandaardiseerde leerwinst (de verbetering in het aantal gestandaardiseerde fouten PHW). Deze correlatie bedroeg -.004 (p=.98, tweezijdig).

6.3.7 Effect TAVAN op inschatting eigen schrijfvaardigheid

Het aantal gestandaardiseerde fouten PHW in de begintekst correleerde significant met het oordeel over de eigen schrijfvaardigheid bij de begintekst, de 3SW-schaal (r= -.67, p=.000, beginmeting). Voor de eindtekst was de correlatie lager, maar nog steeds significant (r= -.43, p=.008, tweezijdig, eerste afname 3SW-schaal). Gemiddeld kwam de correlatie daardoor uit op -.55. Dit verband laat zien dat het eigen oordeel van de studenten globaal overeenstemde met de gemeten basale schrijfvaardigheid, maar niet perfect.

Voor de verandering in de subjectieve inschatting van de eigen schrijfvaardigheid (het verschil tussen de begin- en eindmeting van 3SW) werd een significant verschil (t-toets, p=.03, tweezijdig, t[48]= -2.27) gevonden tussen de groep overige studenten en de TAVAN-groep. De groep overige studenten bleek de eigen schrijfvaardigheid bij de eindtekst significant hoger in te schatten dan bij de begintekst (t-toets gepaarde waarnemingen, p=.01, tweezijdig, t[32]= -2.67). De groep TAVAN-studenten schatte de eigen schrijfvaardigheid bij de eindmeting lager in dan bij de begintekst, maar dit was niet significant (t-toets gepaarde waarnemingen, p=.36, tweezijdig, t[16]=.95). De derde verwachting, dat de TAVAN-studenten zichzelf gemiddeld mogelijk negatiever zouden gaan inschatten door het oefenen en de feedback, werd daarmee niet bevestigd.

De groep overige studenten werd (afgaande op de resultaten van de controlegroep) qua aantal fouten PHW niet beter, maar ging zichzelf wel positiever inschatten. De groep TAVAN-studenten werd in werkelijkheid wel beter, maar ging zichzelf niet positiever inschatten. De eigen inschatting van de schrijfvaardigheid was dus in beide groepen geen goede indicator

voor de werkelijke basale schrijfvaardigheid. In Tabel 6.4 zijn de gemiddelden en SD's voor Schrijfattitude en de eigen inschatting van de schrijfvaardigheid (3SW) per groep vermeld.

Hoewel de subjectieve inschatting van de eigen schrijfvaardigheid bij de TAVAN-groep gemiddeld genomen niet aantoonbaar veranderde, bleek er wel een duidelijk verband te bestaan tussen de TAVAN-score en de verandering in eigen inschatting. Als studenten bij TAVAN laag scoorden (weinig items goed), bleek de subjectieve inschatting te dalen. Omgekeerd was het zo dat bij studenten met een hoge score in het online-programma de subjectieve inschatting steeg (r= -.65, p=.01, tweezijdig, verschilvariabele). Door de score uit het online-programma te vergelijken met de score van andere studenten, ging men zichzelf realistischer inschatten. Dit bleek ook uit de correlatie van de eigen inschatting met de TAVAN-score. Bij de beginmeting was deze .29 (p=.25, tweezijdig), bij de eindmeting .65 (p=.005, tweezijdig).

Tabel 6.4 Gemiddelden met tussen haakjes de standaarddeviaties voor beide metingen van Schrijfattitude (SA), het verschil daarvan, en de subjectieve inschatting van de eigen schrijfvaardigheid (3SW) en het verschil daarvan

	TAVAN	**Overige**
Schrijfattitude afname 1	6.33 (2.28)	7.21 (1.29)
Schrijfattitude afname 2	6.28 (2.35)	7.00 (1.60)
SA (afname 1 - afname 2)	**0.06 (1.73)**	**0.21 (1.49)**
Subjectieve inschatting afname 1	36.24 (4.90)	31.46 (6.44)
Subjectieve inschatting afname 2	35.32 (7.11)	33.08 (7.01)
3SW (afname 1 - afname 2)	**1.22 (5.32)**	**-1.61 (3.47)**

6.4 Conclusies en discussie

Welke conclusies kunnen uit het onderzoek naar de effectiviteit van het TAVAN-programma worden getrokken? Een aantal conclusies heeft een meer algemeen karakter en een aantal heeft te maken met de verdere ontwikkeling en verbetering van het TAVAN-programma.

Het nieuwe schrijfvaardigheidsprogramma werd ontwikkeld om het grote aantal fouten dat eerstejaars hbo-studenten produceren in hun Nederlandse teksten terug te dringen. Het effectiviteitsonderzoek laat zien dat het programma op dit punt zeer succesvol was. Het aantal fouten van de TAVAN-groep daalde met 20%, terwijl de controlegroep die het traditionele programma volgde in de eindtekst niet verbeterde. De eerste verwachting werd daarmee bevestigd.

Het TAVAN-programma toont aan dat de hoge aantallen fouten die gevonden worden in schrijfproducten van studenten met een relatief korte training (twintig lesuur) aanzienlijk gereduceerd kunnen worden. Basale schrijfvaardigheid blijkt belangrijk en snel verbeterd te kunnen worden door te oefenen met het herschrijven van foute zinnen en teksten die veel fouten bevatten.

Het TAVAN-programma is geconstrueerd vanuit het ABC-leermodel dat er vanuit gaat dat basale schrijfvaardigheid (net als andere vaardigheden) ingeoefend moet worden. Studenten leren door te doen en snelle en duidelijke feedback is daarbij van groot belang. De positieve resultaten van het TAVAN-programma wijzen erop dat het ABC-model een goed uitgangspunt kan zijn voor het ontwikkelen van effectief schrijfonderwijs.

In het TAVAN-programma speelde het online-programma een belangrijke rol. Het online-programma werkte structurerend, doordat het steeds automatisch de volgende foute zin aanbood aan de student. Het zorgde per student voor snelle en duidelijke feedback en het hield per student de resultaten bij. Door het online-programma konden studenten (binnen bepaalde grenzen) in hun eigen tempo werken. Online-programma's bieden daarmee mogelijkheden voor effectief onderwijs die niet op andere manieren gerealiseerd kunnen worden.

Interessant in dit verband is dat docenten die enthousiast waren over de mogelijkheden van de computer in het taalonderwijs (Jager, 2009, p. 174) belangrijk anders over de voordelen van computers bleken te denken dan hierboven vermeld voor TAVAN. Dit blijkt uit de IILL-enquête (ICT-Integrated-Language-Learning-enquête). Op een lijst van 18 voordelen van computergebruik in het onderwijs gerangordend via de gemiddelde score van laag (niet mee eens) naar hoog (zeer mee eens), scoorden de stellingen die het TAVAN-programma goed lijken te beschrijven (monitoring student progress, repeated practice, raised student level, personal feedback, time on task) zeer laag, respectievelijk op plaats 2, 3, 6, 9 en 10 (p. 162-164). Gemiddeld kwamen deze items daarmee uit op de zesde plaats, ver onder het ranggemiddelde van 9.5.

Voordelen van het gebruik van de computer die van de respondenten de meeste instemming kregen, hadden betrekking op: het werken door de studenten waar en wanneer hen dat het beste uitkomt, de grotere zelfstandigheid van de studenten, de grotere authenticiteit van het onderwijs en de grotere aantrekkelijkheid van het leren (Jager, 2009, p. 162-164). Verder werd de computer niet zozeer in de les gebruikt, maar vooral door studenten buiten de les. Slechts 16% van de respondenten stelde dat de computer de meeste tijd in de klas werd gebruikt (p. 157). Hieruit blijkt een duidelijk verschil tussen de TAVAN-benadering en hoe taaldocenten de voordelen van de computer zien.

De enquête gaf ook informatie over hoe de docenten onderwijs in het algemeen zagen. De stelling dat taalstudenten zelf verantwoordelijkheid moeten nemen voor hun leren werd door 94% van de respondenten beaamd. De stelling: "It is the role of the teacher to impart knowledge to the student through such activities as explanation, example and practice," werd door 75% van de respondenten beaamd (Jager, 2009, p. 169-171). Hoewel de laatste stelling een oefenaanpak niet volledig uitsluit, lijkt het accent daarmee toch vooral te liggen op uitleg en kennisoverdracht. Dit aspect kwam in het TAVAN-programma alleen in de feedbackfase zeer beperkt aan de orde door het verstrekken van het juiste antwoord na een fout antwoord.

Dat de TAVAN-benadering kennelijk weinig populair was, bleek ook uit de formulering van de gebruikte vragen: in totaal werd naar het gebruik van 6 verschillende soorten educatief bruikbare programma's gevraagd (Tutorial Software, Resources, Asynchronous Communication, Synchronous Communication en Web Publication and Collaboration), maar niet naar het gebruik van Drill and Practice (Lamie, 2000), de categorie waar een programma als TAVAN toebehoort.

Het vooraf geformuleerde idee dat studenten veel fouten produceren doordat het hun ontbreekt aan een juiste schrijfattitude, bleek niet te kloppen. De schrijfattitude bleek niets te zeggen over hoe goed men schreef, uitgaande van de TAVAN-score en het aantal fouten PHW in begin- en eindtekst. De verwachting dat de TAVAN-studenten gemiddeld qua schrijfattitude zouden verbeteren, werd niet bevestigd. Wel bleek dat de studenten die slecht scoorden in het online-programma een betere schrijfattitude ontwikkelden, terwijl studenten die goed scoorden een slechtere schrijfattitude ontwikkelden. Deze verandering in schrijfattitude bleek echter niet samen te hangen met de vooruitgang die men boekte, de gestandaardiseerde leerwinst. Schrijfattitude is daarmee niet van invloed op hoe goed men werkelijk schrijft.

De verklaring voor dit resultaat is vermoedelijk als volgt. Als lezer of beoordelaar ziet men in de tekst een irritante fout staan. Vervolgens neemt men aan dat de schrijver die fout ook gezien heeft. Waarom is die fout niet even gecorrigeerd, vraagt men zich af. Kennelijk ontbrak het de schrijver aan inzet en motivatie. Uit het foutenonderzoek blijkt echter dat zelfs expert-beoordelaars soms meer dan vijftig procent van alle fouten niet zien. Het probleem is niet dat de student onwillig was de fout te corrigeren, het probleem is dat hij de fout niet zag. Deze verklaring wordt bevestigd door het gegeven dat studenten die veel fouten maken, juist meer tijd uittrekken voor het herschrijven. De attitude dat men veel tijd wil uittrekken voor het schrijven helpt niet om daadwerkelijk beter te schrijven. Schrijven met weinig fouten is niet een kwestie van een bepaalde houding, maar een bepaalde vaardigheid. Het is niet een kwestie van niet willen, maar van niet geleerd hebben.

De eigen inschatting van de schrijfvaardigheid bleek enigszins overeen te stemmen met de gemeten basale schrijfvaardigheid via de begin- en eindtekst, maar geen valide maat te zijn voor schrijfvaardigheid. Zo bleek de eigen inschatting te kunnen stijgen zonder dat de basale schrijfvaardigheid verbeterde, terwijl omgekeerd de basale schrijfvaardigheid kon verbeteren zonder dat de eigen inschatting steeg. De verwachting dat deelname aan het TAVAN-programma door negatieve feedback zou leiden tot een lagere inschatting van de eigen schrijfvaardigheid, werd niet bevestigd. Studenten die slecht scoorden in het TAVAN-programma bleken hun schrijfvaardigheid lager te gaan inschatten, terwijl studenten die goed scoorden ten opzichte van het gemiddelde zichzelf juist hoger gingen inschatten. Het effect van het TAVAN-programma was daarmee dat de studenten zichzelf realistischer gingen inschatten ten opzichte van het groepsgemiddelde.

Het herschrijven van teksten zoals dat bij de begin- en eindmeting gevraagd werd, bleek niet gemakkelijk te zijn. Studenten maakten bij het herschrijven ten minste evenveel fouten PHW als in de meer vrije teksten. Herschrijfopdrachten in de vorm van een zin met fouten of in de vorm van een tekst met fouten, lijken daarmee een goede manier te zijn om basale schrijfvaardigheid te bepalen.

Basale schrijfvaardigheid bleek op verschillende manieren vastgesteld te kunnen worden: door het beoordelen van door studenten geschreven teksten, door studenten teksten met fouten en gebreken te laten herschrijven en door studenten zinnen te laten herschrijven in het online-programma. Deze laatste manier bleek uitermate betrouwbaar te zijn, zeer hoog te correleren met het aantal fouten PHW in de begin- en eindtekst samen en geen menselijke

beoordelaar(s) te vergen. Het online-programma vormt daarmee een eenvoudige, betrouwbare en valide manier om basale schrijfvaardigheid vast te stellen.

Uit het onderzoek blijkt dat basale schrijfvaardigheid gemeten via de begin- en eindtekst of via de TAVAN-score een belangrijke variabele is, die zich gedraagt zoals men zou verwachten. Zo bleek de eigen inschatting van de studenten enigszins overeen te stemmen met de gemeten basale schrijfvaardigheid. Verder bleek dat studenten met een slechte basale schrijfvaardigheid meer tijd nodig hadden voor het herschrijven van de foute zinnen dan de studenten die goed konden schrijven. Studenten met havo/vwo als vooropleiding bleken een betere basale schrijfvaardigheid te hebben dan studenten met mbo. Ten slotte bleek basale schrijfvaardigheid (gemeten via de TAVAN-score) een goede voorspeller van dropout te zijn.

Het lijkt mogelijk het TAVAN-programma op een aantal punten nog te verbeteren. De online-lessen waren te moeilijk en mogelijk ook vrij lang. Het streefpercentage van 90% items 'goed', werd bij geen enkele student ooit gehaald. Een eerste mogelijkheid om TAVAN te verbeteren, lijkt daarmee het aanpassen en daarna opnieuw testen van de effectiviteit van de online-lessen. De bedoeling daarvan is al te steile 'hellingen' waar studenten op vastlopen minder steil te maken waardoor het doorwerken van de lessen minder frustrerend wordt, terwijl de totale effectiviteit van de lessen daardoor nog groter kan worden.

Tijdens het tweede uur van TAVAN ontbrak rechtstreekse feedback en was mogelijk ook de structurering onvoldoende: de herschrijftaak was vrij omvangrijk. Verder werden de resultaten van de studenten niet duidelijk zichtbaar. Mogelijk zou ook dit uur via het online-programma gestructureerd kunnen worden of zou het nakijken van de herschreven teksten op andere wijze geautomatiseerd kunnen worden. Herstructurering van het tweede uur is daarmee een tweede mogelijkheid om TAVAN te verbeteren. Ook deze aanpassing zal vervolgens onderzocht moeten worden op effectiviteit.

Het TAVAN online-programma mat nu niet rechtstreeks de vooruitgang van de student: de TAVAN-score per les was ook afhankelijk van de moeilijkheid van de desbetreffende les. De enige manier waarop het uiteindelijke leerresultaat werd vastgesteld, was aan het einde van het studie-onderdeel via het aantal fouten PHW in de herschreven eindtekst. Doordat de online-les echter ook voortdurend het niveau van de student meet, is het in beginsel mogelijk rekening te houden met de moeilijkheid van de les en vanuit de TAVAN-score het ver-

wachte eindniveau op de eindtekst te voorspellen. Dit maakt de vooruitgang van de student sneller en duidelijker zichtbaar en maakt de afhankelijkheid van de eindtekst voor het bepalen van de gestandaardiseerde leerwinst, kleiner. Voor studenten zou deze informatie verder extra motiverend kunnen werken.

In het online-programma werd nu geen mastery learning (blokkadesysteem) toegepast. Zo kon het gebeuren dat een student na een gemiste les doorging met de volgende les zonder eerst de gemiste les in te halen. Deze mogelijkheid zou geblokkeerd kunnen worden. Verder kwamen fout beantwoorde items niet terug. Een student werd dus niet gemotiveerd na een fout antwoord het 'goede' antwoord te bestuderen en zich de fout met de correctie eigen te maken. Dit zou in de online-lessen aangepast kunnen worden.

Voor het onderzoek naar de effectiviteit van het TAVAN-programma werden de begintekst en de eindtekst gebruikt waarvan de herschreven teksten door twee beoordelaars werden beoordeeld. Uitgaande van het gemiddelde aantal gesignaleerde fouten PHW in de begintekst en één minuut per fout kostte het beoordelen van deze teksten een vier uur per student. Verder bleek de Light-tekst soms zo moeilijk dat studenten erin vastliepen wat ook de beoordeling bemoeilijkte en mogelijk negatief inwerkte op de betrouwbaarheid. Hoewel het belangrijk lijkt dat studenten soms met complete teksten werken en de beoordeling van de effectiviteit van het TAVAN-programma valide moet zijn, lijkt het mogelijk deze meetmethode te verbeteren. Als eerste zou met name de Light-tekst aangepast kunnen worden. Een tweede optie is mogelijk de beoordeling van dit soort teksten geheel of gedeeltelijk te automatiseren. De betrouwbaarheid en validiteit van de beoordeling zou daarna opnieuw onderzocht moeten worden.

Een mogelijk probleem met betrekking tot het TAVAN-programma lijkt de implementatie. Een eerste punt is dat TAVAN een groot beroep doet op de beschikbare ruimte voor computerpractica. Een vermoedelijk nog belangrijker punt wordt door de begeleiding gevormd. Docenten moeten overstappen op een systeem van lesgeven waarbij zij ogenschijnlijk passief zijn en de studenten actief. Dit lijkt volledig strijdig te zijn met de gebruikelijke opvatting over doceren. De TAVAN-benadering gaat immers uit van een gestructureerde en geïndividualiseerde methode waarbij de docent wel faciliterend aanwezig is, maar ogenschijnlijk niet actief is. Het is de vraag of de betrokken docenten deze nieuwe taakinvulling als een verbetering ervaren en hiertoe bereid en in staat zijn. Tegelijkertijd is hun aanwezigheid en inzet een voorwaarde voor het goed functioneren van TAVAN.

Vervolgens is er op een overkoepelend niveau een docent-coördinator nodig die de werking van het leersysteem organiseert, controleert en evalueert. Deze docent-coördinator moet ervoor zorgen dat het onderdeel ingeroosterd wordt door de roostermaker, de cijfers geadministreerd worden, het online-programma beschikbaar en bereikbaar is, de studenten een geldige inlogcode hebben, de links naar de server werken, de registratiefiles gebackupt worden, de lessen bijgesteld worden, de toetsing geregeld wordt, de studenten gemotiveerd worden, de practicumdocenten aangestuurd worden en door het online-programma niet goed nagekeken items worden gemeld en gecorrigeerd. Verder moet de effectiviteit van het totale leersysteem gecontroleerd worden. Dit soort activiteiten kosten wel veel tijd, maar vormen geen contacturen zodat hier vaak geen uren voor ter beschikking worden gesteld. Verder kan het verschil in taak tussen de docent-coördinator en de practicumdocenten problemen opleveren.

Het werken met het TAVAN-programma doet dus een groot beroep op de computerfaciliteiten, vraagt een andere onderwijsopvatting van de docenten dan gebruikelijk is en vergt een belangrijk andere organisatie van het onderwijs. Deze factoren kunnen de invoering bemoeilijken.

7

Deelstudie 4

Effect van fouten

7.1 Inleiding

Hoe erg is een taalfout? De meningen hierover verschillen. Sommige mensen zijn er rotsvast van overtuigd dat taalfouten een groot negatief effect hebben, dat zich niet alleen beperkt tot de inhoud van de tekst, maar zelfs kan terugslaan op degene die de fout maakte. Vanuit dit oogpunt is iedere fout een fout te veel. Docenten kijken bijvoorbeeld een tekst na en strepen fouten aan. Hun oordeel is normaal niet vatbaar voor discussie: een fout is iets absoluuts waar niet over te twisten valt en iets dat volledig verkeerd is.

Andere docenten denken dat taalfouten er niet zoveel toe doen, omdat het om de inhoud en het doel van de tekst gaat. Een paar taalfouten meer of minder zouden daarbij niet zo'n belangrijke rol spelen. Taalfouten maken weliswaar soms een zin ambigu en daardoor de interpretatie lastig, maar vaak ook niet. Zo doet een interpretatieprobleem zich niet voor in een zin als 'Het product word gekocht'. Het lijkt hoe dan ook moeilijk om op een neutrale manier naar taalfouten te kijken. Wall en Hull formuleerden het als volgt: "Error and correctness in writing, as topics of research and components of language instruction, have the uneasy characteristic of being quite value-laden" (Wall & Hull, 1989, p. 261).

In de eerste drie delen van dit onderzoek spelen taalfouten in teksten van studenten en het verbeteren van de schrijfvaardigheid op dat punt een belangrijke rol. Daarom leek het zinvol in dit vierde deel de invloed van fouten te onderzoeken en na te gaan in welke mate taalfouten de tekst aantasten.

De vraag wat taalfouten zijn, kwam in deelstudie 1 (Foutenonderzoek) ook aan de orde. Het is een vraag die niet zo gemakkelijk te beantwoorden is als op het oog lijkt. Uit bestaand foutenonderzoek komt namelijk niet het beeld naar voren dat een fout een onbetwistbaar iets is dat door iedereen herkend zal worden. Wall & Hull (1989) verbaasden zich over de zekerheid van wat zij 'ervaren lezers' noemden over wat voor een fout doorgaat en hoe zwaar die aangerekend moet worden: "the assumption is that sentence-level error in writing is a simple matter to measure" (Wall & Hull, 1989, p. 262). Het onderzoek van Wall en Hull liet juist zien hoe gering de overeenstemming tussen beoordelaars over fouten was. Taalfouten kunnen door hun afwijking van hoe het hoort ook een emotioneel effect hebben op de lezer. Zo constateerde Williams (1981) in zijn bespiegeling over wat een 'fout' was: "I am puzzled why some errors should excite this seeming fury" (Williams, 1981, p. 152). Rose (1985) ergerde zich aan het beoordelen van schrijfvaardigheid op basis van het aantal fou-

ten in teksten, want dat noemde hij: "an inaccurate assessment of student ability and need" (Rose, 1985, p. 341). Robinson (1998) zag fouten in teksten van studenten niet zozeer als obstakels die een ideale tekst in de weg staan en afgestraft zouden moeten worden, maar als fases in een leerproces. Anson (2000) beschouwde het begrip 'fout' als een product van regels die voortdurend veranderen. Nieuwenhuijsen (2011) maakte een onderscheid tussen taalfouten en vergissingen. In verzorgde teksten horen volgens hem geen vergissingen te staan, maar er kunnen wel taalfouten in staan, "want dat etiket dekt een aantal uiteenlopende ladingen" (Nieuwenhuijsen, 2011, p. 210).

Ook lijkt een fout tijdgebonden te zijn. 'Wrong word' was een fout die bij Lunsford en Lunsford in 2008 bovenaan de lijst stond, maar in 1988 stond dezelfde fout bij Connors en Lunsford op de vierde plaats. 'Onvolledige of ontbrekende bronvermelding' werd in 2008 beschouwd als een formele fout die op de derde plaats stond, maar in eerder onderzoek kwam deze fout helemaal niet voor. Connors en Lunsfor (1988) vonden echter dat fouten er wel toe doen: "The world judges a writer by her mastery of conventions, and we all know it" (Connors & Lunsford, 1988, p. 396). In paragraaf 7.1.3. wordt verder ingegaan op het betrekkelijke karakter van taalfouten.

Met het 'effect van fouten' kunnen verschillende dingen bedoeld worden. Het is duidelijk dat teksten iets moeten overbrengen, zoals informatie, overtuigingen, gevoelens, want teksten worden niet in het luchtledige geschreven. Steehouder et al. (2006) noemden vier soorten boodschappen: de appellerende boodschap (wat wil de zender bereiken), de referentiële boodschap (wat wordt over het onderwerp gezegd), de expressieve boodschap (wat zegt de uiting over de zender) en de relationele boodschap (hoe kijkt de zender tegen de ontvanger aan). Daarnaast onderscheidden zij in het communicatieproces vijf soorten, elkaar overlappende, doelen die de zender kan hebben met zijn boodschap: informeren, instrueren, overtuigen, motiveren en gevoelens opwekken. Onderzocht zou kunnen worden of een tekst met veel fouten daarin minder succesvol is. Ook kan gekeken worden of taalfouten in teksten invloed hebben op de intentie van de lezer om iets al dan niet te doen - waarbij het overigens de vraag blijft of die intentie inderdaad in gedrag omgezet wordt - of naar de invloed van fouten op de status van degene die de tekst schrijft. In 7.1.1 en 7.1.2 wordt hierop ingegaan.

Dit onderzoek richt zich op de waardering van een tekst door de lezer. De reactie van de lezer op de tekst bepaalt wat de tekst overgedragen heeft, zodat het effect van taalfouten een

kwestie is die in beginsel empirisch onderzocht kan worden. Het gaat hier dan ook om de vraag in hoeverre taalfouten in een tekst effect hebben op de waardering van die tekst door de lezer.

Deze vraag is op drie manieren uitgewerkt met correlationeel en experimenteel onderzoek. Uit het foutenonderzoek (deelstudie 1) bleek dat het aantal bevestigde fouten per honderd woorden relatief betrouwbaar kan worden vastgesteld, maar dit hoeft nog niet te betekenen dat het aantal fouten in een tekst van invloed is op het holistische oordeel over die tekst. Dit leidde tot de vraag: wat is het verband tussen het aantal fouten per honderd woorden in teksten en het holistische oordeel over die teksten? Bij holistische beoordeling wordt een tekst in beginsel relatief snel en impressionistisch beoordeeld worden, dat wil zeggen, afgaande op de algehele indruk die de tekst maakt. Het is een door docenten veelgebruikte manier om de kwaliteit van teksten te beoordelen en de rangorde van teksten vast te stellen (Breland, 1983; Charney, 1984; Cooper, 1984; Greenberg, 1992; Rossen-Knill & Lynch, 2000). De vraag die hieruit volgde, was in hoeverre de holistische beoordeling van teksten door studenten afwijkt van de holistische beoordeling van teksten door de onderzoekers. Om te weten of de vermindering van het aantal fouten per honderd woorden in een tekst de lezer positief beïnvloedt, was de derde vraag: in hoeverre leidt het corrigeren van de fouten in een tekst tot een positiever oordeel bij lezers van die tekst? Deze vraag werd met behulp van experimenteel onderzoek beantwoord.

In het correlationele deel van het onderzoek werd voor een aantal door studenten geschreven teksten nagegaan, wat het verband was tussen het aantal fouten per honderd woorden en het holistische oordeel over die teksten volgens expert-beoordelaars. Hierbij gebruikten de onderzoekers zichzelf als expert-beoordelaars en werden de teksten eerst holistisch beoordeeld en pas daarna op aantallen fouten (zie 7.3.1). Ondanks deze werkvolgorde bestond toch de mogelijkheid dat de onderzoekers zich onbewust bij hun holistische oordeel te veel zouden laten beïnvloeden door fouten in de teksten. Voor een harde toetsing van het verband tussen het aantal fouten per honderd woorden en het holistische oordeel was het wenselijk andere beoordelaars te gebruiken voor het holistische oordeel, zodat de beoordelaars als bron van het verband konden worden uitgesloten.

Daarom werden de door de studenten geschreven teksten in het tweede deel van het correlationele onderzoek voorgelegd aan een groep andere studenten met het verzoek de voorgelegde teksten te rangordenen van slecht naar goed (zie 7.3.2). Op deze wijze werd een twee-

de holistische oordeel verkregen, nu niet afkomstig van de onderzoekers, maar van studenten die zelf soortgelijke teksten geschreven hadden. Verder maakte het holistische oordeel van de studenten het mogelijk na te gaan of de ordening van de teksten door de studenten anders was dan de ordening volgens de 'expert-beoordelaars'.

In het derde deel van het onderzoek, het experimentele deel, is ten slotte nagegaan voor een drietal teksten die relatief veel fouten per honderd woorden bevatten of correctie van de fouten leidde tot een positiever oordeel bij lezers van die teksten (zie 7.3.3). De ene beoordelaar kan meer fouten signaleren dan de andere beoordelaar en mogelijk ook een eigen stijl van corrigeren hanteren en daarom zijn de drie originele teksten door beide onderzoekers onafhankelijk van elkaar gecorrigeerd. Op deze manier ontstonden van iedere tekst drie verschillende versies: de originele versie, de door onderzoeker A herschreven versie en de door onderzoeker B herschreven versie. De lezers, een nieuwe groep hbo-studenten en universitaire studenten, die gevraagd werden de teksten te lezen en te beoordelen kregen steeds slechts één tekst in één bepaalde versie te lezen, zodat ze niet konden focussen op verschillen tussen de teksten.

Ook werd bij de instructie aan de studenten niet de aandacht gevestigd op mogelijke fouten in de tekst. Dit was een belangrijk punt, want om te weten wat een tekst heeft overgedragen, kan de context waarin die tekst gegeven wordt, evenals de manier waarop naar het oordeel gevraagd wordt, bepalend zijn voor de uitkomsten. Uit het onderzoek van Tversky en Kahneman (1981) bleek dat het oordeel van mensen in sterke mate afhangt van de manier waarop een probleem geformuleerd wordt. Bij een identiek probleem dat op verschillende wijze geformuleerd werd, kon dat leiden tot tegenovergestelde uitkomsten. Vertaald naar het effect van taalfouten waarbij de reactie van de lezer op een tekst bepalend is, moet de aandacht niet gericht worden op de taalfouten in de teksten om beïnvloeding te voorkomen.

7.1.1 Geen invloed taalfouten

Voorafgaand aan het eigen onderzoek worden eerst praktijkvoorbeelden en onderzoeksuitkomsten besproken over het effect van taalfouten. In paragraaf 7.1.1 komt naar voren dat taalfouten geen of weinig effect hebben en in 7.1.2 dat dit juist wel het geval is. Opvattingen en voorbeelden op dit gebied zijn er overigens genoeg, maar empirisch onderzoek is veel schaarser. In 7.1.3. wordt ingegaan op het relatieve karakter van het begrip 'fout'.

'Ontdooi het Nederlands' was de titel van een opiniërend artikel in NRC Handelsblad over het taalniveau van studenten. "Als maar liefst 83 procent van de eerstejaarsstudenten aan de Nederlandse universiteiten voor een taaltoets zakt . . . moeten we ons dan niet in gemoede afvragen of er toevallig iets met de taal aan de hand is" (Van de Laar, 18 april 2009). Die ontdooide opvatting bleek in te houden dat het misschien niet zo belangrijk is hoe je 'onmiddellijk' schrijft en of je de betekenis weet van 'malafide'. Dit sloot aan bij de opvatting van Daniëls (Sanders, 2008) dat we beter goed kunnen rekenen, wat heel vaak fout gaat. We zouden dan van een hoop taalfouten af zijn. Deze opvatting komt ook terug in de Taaltrivia (mei 2009): 'De taal is van ons allemaal'. De strekking van deze voorbeelden is dat we ons niet druk moeten maken over taalfouten en dat er niets op tegen is als iedereen zijn 'eigen' taal gebruikt.

Spellingsregels worden beschouwd als 'ketenen van de schrijftaal' (Jansen, 2007), ook door studenten zelf. Zij hoeven geen voldoende te halen op een dictee, maar moeten een zakelijke tekst schrijven. Studenten zeggen: 'Het staat er toch' (Van Straalen, 2009). Dezelfde houding was merkbaar bij het eigen onderwijs aan de Hanzehogeschool Groningen. Een tweedejaarsstudente Commerciële Economie vond het onterecht, dat ze bij het schrijven van een sollicitatiebrief ook op taalfouten werd beoordeeld. Ze zag taalfouten niet als 'iets echts'.

Empirisch onderzoek op dit gebied is verricht door Kloet, Renkema en Van Wijk (2003). Zij hebben onderzoek gedaan naar het effect van taalfouten op de lezer. De gangbare opvatting is dat taalfouten een negatief effect hebben op de tekst zelf, het doel van de tekst en op het imago van de afzender, maar hiervoor is volgens hen weinig empirische onderbouwing. In deze lacune willen Kloet et al. voorzien.

Zij gingen uit van de theoretische opvattingen van Burgoon en Miller (1985) en Petty en Caccioppo (1986) over de invloed van taalfouten. De Language Expectancy Theory van Burgoon en Miller tracht een verklaring te geven voor de negatieve invloed van taalfouten op het imago van de afzender. De lezers hebben bepaalde verwachtingen over het taalgebruik die ze baseren op hun inschatting van de afzender. Bij een deskundige zender zouden taalfouten zwaarder aangerekend worden. Een schending van de verwachtingen van de lezer die veroorzaakt wordt door taalfouten, kan volgens deze theorie geprojecteerd worden op de zender zelf. Een alternatieve verklaring is hier overigens mogelijk, waarbij het niet om een schending van verwachtingen gaat, maar om de vraag of sprake is van nieuwe informatie. Als de lezer weet dat de afzender ongeletterd is, bevatten de taalfouten geen nieuwe informatie. Dat is wel het geval bij een deskundige afzender.

Volgens het Elaboration Likelihood model van Petty en Caccioppo gaat tekstverwerking via de centrale route of de perifere route. De lezer die de centrale route doorloopt, leest de tekst grondig en is gemotiveerd om voor zichzelf een kritische afweging te maken van de gebruikte argumenten in de tekst. Daarnaast of afzonderlijk van de centrale route kan de lezer de perifere route doorlopen, waarbij hij terugvalt op algemene zekerheden, bijvoorbeeld de deskundigheid van de bron en de hoeveelheid argumenten. Hoe gemotiveerd men de tekst leest ofwel hoe grondig of oppervlakkig, zou mogelijk invloed hebben op het effect van de fouten in de tekst. De zogenaamde presentatiefouten die vallen in de categorie 'lichte' taalfouten waarvan wordt aangenomen dat ze de inhoud van de tekst niet aantasten, zoals spelfouten en interpunctiefouten, zouden meer impact hebben op de perifere lezer. De 'zwaardere' markeerfouten, zoals een onjuist gebruik van connectieven, veranderen de inhoud wel en die zouden juist de centrale lezer beïnvloeden.

Deze theorieën gaan over hoe het negatieve effect van taalfouten tot stand komt, waarbij het opvallend is dat van die negatieve invloed werd uitgegaan. Een deel van de bovenstaande noties werd door Kloet et al. overgenomen. Zij onderzochten het effect van spelfouten en markeerfouten op de tekstwaardering, de overtuigingskracht van de tekst en het imago van de schrijver van de tekst. Bij de twee experimenten die zij deden, werden door de onderzoekers twee zelf geschreven brieven herschreven in vier versies van een foutloze versie tot een versie met het maximale aantal fouten, dat verschilde per experiment. Bij het eerste experiment werden vijf markeerfouten (voegwoorden, bijwoorden) en vijf spelfouten in de brieven gestopt en is nagegaan in hoeverre dit soort fouten apart of in combinatie de lezer beïnvloedden. In totaal 323 proefpersonen, hbo- en mbo-studenten, beoordeelden de brieven. Bij het tweede experiment werden maximaal drie spel- en drie markeerfouten in de briefteksten aangebracht. Deze proefpersonen in de leeftijd van 18 tot 45 jaar en ouder (in totaal 200) hadden een opleiding op vwo-niveau afgerond. Op basis van deze experimenten concludeerden Kloet et al. dat spelfouten geen, maar markeerfouten wel invloed hadden op de tekstwaardering, met name op de begrijpelijkheid van de tekst. Beide soorten fouten bleken volgens hen 'nauwelijks effect' te hebben op de overtuigingskracht en het imago van de afzender. De strekking van hun onderzoek was dat taalfouten een gering bereik hebben.

Bij dit onderzoek zijn een paar kanttekeningen te plaatsen. In de twee experimenten was sprake van een klein aantal, niet erg ingrijpende en soms vage taalfouten. De fouten waren door de onderzoekers aangebracht, wat het gevaar met zich meebrengt van gekunstelde fouten. De proefpersonen beoordeelden twee keer een brief, van beide brieven één versie,

waarbij dit dubbel gebruiken van proefpersonen het risico van een onnatuurlijke situatie in zich draagt. Het is niet onwaarschijnlijk dat de proefpersonen in het eerste experiment door hun opleidingsniveau een matige beheersing van het Nederlands hadden, wat van invloed kan zijn geweest op hun beoordelingsvermogen van taalfouten. Als zij bijvoorbeeld zelf slecht spellen, zien ze spelfouten in een tekst vermoedelijk niet en die kunnen dan ook geen effect hebben. Onderdeel van beide experimenten was dat de proefpersonen alle fouten in de tekst moesten aanstrepen, maar de resultaten daarvan zijn niet vermeld. De onderzoekers gaven alleen in het algemeen aan: "De taalfouten worden wel opgemerkt" (Kloet et al., 2003, p. 278). In het tweede experiment hadden de versies van een van beide brieven met alleen spelfouten en die met spel- en markeerfouten wel een negatief effect op het imago van de afzender. Bij de foutloze versies vond men de afzender deskundiger. De conclusie van de onderzoekers dat taalfouten weinig nadelige effecten hoeven te hebben voor het doel van de tekst of het imago van de afzender kan in het licht van het bovenstaande niet onderschreven worden. Hun eigen experimenten bevestigden de conclusie over het geringe bereik van taalfouten niet zonder meer.

7.1.2 Wel invloed taalfouten

Voorbeelden die suggereren dat taalfouten een verstrekkende invloed kunnen hebben, zijn er genoeg. Wie een sollicitatiebrief verstuurt met taalfouten, verhoogt in het algemeen niet zijn kansen om uitgenodigd te worden voor een gesprek, stelde Stroop (2007). Een ingezonden brief in een krant met spelfouten werd door de redactie van die krant minder serieus genomen (Spits, 2007). Uit gesprekken met ondernemers (persoonlijke communicatie) tijdens het Aiesec-congres (Groningen, 2009) bleek, dat zij negatief dachten over verslagen van studenten met taalfouten en slecht geformuleerde zinnen. Een van de sprekers van het congres die uitgebreid was ingegaan op cultuurverschillen en mensenrechten vond dat slecht geformuleerde teksten storen en dat als je de vorm niet beheerst dat iets zou zeggen over wat je verder kon. Slecht schriftelijk taalgebruik bleek zelfs tot geldboetes te kunnen leiden. Een Amerikaanse advocaat kreeg in 2004 een boete van ruim dertigduizend dollar voor zijn slechte schriftelijke taalgebruik. De rechter beschouwde de stukken van de advocaat met de vele taal- en spelfouten als minachting tegenover het hof (Sanders 2007).

Van Bezooijen (2003) onderzocht ergernissen over taalfouten aan de hand van een enquête met zeventien taalfouten die door haarzelf geselecteerd waren uit materiaal van een taalru-

briek en een eerdere enquête waarbij ruim honderd mensen gevraagd was om met taalergernissen te komen. Deze enquête met de zeventien grootste ergernissen in haar ogen werd ingevuld door 222 respondenten die konden aangeven welk verschijnsel het meest stoorde. Bovenaan de lijst kwam foutief gebruik van 'kennen' en 'kunnen' te staan. 'Hun' als onderwerp scoorde ook hoog, evenals verkeerd gebruik van 'dan' en 'als' na de stellende en vergrotende trap. De bedoeling van dit onderzoek was alleen om de grootste ergernissen in kaart te brengen.

Harm (2008) ging op basis hiervan een stap verder. Zij gebruikte teksten die ze aanbood in een variant met een bepaalde fout en één zonder die fout. Haar enquête naar het effect van taalfouten kreeg ze ingevuld terug van 110 mensen. Ze maakte zelf de fouten in de teksten, en bovendien manipuleerde ze de inhoud wat invloed kan hebben op de geloofwaardigheid van de teksten. Harm stelde op grond van haar enquête vast dat fouten met 'als', 'dan' en 'hun', een negatief effect hebben op de geloofwaardigheid van de tekst, de schrijver van de tekst en de bronnen die hij gebruikte. Deze vaststelling lijkt meer een opinie te zijn dan gebaseerd op harde waarnemingen.

Beason (2001) gebruikte een vragenlijst en een interview om te achterhalen hoe mensen in het bedrijfsleven dachten over fouten in geschreven teksten. Net als Harm bracht hij zelf fouten aan in een bestaande tekst. Hij vroeg aan 14 ondernemers uit de Verenigde Staten om die fouten te rangordenen. Daarna interviewde hij hen met als uitgangspunt de aangegeven volgorde van fouten en op basis van deze gesprekken stelde hij vast dat fouten verantwoordelijk zijn voor een verkeerde interpretatie van de tekst. Bovendien zouden fouten bevorderlijk zijn voor een negatieve uitstraling naar de schrijver van de tekst en mogelijk ook naar de organisatie waarvan de schrijver deel uitmaakt. Ook deze conclusies zijn niet gebaseerd op experimenteel onderzoek.

Paulson, Alexaner en Armstrong (2007) toonden aan dat lezers significant vaker en significant langer hun blik fixeerden op foute woorden dan op vergelijkbare niet-foute woorden (p. 319). Verder vonden ze dat hoewel de lezers wel vaker en langer keken naar de foute passages, ze de fouten niet of alleen zeer globaal vermeldden in hun commentaar op de tekst (p. 304).

Loerts (2012) gaf in haar onderzoek een samenvatting van EEG-onderzoek bij taalverwerkingstaken. Hieruit kwam naar voren dat alle inhoudswoorden een N400 genereerden (een

negatieve piekspanning die normaal optreedt 400 milliseconden (ms) na het begin van een semantische fout), maar bij woorden die semantisch niet in de zin pasten, zoals 'de bakker bakt de vogel', was de N400 groter. De LAN (left anterior negativity) die vernoemd is naar de plek op de schedel waar hij gemeten wordt, bleek vooral op te treden bij syntactische overtredingen, zoals onderwerp en werkwoord die niet overeenstemmen en bleek vooraf te gaan aan de P600. De P600 (een positieve spanningsgolf vooral aan de achterkant van de schedel) is voor een groot aantal syntactische fouten gevonden. Uit dit EEG-onderzoek blijkt dus dat proefpersonen bij het luisteren naar zinnen reageren op zowel semantische fouten als op vormfouten. Loerts constateerde aan de hand van haar eigen onderzoek dat Nederlandse moedertaalsprekers sterk reageerden op grammaticale fouten in Nederlandse zinnen, zoals verkeerd gevormde werkwoorden of verkeerde lidwoorden. De proefpersonen kregen zinnen te horen die correct of incorrect waren. De stroompjes op de schedels van de proefpersonen lieten een duidelijke piekspanning (P600) zien bij het horen van een foute zin. De hersenactiviteit veranderde spontaan door het luisteren naar niet-goedgevormde zinnen.

Onderzoek naar het effect van taalfouten in een reisfolder van Elegant Travel '*Mauritius vakantie paradijs'* (Ottjes, 2009) liet zien dat de folder zonder taalfouten positiever werd beoordeeld. In de bestaande Nederlandse tekst van circa 450 woorden stonden 42 taalfouten, dus bijna 1 taalfout per 10 woorden. De tekst maakte beperkt deel uit van deze folder, want de foto's overheersten. Toch hadden taalfouten invloed op de overtuigingskracht van de folder. Precies dezelfde folder met tekst en afbeeldingen is op twee manieren aan lezers voorgelegd die niet wisten dat taalfouten een rol speelden in dit onderzoek. Honderd mensen beoordeelden de folder met taalfouten en honderd mensen beoordeelden dezelfde folder, maar dan met de herschreven tekst zonder fouten. Mensen die de folder zonder taalfouten te zien kregen, oordeelden positiever over het boeken van de vakantiebestemming uit de folder. Ottjes heeft overigens af en toe wel iets meer aan de tekst verbeterd dan alleen maar de duidelijke fouten.

Op het punt van de invloed van fouten is een vergelijking te maken met experimenteel onderzoek dat is gedaan naar het effect van accent bij een presentatie op de toehoorders (Heijmer & Vonk, 2002). Deze toehoorders moesten een aantal mensen beoordelen via praatjes die dezelfde inhoud hadden, maar met een verschillend accent werden uitgesproken. Het kon daarbij gaan om Gronings, Twents, Limburgs of Haarlems. Dat laatste accent wordt doorgaans beschouwd als de norm voor algemeen beschaafd Nederlands. De spreker met

het Haarlemse accent werd door de toehoorders belangrijk positiever gewaardeerd dan de overige sprekers, hoewel de inhoud van de presentaties gelijk was. Bij de overige accenten werden zowel de presentaties als de sprekers negatiever beoordeeld. Dit onderzoek laat zien dat kleine afwijkingen van de sociale norm (een Haarlems accent) leiden tot een minder positieve waardering. Voor teksten die afwijken van de gangbare taalnorm zou iets soortgelijks kunnen gelden.

Deze voorbeelden laten zien dat (kleine) afwijkingen van de standaardtaal tot een negatiever oordeel kunnen leiden bij de lezer of toehoorder. Taalfouten bleken ook van invloed op de leessnelheid. Verder bleken taalfouten te leiden tot kenmerkende patronen in het EEG die door correcte zinnen niet werden opgeroepen.

7.1.3 Relativering foutbegrip

Betekent dit dat iedere taalfout even 'fout' is gezien vanuit de lezer? Dat is vermoedelijk niet het geval. Niet iedereen is bijvoorbeeld op de hoogte van de nieuwste spellingregels of vindt die noodzakelijk. Dat laatste bleek na publicatie van het meest recente Groene Boekje (2005), waarin veranderingen werden doorgevoerd op het gebied van tussenklanken, het gebruik van hoofdletters, kleine letters en leestekens. De schrijfwijze 'pannekoek' versus 'pannenkoek' zal in dat licht vermoedelijk minder beschadiging geven aan een tekst dan het geval is bij fouten in werkwoordsvormen. Dit maakt dat sommige taalfouten minder erg zijn dan andere en het relativeert de officiële spellingsvoorschriften. De lezer lijkt de tekst vanuit zijn eigen referentiekader te beoordelen en focust op wat hij ziet als taalfout. De Volkskrant, Trouw, NRC Handelsblad en HP/De Tijd boycotten de officiële spelling en volgden de zogenaamde witte spelling, vastgelegd in het Witte Boekje (2006). Empirisch gezien is er dan niet zoiets als een officiële standaardspelling, maar is het zaak om zo te spellen dat de minste commotie ontstaat bij de lezer. Opvallend hierbij is echter dat beide spellingboekjes niet gebaseerd zijn op enig empirisch onderzoek onder lezers.

Bestaand foutenonderzoek leert ook te relativeren wat mensen zien als fouten. Connors en Lunsford (1988) en Lunsford en Lunsford (2008) hebben omvangrijke foutenonderzoeken gedaan naar fouten in het werk van Amerikaanse eerstejaarsstudenten. Uit deze onderzoeken bleek dat men zich in de Verenigde Staten dezelfde zorgen maakte als in Nederland over het schrijfniveau van studenten. In hun onderzoek kreeg iedere beoordelaar een aantal

teksten te beoordelen, maar dezelfde tekst is niet door meerdere beoordelaars beoordeeld. Het oordeel van de beoordelaar stond bij Lunsford en Lunsford niet ter discussie en de vraag naar de betrouwbaarheid van de beoordelingen werd niet gesteld. Wall en Hull (1989) noemden dit de 'certainty factor' (p. 261). Mensen zouden ervan overtuigd zijn dat een fout een fout is: aan het bestaan en het verkeerd zijn van een fout kan niet getwijfeld worden. Toch signaleerden Connors en Lunsford (1988) dat beoordelaars op verschillende dingen bleken te letten. "The most prevalent 'error,' failure to place a comma after an introductory word or phrase, was a *bête noir* for some teachers but was ignored by many more" (Connors & Lunsford, 1988, p. 164). Het is daarom opvallend dat zij tegelijkertijd veronderstelden dat beoordelaars alle fouten zien en dat het niet nodig was om naar de overeenstemming tussen beoordelaars te kijken. Aan de ene kant werd het betrekkelijke van een taalfout ingezien: "how arbitrary and context-bound our judgments of formal error are" (Connors & Lunsford,1988, p.158), maar dat inzicht speelde verder geen rol in hun beoordelingsmethode.

Wall en Hull (1989) keken wel naar de overeenstemming tussen beoordelaars en constateerden dat beoordelaars het slechts over een klein aantal fouten eens waren. Een essay van vierhonderd woorden van een student werd voorgelegd aan 55 beoordelaars: 10 universitaire docenten, 25 docenten uit het voortgezet onderwijs en 20 docenten uit het basisonderwijs. De opdracht was om alle fouten op het gebied van interpunctie, spelling, syntaxis en grammatica eruit te halen. Het aantal gesignaleerde fouten per beoordelaar liep uiteen van 9 tot 56. Een minderheid van de beoordelaars (20 procent of minder) signaleerde 75 procent van alle fouten in de tekst. Dit sprak niet voor een gemeenschappelijk referentiekader. Dat kader ontbrak volgens Wall & Hull ook tussen docenten die fouten in teksten van studenten markeerden en de studenten die de teksten moesten verbeteren: "we can no longer take the existence of such a common language for granted" (Wall & Hull, 1989, p. 286).

Het onderzoek van Wall & Hull was vaag over de precieze analysemethode en de uitkomsten. De onderlinge overeenstemming tussen de beoordelaars of het gebrek daaraan werd niet systematisch onderzocht. Wall en Hull vermeldden het totale aantal foutsignaleringen van de beoordelaars (1800 signaleringen), maar niet tot hoeveel verschillende fouten dat uiteindelijk leidde. Zij presenteerden percentages, zonder dat dus duidelijk was op welk totaal aantal fouten die sloegen. In de tekst werden veel voorbeelden gegeven, maar een overzicht van de foutcategorieën met aantallen ontbrak. Hoewel het gemiddeld aantal gesignaleerde fouten per groep docenten werd gegeven, kwantificeerden ze de overeenstemming tussen de beoordelaars verder niet.

In totaal vonden zij 25 'high consensus errors' (Wall & Hull, 1989, p. 269). Dat waren fouten die door 41 procent of meer beoordelaars gesignaleerd werden. Over de 'medium consensus errors' (p. 270) werden wel veel voorbeelden gerapporteerd, maar geen aantallen. Ook de grenzen voor wat een medium consensus error precies was, ontbraken. Die begrenzing viel wel af te leiden uit de 'low consensus errors' (p. 272) die door 20 procent of minder van de beoordelaars gesignaleerd werden.

Wall en Hull wilden ook weten waarom beoordelaars bepaalde fouten ernstig vonden. Het antwoord kwam er in 75 procent van de gevallen op neer dat deze fouten een belemmering waren voor de effectieve communicatie van betekenis. Deze ernstige fouten bleken heel uiteenlopende fouten te zijn, variërend van fouten in interpunctie en spelling tot grammaticale fouten, stijlfouten en fouten die te maken hadden met helderheid van de redenering.

Universitaire docenten zouden het meest kritisch zijn en de docenten uit het basisonderwijs het minst kritisch. Verschillen tussen de drie groepen beoordelaars: 'college professors', 'English teachers in secondary school' en 'English teachers in elementary school' (Wall & Hull, 1989, p. 266) werden geconstateerd, maar een significantietoetsing werd niet uitgevoerd, terwijl op basis van de gerapporteerde gegevens een significante uitkomst weinig plausibel lijkt. De auteurs stelden echter: "we feel justified in describing these three groups as different interpretive communities. Global counts of the errors . . . suggest these differences" (Wall & Hull, 1989, p. 278). Dit onderscheid bleef daarna in de literatuur voortleven. Anson (2000) nam het zonder kwalificatie over: "elementary, secondary, and college teachers' patterns of labeling and identifying errors differed " (Anson, 2009, p. 9).

Deze bevindingen relativeren het begrip 'fout'. Een fout bleek, afgaande op het hier besproken onderzoek, niet een vaststaand iets te zijn. Verder bleek een fout tijdgebonden te zijn (Anson, 2000; Connors & Lunsford, 1988; Lunsford & Lunsford, 2008). Renkema (2005) noemde in de *Schrijfwijzer* zeven verschillende normen op basis waarvan een bepaald taalverschijnsel goed of fout kan worden gevonden. Hij gaf daarbij zelf aan dat het bij die normen vaak niet duidelijk is wanneer ze van toepassing zijn en dat ze soms zelfs leiden tot verschillende resultaten. Wie bijvoorbeeld bij de vergrotende trap 'als' gebruikt, overschrijdt de autoriteitsnorm, maar niet de historische norm. Zelf hanteerde Renkema vooral de effectnorm. "Het gaat erom dat de boodschap overkomt zoals die bedoeld is, in een formulering waardoor de lezer gestimuleerd wordt om kennis te nemen van de inhoud" (Renkema, 2005, p. 17).

Dit geeft aan dat het begrip 'fout' rekkelijk is. Hoe zwaar een taalfout telt, zal vermoedelijk ook te maken hebben met het doel van de tekst en de situatie waarin gecommuniceerd wordt. Als in een mailtje voor een hotelreservering afkomstig van buitenlandse gasten een paar opvallende taalfouten staan, zal de hotelhouder die wel door de vingers zien en graag een kamer willen verhuren. Heel anders wordt het als de hotelhouder zijn hotel probeert te slijten en gekozen kan worden uit tientallen andere hotels. In dat geval zal de lezer niet echt gemotiveerd zijn om zijn folder of website vol taalfouten te ontcijferen. Iets dergelijks bleek ook uit het hiervoor genoemde onderzoek van Ottjes (2009).

Door de uitkomsten van het Foutenonderzoek (deelstudie 1) wordt het begrip 'fout' eveneens gerelativeerd. Iedere beoordelaar bleek fouten te signaleren die niet door andere beoordelaars werden waargenomen. Verder werd duidelijk dat zelfs de beste beoordelaars nog een belangrijk percentage van alle bevestigde fouten in een tekst over het hoofd zien. Het Foutenonderzoek laat echter tevens zien dat er 'echte' fouten bestaan in de zin van taalfouten die door meerdere beoordelaars onafhankelijk van elkaar werden waargenomen (bevestigde fouten). Het aantal fouten per honderd woorden was vervolgens een betrouwbare graadmeter voor de kwaliteit van teksten.

Voorafgaand aan dit onderzoek naar het effect van taalfouten was de verwachting dat taalfouten uitmaken en dat de lezer zich bij een tekst met taalfouten bewust of onbewust focust op die fouten en zich daardoor in negatieve zin laat beïnvloeden. Taalfouten beschadigen de lading van de tekst. Deze verwachting was gebaseerd op de eigen reactie bij het beoordelen van schrijfproducten van hbo-studenten en op het in 7.1.2 vermelde onderzoek naar de invloed van taalfouten.

7.2 Methode

Bij het beantwoorden van de vraag in hoeverre fouten een tekst aantasten, zijn teksten gebruikt die door studenten geschreven zijn. Ter voorkoming van onnatuurlijke fouten is er niet voor gekozen om zelf fouten aan te brengen in bestaande, goed geformuleerde teksten.

Materiaal

Twee groepen eerstejaars hbo-studenten van de opleiding Commerciële Economie (CE) en de opleiding International Business and Languages (IBL) van de Hanzehogeschool Groningen kregen in maart 2013 een schrijfopdracht om een evaluatieve tekst te schrijven van maximaal één A4 over TAVAN (zie bijlage 12). Deze studenten hadden in het cursusjaar 2012-2013 meegedaan aan het TAVAN2-programma (de tweede keer dat TAVAN aangeboden werd aan eerstejaars hbo-studenten). Ze hadden dus informatie over over dit onderwerp en tot op zekere hoogte dezelfde voorkennis. Door deze opdracht werden twee problemen vermeden die vaak voorkomen bij schrijfopdrachten: bij onbekende onderwerpen moet informatie meegeleverd worden en bij veel onderwerpen kan tekst verzameld worden via Internet. In beide gevallen gaan studenten goed geformuleerde informatie overnemen en gebruiken. Bij deze TAVAN-opdracht was dit uitgesloten. Via internet was weinig te vinden over TAVAN en alleen de structuur werd gegeven met een aantal hoofdpunten in de vorm van vragen die behandeld moesten worden. De teksten werden tijdens een regulier werkcollege Bedrijfscommunicatie Nederlands gemaakt in een computerlokaal of in een gewoon lokaal met eigen laptops. Na afloop mailden de studenten het resultaat als Word-document naar de docent die tegelijk een van de onderzoekers was. Het was de bedoeling om op deze manier zestig teksten te verzamelen: dertig van de IBL-groep en dertig van de CE-groep.

Beoordeling

De teksten werden door beide onderzoekers onafhankelijk van elkaar holistisch beoordeeld. Dit oordeel kwam tot uitdrukking in een cijfer variërend van 1 tot 10. Vervolgens werden de teksten door beide onderzoekers onafhankelijk van elkaar beoordeeld op fouten door de gesignaleerde fouten te onderstrepen en het aantal fouten per tekst te tellen. Om volgorde-effecten te voorkomen, werd de stapel teksten per onderzoeker geschud en is vervolgens steeds blind een tekst uit de stapel geselecteerd. De onderzoekers waren bij hun holistische oordeel vrij in de keuze van de beoordelingsschaal en ze waren ook vrij in wat ze als fout wilden signaleren. Wel gold dat een hoog holistisch oordeel betekende dat de tekst goed werd gevonden, terwijl een laag oordeel overeenkwam met het tegenovergestelde. De lengte van de teksten is bepaald via de optie 'Woorden tellen' in OpenOffice Writer. Het aantal fouten per honderd woorden is vervolgens per tekst en per onderzoeker berekend met SPSS versie 20.

Bij het combineren van holistische oordelen door de cijfers op te tellen of te middelen, doet zich het probleem voor dat de spreiding (SD) tussen beoordelaars sterk kan verschillen. De ene beoordelaar kan veel hoge en veel lage cijfers geven, een andere beoordelaars kan vooral in het midden van de schaal gaan zitten en slechts beperkt van het gemiddelde afwijken. Wanneer vervolgens beide oordelen opgeteld of gemiddeld worden, zal het uiteindelijke oordeel vooral bepaald worden door de beoordelaar die de grootste spreiding (de grootste SD) had. Dit is normaal niet de bedoeling bij het combineren van de oordelen van twee beoordelaars en daarom kunnen de SD's van beide beoordelaars gestandaardiseerd worden (gelijk gemaakt worden) door te werken met bijvoorbeeld z-scores. Hetzelfde probleem doet zich voor bij aantallen fouten per honderd woorden. De ene beoordelaar kan een belangrijk andere SD hebben dan de andere. Bij het combineren (optellen) van beoordelingen en aantallen fouten is dan ook steeds gestandaardiseerd, tenzij anders aangegeven.

De teksten over TAVAN zijn als volgt door de studenten beoordeeld. De teksten werden in beginsel per groep in series van zes bij elkaar genomen en zesmaal geprint. Dit leverde per serie zes mappen op met per map steeds dezelfde zes teksten. Per map werd een beoordelingsformulier toegevoegd (zie bijlage 13). Iedere student kreeg bij een volgende les vervolgens één map met zes teksten uit de andere groep ter beoordeling. Door te werken met twee groepen werd voorkomen dat een student zijn eigen tekst moest beoordelen. De teksten werden gecodeerd met behulp van een lettercode. Het was daardoor bij de beoordeling van de teksten niet te zien van wie de teksten afkomstig waren. De studenten moesten deze lettercode ook gebruiken op het beoordelingsformulier. Het aantal beschikbare teksten per groep bij de beoordeling was nooit een veelvoud van zes, zodat teksten soms in meer dan één serie moesten worden opgenomen om voldoende series te krijgen. Ook moesten meer mappen voor beoordeling beschikbaar zijn, omdat bij de volgende les meer studenten aanwezig konden zijn dan bij de eerdere les.

De reden om met niet meer dan zes teksten te werken per student was dat de beoordelingstaak voor de studenten niet te omvangrijk en te moeilijk mocht worden. Het was de bedoeling dat ze het serieus zouden doen. Aan de andere kant mocht het aantal teksten dat beoordeeld werd ook niet te klein zijn, omdat anders per tekst te weinig beoordelingen werden verkregen. Een reden om met verschillende series teksten te werken, was te voorkomen dat studenten bij de beoordeling gingen samenwerken en elkaars oordeel zouden overnemen. De mappen met teksten werden zo uitgedeeld dat studenten met dezelfde map niet naast elkaar zaten. Het was daardoor niet mogelijk om samen te werken en elkaars beoordeling over te nemen.

Door met vaste series teksten te werken in plaats van met volledig wisselende series was het mogelijk om naar de overeenstemming tussen studenten te kijken. De indeling van teksten door studenten binnen elke groep is namelijk achteraf met elkaar vergeleken om te zien of er overeenstemming bestond over de rangordening. Het oordeel van een student over een tekst werd vergeleken met de gemiddelde rang die de tekst kreeg bij de overige vijf beoordelingen. Door deze procedure werd duidelijk welke tekst als slechtste en beste tekst naar voren kwam binnen de groep. Overkoepelend kon op grond daarvan vastgesteld worden wat de slechtste (6 punten) en beste teksten (36 punten) waren.

De totale procedure leverde per tekst drie uitkomsten: het gemiddelde holistische oordeel van de onderzoekers, het gemiddelde aantal fouten per honderd woorden en een gemiddeld studentenoordeel.

Door deze aanpak werden de teksten over TAVAN in eerste instantie beoordeeld door studenten van de andere groep die zelf ook de tekst over TAVAN geschreven hadden en in dezelfde onderwijssituatie zaten. Deze studenten hadden daarmee een schrijversperspectief en niet het perspectief van een normale lezer. Het oordeel van 'normale' lezers over taalfouten was echter ook nodig. Om dat te achterhalen, is een aantal teksten door een nieuwe groep studenten beoordeeld. Drie teksten van hbo-studenten werden geselecteerd die door beide onderzoekers als slecht bevonden waren op grond van het aantal fouten. Deze drie teksten zijn in twee verschillende versies herschreven, zodat er van iedere tekst drie versies beschikbaar waren: de originele tekst met fouten en twee foutloze versies van de beide onderzoekers (zie bijlage 14). Dit leverde negen condities op. De foutloze versies van de onderzoekers verschilden van elkaar, doordat de ene onderzoeker meer fouten in de teksten signaleerde dan de andere.

Op basis van de negen tekstversies zijn negen stapels gemaakt waarin 25 keer een tekst, met een instructie en een beoordelingsformulier voorkwamen. Op de teksten stond de oorspronkelijke code, maar elke tekst was bovendien voorzien van een datum die fungeerde als volgnummer. Vervolgens is één grote stapel gemaakt waarin de teksten op basis van het versienummer zo geordend zijn, dat studenten die naast elkaar zaten niet een versie van dezelfde tekst te beoordelen kregen.

In mei 2013 werden de teksten tijdens reguliere colleges voorgelegd aan studenten Toegepaste Taalkunde van de Rijksuniversiteit Groningen en aan studenten van de opleiding

Small Business and Retail Management (SBRM) van de Hanzehogeschool. Per student werd één tekst beoordeeld. Het was de bedoeling om dezelfde tekst te laten beoordelen door 25 studenten. In totaal 225 studenten werden op die manier bij het onderzoek betrokken. Het effect van iedere tekst werd gemeten door de studenten na lezing van de tekst op een beoordelingsformulier tien schalen te laten aankruisen, waarmee de tekstwaardering gemeten werd (zie bijlage 15).

De tien schalen waarop de studenten gevraagd werd een oordeel te geven, waren in deze volgorde: saai - leuk, onduidelijk - duidelijk, niet informatief - wel informatief, slordig - verzorgd, slecht geschreven - goed geschreven, vervelend - interessant, subjectief - objectief, zwak - sterk, ondeskundig - deskundig, ongeschikt voor publicatie - geschikt voor publicatie. Bij de laatste schaal werd gevraagd of de tekst geschikt was voor publicatie, omdat in de instructie aan de studenten was aangegeven dat de tekst bedoeld was voor een studentenblad (zie bijlage 16). Aan de hand van deze schalen beoordeelden de studenten niet de fouten in de tekst, maar hoe de tekst overkomt.

De studenten gaven antwoord door een kruisje te zetten op de afgebeelde schaal. Het aantal millimeters van het begin van de schaal (links), vormde de score. Vervolgens is het aantal millimeters gedeeld door de totale lengte van de schaal en vermenigvuldigd met 100. Hierdoor ontstond per schaal een score tussen 0 en 100.

Deze methode heeft voordelen boven een benadering waarbij de totale schaal wordt onderverdeeld in een aantal stukken, een werkwijze die oorspronkelijk vermoedelijk is ingevoerd om het coderen te vergemakkelijken. Allereerst is het werken met een niet-onderverdeelde schaal veel gevoeliger: de totale schaal bestaat in de praktijk nu uit ruwweg 100 meetpunten tegen anders bijvoorbeeld 7. Ten tweede is de schaal continu, terwijl de respondent anders beperkt wordt tot een reeks geordende categorieën waar hij uit moet kiezen, zodat het antwoord in beginsel intuïtiever gegeven kan worden. Ten derde is de schaal minder ambigu. Veel respondenten zullen immers bij een schaal met onderverdeling antwoorden door een kruisje te zetten in het midden van een bepaald vak. Sommige respondenten kunnen echter antwoorden door het kruis op de vakmarkering te plaatsen of op bijvoorbeeld een derde van een bepaald vak. In het eerste geval kan de onderzoeker het antwoord nog coderen als bijvoorbeeld 3.5 wanneer men zich realiseert dat het statistisch programma dit toelaat, maar in het tweede geval blijft onduidelijk of de respondent nu bedoelde dat hij iets onder de 4 wilde antwoorden of gewoon 4 koos. Het nadeel van het werken met een niet-

onderverdeelde schaal is dat ieder antwoord op een millimeter nauwkeurig bepaald moet worden, wat extra tijd kost.

7.3 Resultaten

In totaal werden van de twee groepen studenten 48 bruikbare teksten per mail retour ontvangen: 22 uit de groep CE-studenten en 26 uit de groep IBL-studenten. Het minimum aantal woorden dat een tekst telde, bedroeg 315, het maximum aantal woorden 647. Gemiddeld telden de teksten 439.7 woorden (SD=46.3 woorden). De gemiddelde woordlengte van de teksten, berekend als het aantal tekens gedeeld door het aantal woorden, varieerde van 5.3 tekens tot 6.3 tekens per woord. Gemiddeld bedroeg de woordlengte 5.8 tekens per woord (SD=0.2 tekens).

Doordat sommige teksten in meerdere series voorkwamen en doordat het aantal beoordelingen per serie varieerde, werden niet alle teksten even vaak beoordeeld. Het laagste aantal beoordelingen van een tekst was vier, het hoogste vijftien. Gemiddeld werden de teksten door 6.75 studenten beoordeeld (SD=2.32). In totaal werden 22 teksten door minder dan zes studenten beoordeeld en 26 teksten door ten minste zes studenten.

In totaal brachten in de beoordelingsfase 54 studenten hun oordeel uit, waarbij iedere student een serie van zes teksten beoordeelde (324 beoordelingen). In totaal werden elf verschillende series (A-K) teksten gebruikt. Serie A werd door zes studenten beoordeeld, serie B door drie, serie C door zes, serie D door drie, serie E door zes, serie F door zes, serie G door vijf, serie H door vijf, serie I door vijf, serie J door vijf en serie K door vier studenten.

7.3.1 Holistisch oordeel en aantal fouten volgens de onderzoekers

In de eerste plaats is gekeken naar de overeenstemming tussen de expert-beoordelaars (de beide onderzoekers) bij hun holistische oordeel over de teksten en het aantal fouten dat zij in de teksten signaleerden.

Overeenstemming tussen expert-beoordelaars

De beide onderzoekers stemden significant overeen in hun holistisch oordeel over de 48 teksten met een productmoment-correlatie van 0.65 (p=0.000, 2-zijdig). Na standaardisatie op dezelfde SD, om iedere onderzoeker evenveel invloed te geven, bedroeg de betrouwbaarheid van het gezamenlijke holistische oordeel (de gestandaardiseerde beoordelaarsalfa) 0.79. Voor de overeenstemming tussen onafhankelijk werkende holistische beoordelaars kan dit opgevat worden als een hoge waarde.

In verhouding tot de te bespreken overeenstemming tussen de studenten (zie 7.3.2) deden de beide onderzoekers het qua overeenstemming belangrijk beter. Om bij de holistische beoordeling dezelfde betrouwbaarheid als die van de beide onderzoekers te bereiken, bleken - uitgaande van de gemiddelde onderlinge correlatie - meer dan 12 studenten nodig te zijn. Het oordeel van een enkele onderzoeker woog in termen van betrouwbaarheid ongeveer even zwaar als het oordeel van ruim zes studenten.

Pas nadat dat het holistische oordeel was uitgebracht, hebben de beoordelaars het aantal fouten in de teksten gesignaleerd. De overeenstemming tussen de beide onafhankelijk werkende onderzoekers over het aantal fouten PHW (per honderd woorden) in de teksten resulteerde in een significante onderlinge correlatie van 0.78 (p=0.000, 2-zijdig). De betrouwbaarheid van het gezamenlijk bepaalde aantal fouten PHW bedroeg na standaardisatie per beoordelaar op dezelfde SD (de gestandaardiseerde beoordelaarsalfa) 0.88. Het aantal fouten PHW bleek daarmee betrouwbaarder te kunnen worden vastgesteld dan het holistische oordeel.

Verband holistisch oordeel en het aantal fouten per honderd woorden

Wat was het verband tussen het het holistische oordeel van beide onderzoekers over de 48 teksten en het later bepaalde aantal fouten per honderd woorden? De correlatie tussen deze twee variabelen bedroeg -0.74 (p=0.000, 2-zijdig) en was daarmee hoog negatief en significant. Teksten met veel fouten werden door de onderzoekers belangrijk negatiever beoordeeld dan teksten met weinig fouten.

Indien rekening gehouden wordt met de betrouwbaarheid waarmee beide variabelen gemeten werden en de gevonden correlatie hiervoor corrigeert (correctie voor onbetrouwbaarheid of attenuatie) wordt een voor onbetrouwbaarheid gecorrigeerde correlatie van -0.89 gevonden voor het verband tussen het aantal fouten PHW en het holistische oordeel van de onderzoekers.

7.3.2 Studenten als holistische beoordelaars

In de tweede plaats is gekeken naar de overeenstemming tussen de studenten bij hun holistische oordeel over de teksten en naar de betrouwbaarheid en validiteit van hun oordeel.

Betrouwbaarheid holistisch oordeel studenten

Waren de studenten het onderling eens over de rangordening van de teksten? Per groep studenten die een bepaalde serie van zes teksten had beoordeeld, is via SPSS de beoordelaarsalfa met de gemiddelde onderlinge (product-moment) correlatie berekend. De laagste gevonden gemiddelde onderlinge correlatie tussen de studenten als beoordelaars bedroeg per groep -.03, de hoogste bedroeg 0.62. Het gemiddelde van alle onderlinge correlaties bedroeg 0.22 (gewogen naar het aantal betrokken correlaties per groep).

Een gemiddelde onderlinge correlatie van 0.22 stemt overeen met een beoordelaarsbetrouwbaarheid van 0.36 voor twee beoordelaars, 0.63 voor zes beoordelaars en 0.77 voor twaalf beoordelaars (Spearman-Brown formule voor testverlenging). Studenten stemden dus onderling enigszins overeen over de vraag wat de beste teksten waren, maar voor een betrouwbaar oordeel was een groot aantal (onafhankelijk van elkaar werkende) studenten vereist.

Hoewel de betrouwbaarheid van holistische beoordelingen bekend laag is, lijkt een waarde van gemiddeld 0.22 nog weer lager te zijn dan de waarden die normaal tussen expert-beoordelaars worden gevonden. Coffman (1966) vermeldde een gemiddelde correlatie van 0.386 op grond van het onderzoek uit hetzelfde jaar van Godshalk, Swineford en Coffman, voor beoordelaars die teksten met hetzelfde topic beoordeelden. Die correlatie werd al laag gevonden. De studenten deden het dus nog slechter dan deze beoordelaars. Via de formule

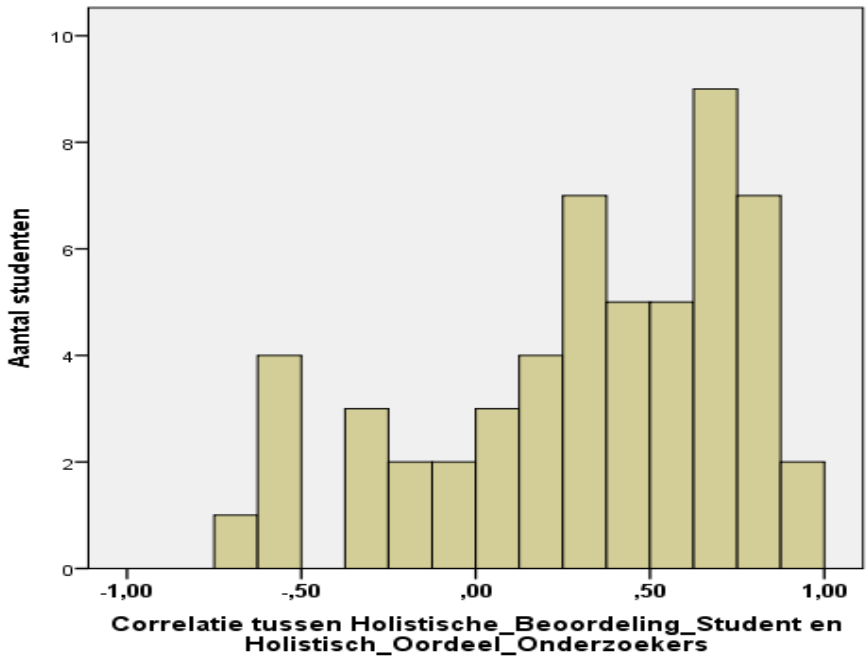
voor testverlenging valt af te leiden dat ongeveer twee studenten nodig zouden zijn om dezelfde correlatie als in het onderzoek van Godshalk et al. te bereiken. Waar zij vijf beoordelaars per tekst gebruikten, zouden er uitgaande van de studenten in dit onderzoek elf studentbeoordelaars nodig zijn om dezelfde beoordelaarsbetrouwbaarheid te bereiken.

Het kleine aantal per student beoordeelde teksten (zes) kan de lage waarde van 0.22 niet verklaren, omdat de hoogte van de correlatie (dit in tegenstelling tot de stabiliteit) niet beïnvloed wordt door de grootte van de steekproef. Kennelijk vormde het holistische beoordelen van de inhoudelijk sterk op elkaar lijkende teksten die soms weinig van elkaar leken te verschillen, voor de studenten een moeilijke taak.

Niet iedere student deed het als beoordelaar even goed. Door de zes tekstbeoordelingen te correleren met het relatief betrouwbare gemiddelde holistische oordeel van beide onderzoekers werd per student een index verkregen voor de kwaliteit van de beoordeling. Figuur 7.1 laat zien dat een relatief grote groep studenten rond de 0 scoorde: men deed het niet beter dan toeval. Verder was er nog een kleine groep die zelfs onder de -0.50 scoorde. Hoewel dit toeval zou kunnen zijn, suggereert de top op dat punt in het frequentiediagram anders. Het is mogelijk dat een aantal studenten de taak zo lastig en inspannend vond, dat men lukraak wat heeft ingevuld en dat enkelen zelfs opzettelijk of bij vergissing de rangorde hebben omgedraaid. Het resultaat is dan dat de gemiddelde overeenstemming relatief laag uitvalt. Dit doet overigens aan de juistheid van het gevonden gemiddelde voor de totale groep studenten niets af.

De uiteindelijke betrouwbaarheid van het studentenoordeel per tekst is voor de teksten die door ten minste zes studenten waren beoordeeld via een split-half methode bepaald. Dit is gedaan, omdat het aantal student-beoordelaars tussen de teksten fluctueerde, zodat de berekening van de beoordelaarsalfa (coëfficiënt alfa voor alle beoordelaars samen) niet mogelijk was. De zes of meer student-beoordelaars zijn op grond van hun plaats in de variabelenlijst (in de SPSS-datafile) verdeeld in twee groepen: de even en de oneven beoordelaars. Vervolgens is per groep beoordelaars de gemiddelde beoordeling van de teksten berekend. De correlatie tussen deze twee gemiddelde beoordelingen leverde de betrouwbaarheid voor de helft van de beoordelaars. Via de formule voor testverlenging kon daarna de betrouwbaarheid berekend worden van de volledige groep beoordelaars.

Figuur 7.1 Frequentiediagram van de correlaties tussen de beoordeling van de studenten en de beoordeling van de onderzoekers



Op deze manier werd een correlatie van 0.45 gevonden (p=.021, 2-zijdig, 26 teksten) tussen beide 'halve' groepen beoordelaars. De beide 'halve' studentbeoordelingen samengenomen, leverden vervolgens via de formule voor testverlenging de uiteindelijke betrouwbaarheid voor het holistische oordeel over de 26 teksten die door ten minste zes studenten waren beoordeeld. De op deze wijze gevonden betrouwbaarheid bedroeg 0.62.

Validiteit holistisch oordeel studenten

Stemde het oordeel van de studenten over de kwaliteit van de teksten overeen met het oordeel van beide onderzoekers? Om dit na te gaan is voor de 26 teksten die door ten minste zes studenten waren beoordeeld, de correlatie berekend tussen het gemiddelde oordeel van de studenten over de teksten en het gemiddelde oordeel van de onderzoekers. De correlatie bedroeg 0.69 (p=.000, 2-zijdig, N=26) en was daarmee significant en hoog.

De alfa-betrouwbaarheid van het holistisch oordeel van de onderzoekers was 0.79, zoals in 8.3.1 vermeld is. Voor de split-half betrouwbaarheid van het holistisch oordeel van de ten minste zes studenten werd eerder 0.62 gevonden. Uitgaande van deze betrouwbaarheden

bedroeg na correctie voor onbetrouwbaarheid de gecorrigeerde correlatie 0.99 voor 26 teksten. Dit resultaat betekent dat de studenten de teksten bij de holistische beoordeling op dezelfde criteria beoordeelden als de onderzoekers.

Wanneer de holistische oordelen van beide onderzoekers en de studenten werden gecombineerd (gemiddelde van 3 z-scores: onderzoeker A, onderzoeker B en de zes of meer studenten) voor de 26 teksten die door ruim zes studenten waren beoordeeld, ontstond een schaal met een alfabetrouwbaarheid van 0.83. Dit was de meest betrouwbare index voor het holistische oordeel in het onderzoek.

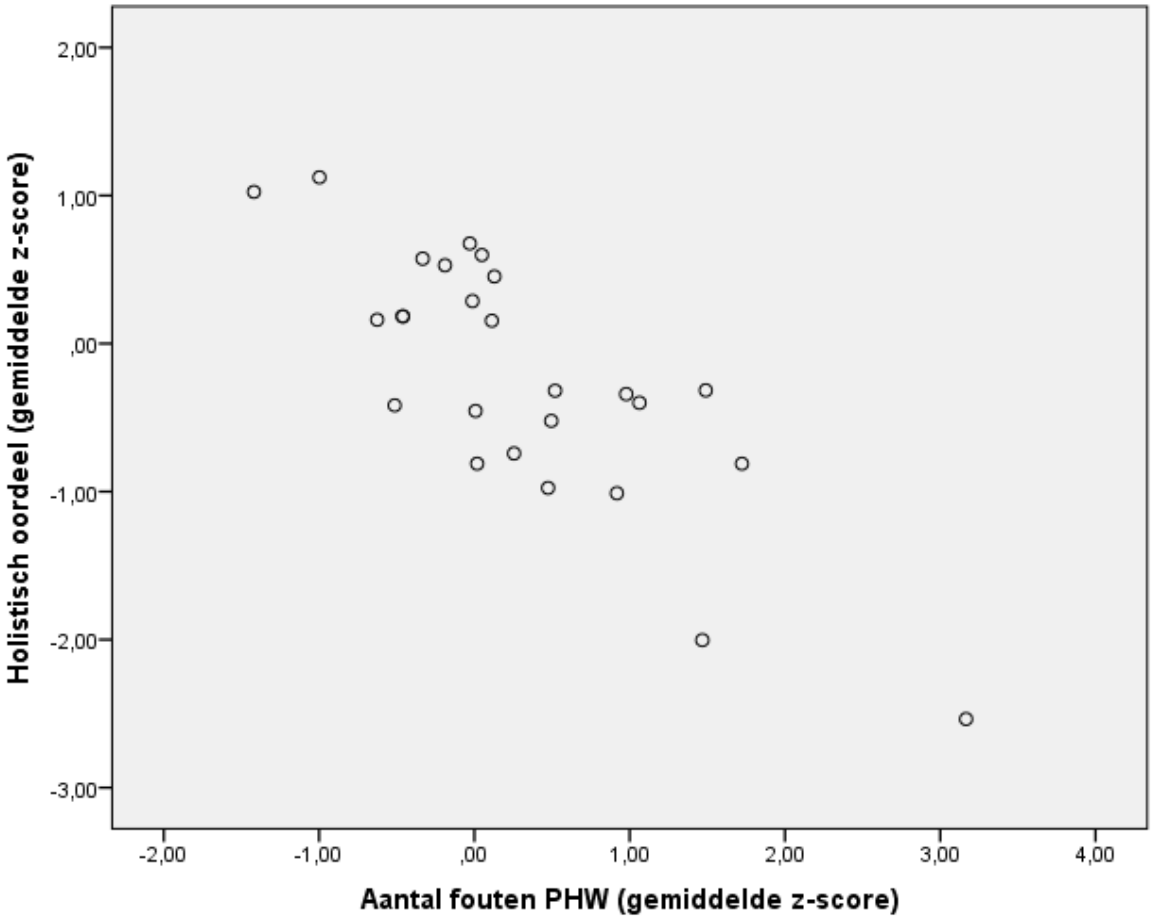
Uit de itemanalyse bleek via de gecorrigeerde item-totaalcorrelaties van 0.72, 0.65 en 0.69 voor respectievelijk onderzoeker A, onderzoeker B en de zes of meer studenten dat de studenten even goed in deze schaal pasten als de onderzoekers. Deze uitkomst laat zien dat het oordeel van de studenten over de teksten even waardevol (betrouwbaar en valide) was als het oordeel van beide onderzoekers, mits het aantal studenten voldoende groot was.

Verband aantal fouten PHW en holistisch oordeel studenten

Het aantal fouten per honderd woorden (gemiddelde z-score van beide onderzoekers) bleek significant negatief te correleren met het holistisch oordeel van de studenten over de teksten. Op basis van de 26 teksten die door ten minste zes studenten holistisch waren beoordeeld, bedroeg de correlatie -0.66 (p=.000).

Bij het interpreteren van deze uitkomst dient rekening gehouden te worden met de beperkte betrouwbaarheid van het holistische studentenoordeel. In beginsel kan een variabele nooit hoger met een andere variabele correleren dan de vierkantswortel uit zijn betrouwbaarheid. Na toepassing van de correctie voor onbetrouwbaarheid kwam de correlatie tussen het aantal fouten PHW en het holistisch oordeel van de studenten uit op -0.89. Dit is dezelfde waarde als de waarde die gevonden werd voor beide onderzoekers. Deze uitkomst laat zien dat er ook bij de studenten een zeer sterk verband bestond tussen het aantal fouten PHW en het holistisch oordeel. De voor onbetrouwbaarheid gecorrigeerde correlatie was dermate hoog dat beide maten kennelijk overwegend dezelfde factor maten.

Figuur 7.2 Verband tussen het gemiddelde aantal fouten PHW (gemiddelde z-score beide onderzoekers) en het holistische oordeel (gemiddelde z-score van onderzoeker A, onderzoeker B en zes of meer studenten) voor 26 teksten (r= -0.82, p=0.000)



In Figuur 7.2 is het scatterdiagram weergegeven voor het aantal fouten PHW in de 26 teksten (gemiddelde z-score van beide onderzoekers) en de meest betrouwbare index voor het holistische oordeel (gemiddelde van 3 z-scores: beide onderzoekers en de zes of meer studenten). De correlatie bedroeg -0.82 (p=0.000). Naarmate een tekst meer fouten bevatte per honderd woorden, werd de tekst als slechter beoordeeld.

Hierbij moet wel worden opgemerkt dat alle teksten betrekking hadden op hetzelfde onderwerp (de evaluatie van TAVAN), dat alle teksten ongeveer even lang waren en dat alle teksten dezelfde structuur volgden. Veel van de mogelijke factoren waardoor een holistische beoordelaar beïnvloed en afgeleid zou kunnen worden, waren in dit onderzoek constant en konden daardoor de holistische beoordeling niet beïnvloeden.

Het verband tussen schrijven en beoordelen

De studenten die de teksten schreven, beoordeelden ook teksten, zodat het mogelijk was na te gaan of de studenten die betere teksten schreven ook betere beoordelaars waren. Hiertoe werd per student eerst de correlatie berekend tussen zijn zes beoordelingen en de gemiddelde holistische beoordeling van beide onderzoekers als maat voor de kwaliteit van de beoordeling. Vervolgens werd de waarde van deze correlatie per student ingevoerd in de datafile. De correlatie tussen deze variabele, kwaliteit van de beoordeling, en de kwaliteit van de geschreven tekst (gemeten via het holistisch oordeel van beide onderzoekers) bleek 0.31 te bedragen (p=.041, 2-zijdig, 44 teksten). De studenten die beter schreven, bleken ook beter te beoordelen dan de slecht schrijvende studenten.

Om de hoogte van de correlatie goed te kunnen interpreteren, dient in de eerste plaats weer rekening gehouden te worden met de (on)betrouwbaarheid. De correlatie die als criterium werd gebruikt, was slechts gebaseerd op zes beoordeelde teksten. Voor een betrouwbare schatting zouden aanzienlijk meer teksten wenselijk zijn. Om een idee te krijgen van de betrouwbaarheid van de beoordeling is een split-half methode gebruikt met als maat de som van de afstanden. Het aantal waarnemingen per helft (drie) was te klein om in dit geval met correlaties te werken. De drie teksten die het hoogst in de datafile stonden, vormden de eerste helft, de drie teksten daaronder de tweede helft van de items. Wanneer bijvoorbeeld volgens de gemiddelde holistische beoordeling van beide onderzoekers de juiste rangorde van de zes teksten van boven naar beneden respectievelijk '3, 1, 6, 5, 4, 2' was en de beoordeling van de student was '3, 1, 5, 6, 2, 4', dan waren de respectievelijke afstanden '0, 0, 1, 1, 2, 2'. De som van de afstanden per helft was dan 1 en 5 en de totale afstand 6.

Deze procedure resulteerde in een split-half betrouwbaarheid van 0.42 waarbij de som van de afstanden het criterium vormde of de student goed of slecht beoordeeld had. Bij een slechte beoordeling was de totale afstand groot, bij een goede beoordeling klein. Na correctie voor onbetrouwbaarheid werd een gecorrigeerde correlatie van -0.50 gevonden tussen de holistische kwaliteit van de geschreven tekst en de kwaliteit van de beoordeling. Studenten waarvan de tekst hoog beoordeeld was, beoordeelden beter.

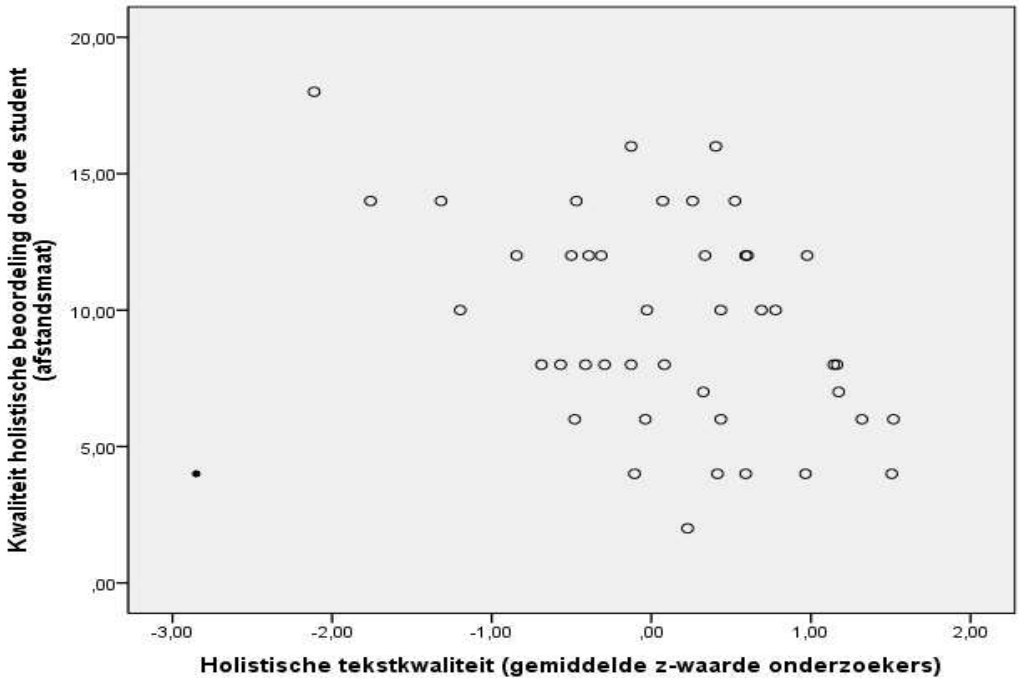
Een tweede reden waardoor de correlatie tussen de kwaliteit van de beoordeling en de kwaliteit van de geschreven tekst relatief laag kan zijn uitgevallen, was dat de beoordelingstaak per groep verschilde. Studenten die pech hadden, konden teksten krijgen die holistisch

soms slechts 0.01 van elkaar verschilden in z-waarde. Studenten die geluk hadden, kregen teksten die begonnen bij -2 of lager en dan keurig opliepen tot +2 in z-waarde met steeds duidelijke verschillen tussen de zes teksten. Deze ongelijkheid in de beoordelingstaak van de groepen maakte de score, in dit geval de afstandsmaat, minder betrouwbaar, maar zat niet verwerkt in de split-half betrouwbaarheid. Het kan dus zijn dat bij een betrouwbaardere en beter gestandaardiseerde meetprocedure een belangrijk hogere correlatie wordt gevonden tussen de kwaliteit van geproduceerde tekst en de mate waarin studenten teksten op kwaliteit kunnen beoordelen.

Een derde reden om te denken dat het verband tussen schrijven en beoordelen mogelijk vaak hoger zal liggen, lag besloten in het scatterdiagram (zie Figuur 7.3). Bij inspectie van het scatterdiagram bleek één punt ver van alle andere punten af te liggen: een zogenaamde 'outlier' of uitbijter. Dit punt betrof een student die extreem laag scoorde met zijn tekst (een z-waarde van 3.1 volgens de holistische beoordeling van de onderzoekers), maar vrijwel perfect scoorde met zijn beoordelingstaak. Wanneer deze ene student buiten beschouwing werd gelaten, steeg de voor onbetrouwbaarheid gecorrigeerde correlatie tussen schrijven en beoordelen naar -0.80.

Wanneer in plaats van het holistisch oordeel van beide onderzoekers het aantal fouten per honderd woorden werd gebruikt - op deze variabele is in 7.3.1 uitgebreider ingegaan -, dan was het verband tussen tekstkwaliteit en beoordelingskwaliteit nog iets hoger met een correlatie van 0.36 (p=0.018, 2-zijdig, 44 teksten). Na correctie voor onbetrouwbaarheid werd dit 0.59.

Figuur 7.3 Verband tussen hoe de studenten schrijven volgens holistische beoordeling onderzoekers en hoe studenten beoordelen. De outlier is zwart aangegeven.



7.3.3 Het effect van fouten op lezers

De studenten die de teksten beoordeelden (zie 7.3.2), hadden zelf soortgelijke teksten geschreven. De mogelijkheid bestaat dat zij daardoor anders naar een tekst gingen kijken. Verder was het voorgaande onderzoek correlationeel van aard. Er werd wel een duidelijk verband gevonden tussen het aantal fouten PHW en het holistische oordeel van zowel de studenten als de onderzoekers, maar strikt genomen werd daarmee nog niet een causaal verband aangetoond in de zin dat het verminderen van het aantal fouten PHW in een tekst tot gevolg heeft dat het oordeel van lezers over die tekst positiever wordt. Voor een dergelijke conclusie is een experimentele opzet vereist. De volgende vraag werd daarbij beantwoord: in hoeverre leidt het corrigeren van de fouten in een tekst tot een positiever oordeel bij lezers van die tekst?

Uit de 48 beschikbare en door studenten geschreven teksten over TAVAN werden drie teksten geselecteerd met veel fouten PHW, op basis van de aantallen fouten PHW die vastgesteld waren door een van beide onderzoekers. Deze teksten worden hierna aangeduid als A, B en C. Uitgaande van het gemiddelde aantal fouten PHW gebaseerd op beide onderzoe-

kers (zonder standaardisatie) bevatten de drie teksten A, B en C respectievelijk 13.4, 9.2 en 9.0 fouten PHW. De respectievelijke z-scores van deze aantallen fouten gebaseerd op beide onderzoekers waren 3.1, 1.3 en 1.2. In gemiddelde aantallen over beide onderzoekers ging het per tekst om respectievelijk 84.5, 84.0 en 79.5 fouten. De lengte van de teksten A, B en C bedroeg respectievelijk 315 woorden, 468 woorden en 434 woorden. De gemiddelde woordlengte bedroeg respectievelijk 4.5, 4.8 en 4.9 letters.

Beide onderzoekers werkten bij het nakijken op fouten onafhankelijk van elkaar en waren vrij in het bepalen van wat ze als 'fout' wilden signaleren. Per onderzoeker verschilden de aantallen gesignaleerde fouten. Onderzoeker 1 signaleerde in de teksten A, B en C respectievelijk 46, 50 en 39 fouten. Onderzoeker 2 signaleerde respectievelijk 123, 118 en 120 fouten. Beide onderzoekers herschreven de teksten A, B en C vervolgens onafhankelijk van elkaar, waarbij het doel was alleen de fouten te corrigeren. Bij het herschrijven van de teksten werden vervolgens nog weer enkele nieuwe fouten ontdekt, die ook gecorrigeerd werden. Op deze wijze ontstonden van de teksten A, B en C drie versies (zie bijlage 14): de originele met fouten (versie 0), de door onderzoeker 1 gecorrigeerde versie (versie 1) en de door onderzoeker 2 gecorrigeerde versie (versie 2). In beide versies 1 en 2 waren in beginsel alle fouten volgens de desbetreffende onderzoeker gecorrigeerd, maar in versie 2 waren belangrijk meer fouten gecorrigeerd dan in versie 1.

De vraag doet zich voor hoe het mogelijk is dat de ene onderzoeker belangrijk meer fouten signaleerde dan de andere. 'Fout' is geen absoluut begrip in dit verband, maar een relatief. De ene beoordelaar kan belangrijk meer of minder fouten signaleren dan de andere. Een fout kan enerzijds door een beoordelaar worden opgevat als een afwijking van het ideaal, dus als een mogelijk verbeterpunt. Anderzijds kan een beoordelaar een fout opvatten als een duidelijke overtreding van een taalnorm. In het laatste geval signaleert een beoordelaar belangrijk minder fouten dan in het eerste geval. Dit verschil in de absolute aantallen fouten die de beoordelaars signaleerden, betekende echter niet dat beoordelaars het oneens waren, in termen van correlatie, over de rangordening van de teksten op basis van het aantal fouten PHW. De overeenstemming tussen beide onderzoekers op dit punt was hoog ($r=0.78$, $p=0.000$, 2-zijdig), zoals reeds eerder werd vermeld (zie 7.3.1).

Verder hoeft het niet automatisch zo te zijn, dat een verbeterpunt dat een beoordelaar in een tekst meent te zien of dat zelfs door meerdere beoordelaars wordt gezien, ook automatisch zal leiden tot een hogere waardering bij de lezer. Gebruikelijk is immers dat docenten, be-

oordelaars en schrijvers hun ideeën over hoe een tekst behoort te zijn, zelden of nooit empirisch toetsen. De vraag is daarmee of fouten inderdaad uitmaken voor hoe de tekst bij de lezer overkomt of dat ze er eigenlijk niet echt toedoen.

Voor dit doel werden de 3x3 tekstversies (A0, A1, A2, B0, B1, B2, C0, C1, C2) voorgelegd aan elf verschillende groepen studenten (in totaal 188 studenten), met het verzoek de tekst te lezen en daarna het beoordelingsformulier (zie bijlage 15) in te vullen. Van de 188 studenten werden twee ingevulde formulieren niet meegenomen in de verwerking. De ene student had Nederlands niet als moedertaal en beheerste het Nederlands volgens eigen zeggen onvoldoende; de andere student had deelgenomen aan TAVAN2.

Om de tekstversies te randomiseren, is niet strikt gerandomiseerd, maar zijn de tekstversies vooraf systematisch geordend in de volgorde van A0 tot C2, zodat ze gelijkmatig verdeeld werden over de achtereenvolgende groepen. Iedere student kreeg steeds slechts één van de negen tekstversies te lezen en te beoordelen (zie bijlage 14).

Van de elf groepen waren acht groepen universitaire studenten en drie groepen hbo-studenten. In totaal deden 33 studenten van het hbo mee en 153 van de universiteit. Het kleinste aantal beoordeelde teksten per tekstversie was 20, het hoogste 23.

Via een itemanalyse van de tien schalen (de items) bleken twee schalen minder goed in de totale schaal te passen ('leuk' en 'objectief') door een in verhouding tot de andere schalen relatief lage gecorrigeerde item-totaalcorrelatie. De score op de overige acht schalen is samengenomen in een gemiddelde. De coëfficiënt alfa van deze acht schalen samengenomen bedroeg 0.91. De gemiddelde onderlinge correlatie tussen de items bedroeg .54. De totale schaal bleek daarmee zeer betrouwbaar.

Een 3x3 variantie-analyse met als onafhankelijke factoren Tekst (tekst A, tekst B, tekst C) en Versie (oorspronkelijke versie 0, herschreven versie 1, herschreven versie 2) leverde significante effecten op voor Tekst en Versie (beide p-waarden: 0.000) en een significant interactie-effect (p=.006). De Levene test op de homogeniteit van de foutenvarianties gaf aan dat de foutenvarianties niet verschilden (p=.386), zodat aan deze voorwaarde voor een variantie-analyse werd voldaan. De interactie bestond eruit dat tekst B door het herschrijven meer verbeterde dan tekst A en tekst C (zie het profieldiagram in figuur 3.4 met de geschatte randgemiddelden).

Tussen versie 0 en versie 1 en tussen versie 0 en versie 2 werden met een post hoc test met Bonferroni-correctie significante verschillen gevonden (p steeds .000), maar tussen versie 1 en versie 2 werden geen significante verschillen gevonden (p=.271). Het corrigeren van de fouten leidde voor beide herschreven versies tot een significant hogere waardering dan de originele versie, maar tussen de twee herschreven versies bestonden geen aantoonbare verschillen.

De percentages verklaarde variantie ('partial eta squared' x 100) van de onafhankelijke factoren waren 10.5% voor Tekst, 32,4% voor Versie en 7,8% voor de interactie tussen Tekst en Versie. Het percentage verklaarde variantie van Versie was daarmee meer dan driemaal zo groot als van Tekst. Omdat er tussen beide herschreven versies geen aantoonbare verschillen bestonden, betekende dit dat het herschrijven een zeer grote invloed had op de waardering van de lezers.

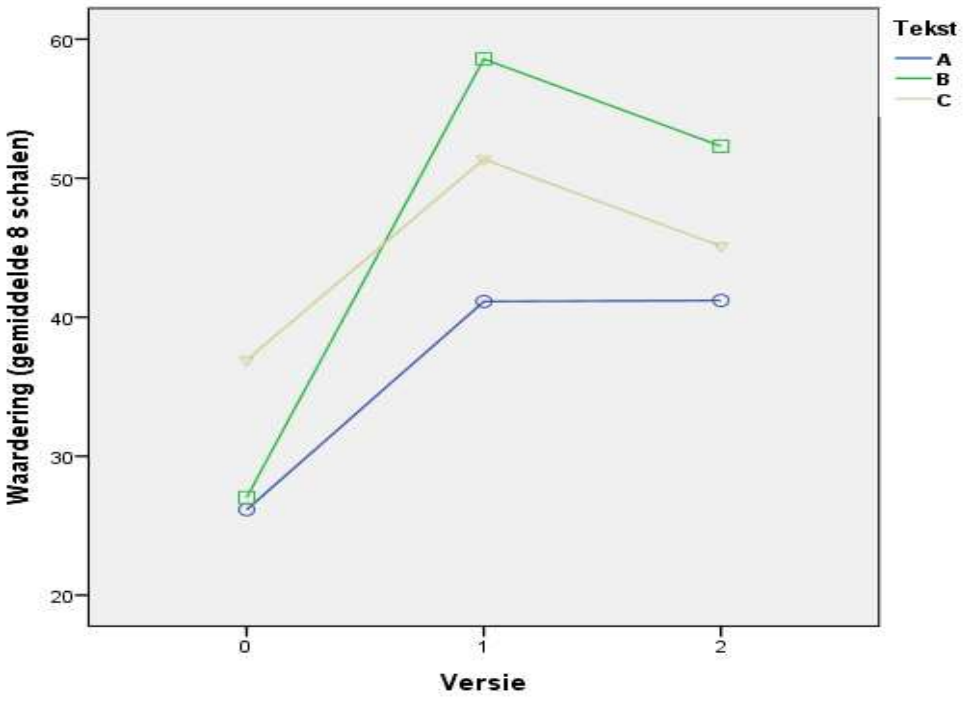
De waardering voor de originele versies met fouten was 29.9, de waardering voor de verbeterde versies was 48.3 of meer dan anderhalf maal zoveel. Dit verschil van 18.4 was belangrijk meer dan de SD van de controleconditie (versie 0): 13.6. Het effect van fouten op de waardering van de lezer was daarmee 1.35 SD, wat geldt als een zeer groot effect.

Voor tekst B werd door het corrigeren van de fouten zelfs meer dan een verdubbeling in de waardering bereikt. De waardering was 27.0 in de originele versie en deze werd in de herschreven condities gemiddeld 55.5. Waarom tekst B meer verbeterde door het verwijderen van de fouten dan de teksten A en C is niet duidelijk (zie Figuur 7.4).

Bij het nagaan per schaal van de verschillen tussen de originele teksten en de herschreven teksten bleken de schalen 'saai - leuk' en 'subjectief - objectief' geen verschillen te vertonen (t-toets, p-waarden respectievelijk 0.936 en 0.928, 2-zijdig, N=185). Alle andere acht schalen lieten significante verschillen zien in het voordeel van de herschreven versies. De hoogste p-waarde was 0.004 (2-zijdig).

De grootste verschillen in waardering werden gevonden op 'slordig - verzorgd' (30.3), op 'slecht geschreven - goed geschreven' (25.6) en op 'ongeschikt voor publicatie - geschikt voor publicatie' (25.3). Maar ook 'zwak - sterk' (18.0), 'ondeskundig - deskundig' (16.5) en 'onduidelijk - duidelijk' (15.2) lieten grote verschillen zien. De schalen 'niet informatief - wel informatief' (9.4) en 'vervelend - interessant' (9.0) lieten iets kleinere verschillen zien in het voordeel van de gecorrigeerde versies.

Figuur 7.4 De geschatte randgemiddelden per versie en per tekst. Tekst B liet een duidelijk interactie-effect zien: door het corrigeren van de fouten verbeterde de waardering meer dan bij de andere teksten. Tekst A scoorde in alle versies het laagst.



Tussen de universitaire studenten en de hbo-studenten werden geen significante verschil in waardering van de verschillende tekstversies gevonden, wanneer dit als extra factor in de variantie-analyse werd opgegeven (p=.261). Ook voor de factor 'groep' (de 11 respondenten-groepen) werd geen significant effect gevonden (p=.203).

7.4 Conclusies en discussie

In hoeverre hebben taalfouten in een tekst effect op de waardering van die tekst door de lezer? Deze vraag is beantwoord aan de hand van correlationeel en experimenteel onderzoek. Twee groepen hbo-studenten schreven in totaal 48 teksten. De beide onderzoekers die fungeerden als expert-beoordelaars beoordeelden de teksten onafhankelijk van elkaar, eerst holistisch en daarna op grond van het aantal fouten per honderd woorden. De hbo-studenten die de teksten schreven, werden zelf ook ingezet als holistische beoordelaars. De ene groep hbo-studenten beoordeelde de andere groep, waarbij iedere student een serie van zes teksten

beoordeelde, wat in totaal resulteerde in 324 beoordelingen van 54 studenten. Ten slotte zijn drie teksten in drie verschillende versies (de originele versie met fouten en twee verbeterde versies) beoordeeld door 'normale' lezers: een groep van 186 studenten (33 studenten van het hbo en 153 van de universiteit) die onbevangen naar de teksten konden kijken, omdat ze zelf geen teksten in dit verband geschreven hadden en niets van het onderwerp wisten. Iedere student beoordeelde aan de hand van een beoordelingsformulier met tien schalen één van de negen tekstversies. Iedere tekstversie werd minimaal 20 keer beoordeeld en maximaal 23 keer.

Een belangrijke uitkomst van dit onderzoek is dat tussen het aantal fouten per honderd woorden (PHW) in een tekst dat was vastgesteld via de beide onderzoekers en het holistische oordeel over een tekst, een zeer sterk verband bestaat bij zowel expert-beoordelaars (r=-0.74) als bij student-beoordelaars (r=-0.66). Deze uitkomst laat zien dat beoordelaars zich bij hun holistische oordeel (bewust of onbewust) sterk laten beïnvloeden door het aantal fouten in een tekst. Het verband is dermate sterk dat gesteld kan worden, dat het aantal fouten PHW en het holistische oordeel in de praktijk dezelfde factor meten. Met andere woorden: in plaats van holistisch te beoordelen kan men ook het aantal fouten PHW tellen om teksten te rangordenen. Een nadeel is dat dit meer tijd kost, een voordeel is dat het betrouwbaarder is.

Dit onderzoek laat ook zien dat studenten ingezet kunnen worden bij de holistische beoordeling van teksten, maar de mate van onderlinge overeenstemming bleek te verschillen van die van de onderzoekers. De gemiddelde correlatie tussen de oordelen van de studenten over dezelfde teksten was 0.22, die van de onderzoekers was 0.65. Voor een holistische beoordeling door studenten die even betrouwbaar was als de holistische beoordeling van één expert-beoordelaar, waren ruim zes studenten vereist.

Studenten bleken echter in beginsel even goed in staat te zijn de kwaliteit van teksten te beoordelen als expert-beoordelaars, mits het aantal student-beoordelaars voldoende groot was. Het holistische oordeel van de studenten stemde, na correctie voor onbetrouwbaarheid, vrijwel perfect overeen met het holistische oordeel van de expert-beoordelaars. Het oordeel van de studenten was met andere woorden minder betrouwbaar, wat inhield dat meer studenten-beoordelaars nodig waren, maar het was perfect valide. Dit resultaat betekent dat studenten teksten bij holistische beoordeling op dezelfde impliciete criteria beoordelen als 'expert-beoordelaars'. Ze hebben hetzelfde idee wanneer een tekst goed is, maar het is een vager idee.

De uitkomst dat teksten met veel fouten slechter beoordeeld werden door de student-beoordelaars dan teksten met weinig fouten, is opvallend. Studenten maken doorgaans veel fouten als ze schrijven en lijken het belang niet in te zien van een foutloze tekst. Toch speelde de hoeveelheid taalfouten een belangrijke rol in hun holistische oordeel over teksten van andere studenten.

Studenten bleken te verschillen in hun vermogen om teksten holistisch te beoordelen. Sommigen waren slecht, anderen waren uitgesproken goed. Doordat iedere student slechts zes teksten beoordeelde, was deze maat niet erg betrouwbaar (een split-half betrouwbaarheid van 0.62, 26 teksten). Er bleek een significant verband te bestaan tussen hoe goed studenten schreven en hoe goed ze holistisch beoordeelden ($r=0.31$, $p=0.041$, 2-zijdig, 44 teksten). Studenten die goed schreven (een hoge holistische beoordeling van hun tekst en weinig fouten PHW in die tekst), waren beter in het holistische beoordelen van teksten dan studenten die slecht of matig schreven.

De vraag of het corrigeren van de fouten in een tekst leidt tot een positiever oordeel bij lezers van die tekst is op grond van de uitkomsten van het experiment positief te beantwoorden. Het corrigeren van teksten met veel fouten, bleek te resulteren in een veel hogere waardering van die teksten door lezers (meer dan anderhalf maal zo hoog). De verwachting dat taalfouten uitmaken en dat de lezer zich in negatieve zin laat beïnvloeden door taalfouten in een tekst, werd bevestigd.

De ene serie teksten was ingrijpender herschreven dan de andere, maar dit leidde niet tot een hogere waardering bij de lezer. Kennelijk heeft het zin duidelijke en opvallende fouten te corrigeren, maar heeft het geen zin de tekst daarna nog verder te vervolmaken, althans niet wanneer de tekst normaal (vrij snel en vluchtig) gelezen wordt. Deze uitkomst suggereert dat er twee typen fouten bestaan. Sommige fouten maken uit voor het lezersoordeel en andere niet. Niet iedere fout heeft kennelijk dezelfde waarde.

De belangrijkste gevolgtrekking op basis van de onderzoeksuitkomsten is dat taalfouten in teksten uitmaken voor lezers. Teksten met veel fouten worden lager gewaardeerd door 'expert-beoordelaars', student-beoordelaars en normale lezers. Het corrigeren van de fouten in teksten bleek een zeer groot positief effect te hebben op de waardering van die teksten.

8

Deelstudie 5

Het meten van basale schrijfvaardigheid

Inleiding

In dit hoofdstuk en het volgende wordt dieper ingegaan op de problemen rond het meten van basale schrijfvaardigheid.[1] Met basale schrijfvaardigheid wordt in dit verband niet bedoeld het kunnen schrijven van een boek of een lang artikel, maar het zonder al te veel fouten kunnen schrijven van een tekst ter lengte van bijvoorbeeld een A4 (500 woorden).

Het uitgangspunt van dit proefschrift is dat wie goed kan meten, de desbetreffende vaardigheid ook kan maximaliseren, bijvoorbeeld door oefeningen aan te bieden. Toetsen en trainen moeten samengaan. De beste voorbereiding op een toets is vaak een andere toets. In onderwijssituaties is een meetinstrument pas echt zinvol op het moment dat het vertaald kan worden naar een trainingsprogramma. Het ontwikkelen van een effectief schrijfprogramma moet daarom beginnen met de vraag: hoe moeten we schrijfvaardigheid meten? Vervolgens is een doorslaggevende vraag of die meetmethode zich laat vertalen in een effectief trainingsprogramma.

In de onderwijspraktijk zijn veel docenten niet bij voorbaat overtuigd van het nut van meten van schrijfvaardigheid. Het kost veel tijd en levert weinig op, is hun overtuiging. Verder zijn metingen vaak niet betrouwbaar en niet valide, vindt men (Bonset & Braaksma, 2008, p. 133; Castagna, 2008; Deygers & Kanobana, 2010; Expertgroep Doorlopende Leerlijnen Taal en Rekenen, 2008, p. 13-14; Gilbert, 2004, p. 361; Inspectie voor het Onderwijs, 2010, p. 12 en p. 21; Peters, Van Houtven & El Morabit, 2010; Van der Westen, 2011a, p. 4). In de eerste drie paragrafen gaan we op deze kritiek in en komen tot de conclusie dat deze gedachtegang niet zonder grond is. Pas op het moment dat er een effectieve methode beschikbaar is om studenten met een tekortschietende schrijfvaardigheid te remediëren, wordt meten werkelijk zinvol.

In de eerste drie paragrafen van dit hoofdstuk speelt de geschiedenis van het meten van schrijfvaardigheid een belangrijke rol. Wat hier volgt, is bedoeld als achtergrondinformatie. De geschiedenis van het meten van schrijfvaardigheid laat men vaak beginnen in 1874 toen Harvard University de eis invoerde dat aspirant-studenten een korte tekst moesten inleveren in het kader van de toelatingsprocedure (Elliot, 2005). In 1900 werd in de VS de College Entrance Examination Board (College Board) opgericht om de toelatingsprocedure van de aangesloten universiteiten en colleges te standaardiseren via tests en toelatingsexamens. In

[1] In bijlage 1 worden een aantal gebruikte statistische en psychometrische begrippen toegelicht.

1921 toonde Hopkins aan dat het schriftelijke toelatingsexamen van de College Board dat uitging van essayantwoorden die holistisch beoordeeld werden, hinderlijk onbetrouwbaar was (Breland, 1983, p. 1). De oplossing van het onbetrouwbaarheidsprobleem kwam toen de College Board ten slotte overging op het gebruik van objectieve tests (meerkeuzevragen).

Deze overgang op objectieve tests leidde echter tot groot onbehagen bij docenten en bij de College Board aangesloten onderwijsinstellingen. Om te weten of een student kon schrijven, moest je hem laten schrijven, vond men. Meerkeuzevragen leken niet valide om schrijfvaardigheid te meten. In reactie op deze kritiek publiceerden Godshalk, Swineford en Coffman in opdracht van de College Board ten slotte het baanbrekende: *The Measurement of Writing Ability* (1966). In deze publicatie lieten de auteurs zien dat resultaten van objectieve tests om schrijfvaardigheid te meten, hoog correleerden met het holistische oordeel. Objectieve tests waren dus valide instrumenten om schrijfvaardigheid te meten, concludeerden de auteurs.

Het onderwijsveld zag dit echter belangrijk anders, omdat men gewend was te denken in termen van inhoudsvaliditeit. De inhoud van de test of toets was bepalend voor de validiteit. Vanuit dat criterium redenerend waren objectieve tests niet valide en dat ze wel in staat waren het holistische oordeel van de 'expert-beoordelaar' (de docent) goed te voorspellen, was dan vooral vreemd en fout.

Een ander bezwaar was dat aan studenten op deze manier niet langer goed viel uit te leggen dat ze moesten oefenen met schrijven, want dit werd in de toelatingsprocedure niet langer gevraagd. Ook de College Board erkende ten slotte de relevantie van dit argument en voerde in 2005 in de SAT (Scholistic Aptitude Test) opnieuw een essaydeel in om schrijfvaardigheid te meten. Daarnaast handhaafde men in verband met de betrouwbaarheid een objectief deel. Hoewel de objectieve tests betrouwbaar en valide waren gebleken, bleek de opsplitsing in twee qua inhoud volledig verschillende methodes waarvan de ene in het onderwijs werd gebruikt en de andere bij de toetsing, uiteindelijk niet langer volledig vol te houden.

Deze uitkomst lijkt onze stelling dat toetsen en trainen moeten samengaan, te bevestigen. In het geval van de SAT gaat het echter om een negatief gemotiveerde keuze: de discrepantie tussen holistische beoordeling en objectieve test leidde tot politieke druk en publieke ver-

ontwaardiging waardoor de testende instantie ten slotte water in de wijn deed, maar het probleem blijft in feite bestaan. Ons uitgangspunt daarentegen is dat het oefenen met toetsitems en verwante items de meest effectieve voorbereiding vormt op de toets.

De vraag die in paragraaf 8.4 aan de orde komt, is: waarom probeerde men het merkwaardige verband dat Godshalk et al. (1966) aantoonden tussen objectieve tests en holistische beoordeling, niet te verklaren? De meest aannemelijke verklaring is dat het verband voor het onderwijsveld zo gevoelig lag, dat de onderzoekers er liever voorzichtig omheen probeerden te manoeuvreren.

In paragraaf 8.5 wordt vervolgens ingegaan op de sterkte en de mogelijke verklaring van het verband. Het verband blijkt uitermate sterk te zijn: de objectieve tests blijken de holistische totaalscore voor een zeer groot deel (qua verklaarde variantie) te overlappen. De verklaring die ten slotte wordt voorgesteld op grond van de inhoud van de items is dat de objectieve tests vooral meten hoe goed studenten zijn in het opsporen en corrigeren van fouten. Wij noemen dit de FOC-factor (Fouten Opsporen en Corrigeren). Bij het schrijven van een tekst komt deze FOC-factor vervolgens tot uiting in het aantal fouten PHW. Vervolgens beïnvloedt het aantal fouten PHW de holistische beoordelaar waardoor het door Godshalk et al. (1966) aangetoonde verband ontstaat.

Het belang van deze paragraaf is niet alleen dat een toetsbaar verklaringsmechanisme voor het aangetoonde verband wordt geschetst (het aantal fouten PHW moet correleren met beide andere variabelen), maar deze verklaring maakt ook duidelijk wat deze drie variabelen die qua inhoud zo verschillend zijn, gemeenschappelijk hebben.

In paragraaf 8.6 wordt ingegaan op eerder onderzoek om het verband te verklaren. We merken hier op dat toen de sterkte van het verband ten slotte wel duidelijk werd gesteld door onderzoekers, de conclusie zo beladen was dat men die niet trok. In plaats daarvan spande men zich in de hoge gevonden waarde als niet belangrijk voor te stellen of zelfs weer terug te gaan naar de oorspronkelijke vraag. Een bepaalde angst voor 'foute' uitkomsten en een bepaalde gedrevenheid 'politieke correcte' conclusies te trekken lijkt het onderzoek in een andere richting geduwd en getrokken te hebben. Dezelfde factoren die eerder Godshalk et al. (1966) verhinderden al te diep in te gaan op het gevonden verband, bleven ook daarna het onderzoek sturen.

In paragraaf 8.7 wordt aangetoond dat het eerder in paragraaf 8.5 geschetste verklaringsmodel bevestigd wordt door waarnemingen. Dit wordt aangetoond door de gegevens uit het fouteneffect-onderzoek te combineren met de resultaten van TAVAN2 (de tweede keer dat het TAVAN-programma werd gegeven). De TAVAN-score wordt gebruikt als objectieve testscore. Het aantal fouten PHW in de teksten blijkt hoog gecorreleerd zijn met enerzijds de TAVAN-score en anderzijds het holistische oordeel over die teksten. In feite blijken alle drie variabelen in hoge mate dezelfde factor te meten (hoog gecorreleerd te zijn).

Dit resultaat toont tegelijkertijd aan dat de TAVAN-score en het aantal fouten PHW criteriumvalide zijn doordat ze hoog correleren met het holistische oordeel. Ten opzichte van de situatie na Godshalk et al. (1966) betekent dit dat het aantal manieren waarop basale schrijfvaardigheid gemeten kan worden, verdubbeld is. Voorheen bestonden er twee manieren (holistische oordeel en objectieve tests) en nu zijn daar nog twee aan toegevoegd (TAVAN-score en aantal fouten PHW). De TAVAN-score vormt wel een objectieve maat, maar wijkt af van de traditionele objectieve tests doordat de student zelf (her)schrijft.

In het fouteneffect-onderzoek (zie hoofdstuk 7) werd echter nog een andere variabele gehanteerd: de kwaliteit van het door de student uitgebrachte holistische oordeel. Naarmate het holistische oordeel dat de student had uitgebracht over zes teksten van medestudenten het 'ideale' holistische oordeel dichter benaderde, scoorde hij hoger op deze maat. Ook deze maat bleek na correctie voor onbetrouwbaarheid[1] hoog te correleren met de andere maten om basale schrijfvaardigheid te meten en daarmee criteriumvalide te zijn. Dit levert ten opzichte van de situatie na Godshalk et al. (1966) een derde nieuwe maat op om basale schrijfvaardigheid vast te stellen.

In paragraaf 8.7 komt ook naar voren dat de factor die alle vier gebruikte schrijfvaardigheidsmaten gemeenschappelijk hebben, betrekking heeft op het opsporen en corrigeren van fouten. Voor de objectieve tests was dit al aangetoond in 8.5. Basale schrijfvaardigheid blijkt daarmee in hoge mate samen te vallen met de FOC-factor.

In 8.8 wordt ingegaan op de vraag in hoeverre de verschillende meetmethodes bruikbaar zijn als trainings- en onderwijsmethode. Het holistische oordeel wordt al heel lang gebruikt in onderwijssituaties, maar lijkt didactisch gezien niet erg effectief. De TAVAN-score is uitgeprobeerd en bleek effectief te zijn. De kwaliteit van het uitgebrachte holistische oordeel is met enige aanpassingen vermoedelijk bruikbaar te maken als trainingsmethode.

[1] Zie bijlage 1 voor de formule van de correctie voor onbetrouwbaarheid.

Het praktische belang van deze paragraaf is dat er naast TAVAN in beginsel een tweede effectieve trainingsmethode mogelijk lijkt. Theoretisch van belang is dat die methode niet uitgaat van daadwerkelijk schrijven, maar van nauwkeurig lezen. Wanneer de methode effectief zou blijken te zijn, gaat de stelling dat men schrijven moet leren door veel te schrijven, niet meer volledig op. Het zou dan blijken dat goed lezen de basis legt voor schrijven.

In 8.9 wordt de vraag gesteld op welke manieren men schrijfvaardigheid in beginsel kan meten. In totaal worden tien verschillende methodes gepresenteerd en besproken. Een uitermate simpele methode die vaak toch betrouwbaar blijkt, is de lengte van de geschreven tekst. De bedoeling van deze paragraaf is te laten zien dat er in beginsel veel meer manieren zijn waarop basale schrijfvaardigheid vastgesteld kan worden dan men doorgaans aanneemt.

8.1 Is het meten van schrijfvaardigheid zinvol?

In dit proefschrift staan twee samenhangende problemen centraal: het meten en het optimaliseren van basale schrijfvaardigheid. Om te weten of onderwijs op het gebied van schrijfvaardigheid succesvol is, moet het schrijfvaardigheidsniveau eerst vastgesteld kunnen worden, dat wil zeggen: het moet gemeten kunnen worden. Zolang er niet goed gemeten kan worden, valt het effect van schrijfvaardigheidsonderwijs niet goed vast te stellen. Verder kunnen in dat geval ook geen eisen gesteld worden aan het niveau van individuele studenten en kan het niveau van groepen studenten niet worden bepaald. Het goed kunnen meten van schrijfvaardigheid heeft daarmee belangrijke gevolgen.

In de praktijk van het onderwijs zal echter niet iedereen onmiddellijk overtuigd zijn van de noodzaak dat schrijfvaardigheid gemeten moet worden. We zullen hierna laten zien dat dit standpunt vanuit de praktijk van het onderwijs nog steeds valt te begrijpen. Verder zullen we laten zien dat de problemen rond het meten van schrijfvaardigheid in het verleden zelfs zo groot waren, dat het niet meten of het niet al te objectief meten, waarschijnlijk goede opties waren.

In het onderwijs hebben we in de praktijk vooral te maken met studenten, docenten en onderwijsinstellingen. Een student zal normaal niet zitten wachten op aanvullende eisen. Wan-

neer er aan zijn schrijfvaardigheid geen eisen worden gesteld, zal hij dat niet zien als een probleem. Een docent schrijfvaardigheid doet zijn best, maar op het moment dat er eisen worden gesteld aan het niveau van zijn studenten, zal hij daar snel verantwoordelijk voor worden gehouden, terwijl zijn mogelijkheden dat niveau daadwerkelijk te beïnvloeden beperkt zijn. Voor onderwijsinstellingen geldt iets soortgelijks.

Alle betrokkenen hebben daarmee in beginsel een bepaald belang bij het niet meten of niet al te objectief meten van schrijfvaardigheid. Voordat de problemen rond het meten van schrijfvaardigheid in de onderwijspraktijk met succes aangepakt kunnen worden, zal eerst een bepaalde consensus moeten ontstaan dat schrijfvaardigheid een belangrijk probleem is dat aangepakt moet worden en dat meten bij de aanpak van dat probleem een essentieel deel van de oplossing vormt. Tekortschietende basale schrijfvaardigheid kan alleen bestreden worden door bij examens en toelatingsprocedures duidelijke eisen te stellen aan het minimaal benodigde niveau. Dat is echter pas mogelijk op het moment dat we dat niveau eenduidig, objectief en betrouwbaar kunnen vaststellen. Voordat we met succes eisen kunnen gaan stellen, moeten we eerst kunnen meten. Maar zolang de samenleving in doorsnee er niet van overtuigd is dat het meten van schrijfvaardigheid mogelijk is en zinvol is, zal dat niet gebeuren. Op dit punt is in zekere zin een cultuuromslag nodig.

We zullen hierna laten zien dat het kunnen meten van basale schrijfvaardigheid wel essentieel is, maar niet voldoende. Zelfs wanneer voor de hand liggende meetproblemen als betrouwbaarheid, validiteit en kosten (in arbeidstijd) zijn opgelost, heeft het meten van schrijfvaardigheid in het kader van een onderwijssysteem weinig zin, zolang een duidelijke methode met aangetoonde effectiviteit ontbreekt om een tekortschietend niveau gericht te verhogen. Meten wordt in onderwijskundige zin pas zinvol op het moment dat we studenten met een tekortschietende schrijfvaardigheid een duidelijk en effectief programma kunnen aanbieden om zich op dit punt te verbeteren.

Uitgaande van dit principe is het niet voldoende dat een maat voor basale schrijfvaardigheid betrouwbaar, valide en betaalbaar is, de maat moet ook bruikbaar zijn als trainingsmethode. Wie hardloopt, zal bij een wedstrijd een bepaalde afstand moeten lopen in een zo'n kort mogelijke tijd. Bij de training komt dezelfde opgave terug. Men loopt een bepaalde afstand in een bepaalde tijd. Meting en training vallen daardoor in hoge mate samen.

Hoewel dit principe misschien voor de hand liggend lijkt voor onderwijskundige maten, is het tot nu toe geen gebruikelijke eis, mogelijk doordat de ontwikkelaars van meetinstrumenten doorgaans een sterk psychometrisch perspectief hebben. Testontwikkelaars proberen in de praktijk bij voorkeur te voorkomen dat de moeizaam ontwikkelde items uitlekken en gebruikt worden als trainingsmateriaal. Verder is men bang dat training de zuiverheid van de meting zal verstoren, doordat sommige studenten wel zullen trainen en anderen niet. Ten slotte heeft men soms ook een sterke overtuiging dat training zinloos is en cognitieve vaardigheden vooral berusten op aangeboren eigenschappen. Er mag wel gemeten worden, maar er moet niet getraind worden, is dan het uitgangspunt. Vanuit psychometrisch standpunt is dit uitgangspunt misschien verdedigbaar, vanuit onderwijskundig standpunt leidt het tot de vreemde en onwenselijke situatie dat meten en trainen volledig worden losgekoppeld.

We gebruiken in dit verband de term 'training' in plaats van 'onderwijs', omdat men bij onderwijs geneigd is te denken aan een docent die zijn kennis probeert over te dragen aan de studenten (Beetsma, 2010, p. 43 en p. 46; Hofstee, 2008, p. 38; Keller, 1968; Lindall & Bolvin, 1967; Van Es, 1985, p. 24; Van der Wagen-Huijskes, 2011, p. 3; Vargas, 2009, p. 4.) Men veronderstelt een cognitief model. De essentie van een vaardigheid zou overdraagbare kennis zijn die de docent al heeft en de student nog niet. Door de student te vertellen wat hij moet weten, zou hij daarna de vaardigheid ook beheersen. De student hoeft als het ware alleen de instructies te volgen die de docent hem geeft. Uitgaande van dit cognitieve model is het niet nodig na te gaan of de student kan schrijven, maar of hij weet hoe hij volgens de docent moet schrijven. Uitgaande van dit cognitieve model is het ook niet nodig daadwerkelijk te oefenen, want het is vooral een kwestie van goed luisteren en weten.

Het antwoord op de vraag of het meten van basale schrijfvaardigheid zinvol is, kan daarmee als volgt luiden. Het meten van basale schrijfvaardigheid is zinvol, wanneer de meetmethode in beginsel ook gebruikt kan worden als trainingsmethode. De toets moet ook bruikbaar zijn als trainingsmiddel.

8.2 Eisen aan het meten van schrijfvaardigheid

Welke argumenten kunnen voorstanders van de 'meten is niet nodig-benadering' aanvoeren? Er zijn verschillende argumenten mogelijk, maar enkele belangrijke in dit verband zijn:

1. de meting is onbetrouwbaar;
2. de meting is niet valide;
3. de meting is duur, arbeidsintensief, tijdrovend;
4. de meting levert weinig bruikbare informatie op.

Een goede meting is dus in ieder geval betrouwbaar, valide, niet-arbeidsintensief en levert bruikbare informatie op. Hierna bespreken we kort deze eisen.

De betrouwbaarheid van de schrijfvaardigheidsmeting vormt bij holistische beoordeling vaak een probleem. Docent X laat zijn studenten een essay schrijven en beoordeelt dit. Een week later geeft docent Y een schrijfopdracht aan dezelfde groep studenten. In statistische zin correleren de uitkomsten enigszins, maar het verband is dermate zwak dat sommige studenten bij de ene docent een goed cijfer kunnen halen en bij de andere docent een slecht. Kennelijk speelde de schrijfopdracht en de docent die het werk beoordeelde een grote rol in plaats van hoe goed de student kon schrijven. De betrouwbaarheid, de overeenstemming tussen de uitkomsten van beide docenten, is voor gebruik in de onderwijspraktijk hinderlijk laag, waardoor de holistische beoordeling in de praktijk niet goed bruikbaar is.

Godshalk, Swineford en Coffman (1966) toonden aan dat het betrouwbaarheidsprobleem van de holistische beoordeling op te lossen is door per student veel teksten (schrijfopdrachten of 'topics') te gebruiken en per topic veel beoordelaars. Ze gebruikten vijf topics per student en vijf 'expert-beoordelaars' per topic. Dit resulteerde in een betrouwbaarheid van 0.841. Dit betekent dat wanneer de betrokken studenten vijf andere topics als schrijfopdracht hadden gekregen, waarna deze nieuwe teksten vervolgens waren beoordeeld door vijfentwintig andere expert-beoordelaars (vijf per topic), beide totaalscores naar verwachting 0.841 zouden correleren. Voor onderzoeksdoeleinden is deze methode van vijf topics en vijf beoordelaars soms bruikbaar; in de onderwijspraktijk niet, omdat de methode te duur en te omslachtig is.

Om het betrouwbaarheidsprobleem te ondervangen ging men na de Tweede Wereldoorlog in de Verenigde Staten over op het gebruik van objectieve toetsen (bestaande uit alleen meerkeuzevragen) om schrijfvaardigheid te meten. (Dit punt komt uitgebreider aan de orde in paragraaf 8.3.) Deze objectieve toetsen bleken inderdaad betrouwbaar te meten. Hierna rees echter de vraag of de score op zo'n objectieve toets nog wel iets kon zeggen over de daadwerkelijke schrijfvaardigheid van de student. Het betrouwbaarheidsprobleem was opgelost door over te gaan op objectieve tests, maar nu was er een probleem met de validiteit. Maten deze objectieve tests met alleen meerkeuzevragen nog wel schrijfvaardigheid?

Validiteitsproblemen kunnen op drie verschillende manieren opgelost worden (Nunnally, 1967, p. 75-102). De eenvoudigste manier is te kijken naar de wijze waarop de score tot stand komt. Om de validiteit van een studietoets te beoordelen, zal men zich een oordeel moeten vormen over de adequaatheid van de vragen. Wanneer de vragen geen betrekking hebben op de in het onderwijs behandelde stof of wanneer de vragen onnodig vaag zijn, kan men tot de conclusie komen dat de toets niet geschikt is, niet valide is. De inhoud van de toets, de vragen, vormt de basis voor het oordeel over de validiteit, daarom wordt dit 'inhoudsvaliditeit' genoemd. Bij een studietoets is het begrijpelijk dat studenten die laag op de toets scoorden vaak een ander oordeel over de toetsvragen zullen hebben dan de docent die de vragen geconstrueerd heeft. Oordelen over de inhoudsvaliditeit hebben daardoor gemakkelijk een opiniërend karakter in plaats van een empirisch karakter. De ene partij vindt de methode goed; de andere partij vindt de methode slecht. Voor schrijfvaardigheid vonden docenten het holistische oordeel de juiste manier om schrijfvaardigheid vast te stellen. Het holistische oordeel was volgens docenten (inhouds)valide, terwijl objectieve tests dat duidelijk niet waren, omdat ze niet gebaseerd waren op schrijfopdrachten, maar op het beantwoorden van meerkeuzevragen.

Om aan te tonen dat de objectieve tests om schrijfvaardigheid te meten wel valide waren, toonden Godshalk et al. in 1966 aan dat objectieve tests hoog correleerden met het criterium: de som van de holistische beoordelingen van vijf essays door steeds vijf beoordelaars per essay. De criteriumvaliditeit van de objectieve tests kon daarmee empirisch aangetoond worden en bleek zeer goed te zijn.

Door aan te tonen dat de objectieve tests schrijfvaardigheid valide maten, werd echter omgekeerd ook aangetoond dat het holistische oordeel valide was. Het holistische oordeel bleek immers hoog te correleren met de score op totaal andere maten die ook bedoeld waren schrijfvaardigheid te meten. Na de publicatie van Godshalk et al. (1966) waren er daarmee in beginsel twee totaal verschillende maten die beide schrijfvaardigheid betrouwbaar en valide konden meten. Voor een derde methode die schrijfvaardigheid beoogt te meten, betekent dit dat deze derde methode bij voorkeur met beide al bestaande meetmethodes moet correleren. Het criterium bestaat niet langer uit alleen het holistische oordeel, maar omvat nu in beginsel ook de objectieve tests. Het criterium is complexer geworden. Doordat schrijfvaardigheid op twee totaal verschillende manieren bepaald kan worden, is het een 'construct' geworden dat niet langer afhankelijk is van de meting van één specifieke varia-

bele. Wanneer een derde variabele gevonden is die schrijfvaardigheid meet, is het daarna wellicht mogelijk een vierde of zelfs vijfde variabele te vinden. Wat begon als een wat vaag begrip waarvan men niet helemaal zeker was hoe het precies gemeten moest worden, is het ondertussen een construct geworden dat niet alleen betrouwbaar, maar op meerdere manieren gemeten kan worden. Deze vorm van validiteit wordt daarom aangeduid als 'construct-validiteit'.

Een belangrijke volgende stap in de ontwikkeling van de validiteit van een maat als basale schrijfvaardigheid is of het construct door een docent of een onderzoeker verhoogd kan worden via een duidelijke en vastliggende procedure. Het construct moet trainbaar zijn. Zolang schrijfvaardigheid alleen gemeten kan worden, zelfs al zou dat op meerdere totaal verschillende manieren kunnen, is de waarde van het construct beperkt. Belangrijk is dat schrijfvaardigheid gericht vergroot kan worden. In deelstudie 2 (hoofdstuk 5) naar de geschiktheid van bestaande onderwijsmethodes om de schrijfvaardigheid van eerstejaars hbo-studenten te verhogen, bleek dat er voor dit doel veel methodes beschikbaar waren, maar dat geen enkele methode een empirisch aangetoonde effectiviteit bezat. Kennelijk is deze stap bij schrijfvaardigheid uiterst lastig. Het nieuw geconstrueerde TAVAN-programma (zie hoofdstuk 6) bleek echter zeer effectief, terwijl de te volgen procedure vrijwel volledig geautomatiseerd is.

De redenen waarom het kunnen verhogen (het maximaliseren) van de schrijfvaardigheid van belang is, zijn allereerst praktisch. Wanneer we wel de schrijfvaardigheid van een student kunnen meten, maar vervolgens alleen kunnen constateren dat die onvoldoende is, zonder daar iets aan te kunnen doen, is dat een onbevredigende situatie. De thermometer doet het wel, maar de verwarming niet.

Een tweede reden waarom het kunnen verhogen van de schrijfvaardigheid van belang is, is dat dit onderscheid mogelijk maakt tussen maten die wel en niet reageren op het verhogen van de schrijfvaardigheid. Zo bleek bij het effectonderzoek van TAVAN (deelstudie 3, hoofdstuk 6) de schrijfvaardigheid gemeten via het aantal fouten per honderd woorden verbeterd te zijn, maar de inschatting van de eigen schrijfvaardigheid (gemiddeld genomen) niet. Dankzij de effectiviteit van het TAVAN-programma was het mogelijk te laten zien dat de inschatting van de eigen schrijfvaardigheid voor het meten van basale schrijfvaardigheid soms geen valide maat is.

Het trainbaar zijn van een maat als basale schrijfvaardigheid is een volgende stap in de ontwikkeling van de validiteit, stelden we. Deze vorm van validiteit sluit aan bij het zinvolheidscriterium uit 8.1 dat stelde dat de toets in beginsel ook trainingsmiddel moet kunnen zijn. We komen op dit punt terug bij de vierde eis: de maat moet bruikbare informatie opleveren.

De derde eis die we stellen aan schrijfvaardigheidsmaten heeft betrekking op de kosten van de meting: in de praktijk vooral docenttijd en tijd benodigd voor de constructie van de maat. Holistische beoordeling heeft als nadeel dat de beoordeling relatief veel tijd kost bij grote aantallen studenten. Objectieve maten (meerkeuzevragen) hebben als nadeel dat het ontwikkelen tijd en deskundigheid vergt.

De vierde eis heeft betrekking op de informatie die de schrijfvaardigheidsmeting oplevert: wat doe men er precies mee? Wanneer we alleen meten om het meten, is het een kostbare aangelegenheid, die weinig zin heeft. De meting moet resulteren in een verantwoorde en zinvolle beslissing. Hoewel we dit punt als laatste eis aan de orde stellen, behoort dit eigenlijk eerst te komen. We moeten ons vooraf afvragen wat het doel van de meting precies is.

In grote lijnen lijken er dan twee verschillende doelen mogelijk. In de situatie van Godshalk et al. (1966) ging het ogenschijnlijk vooral om de selectie van de beste studenten. Voor toelatingsexamens en afsluitende examens gaat het er om de schrijfvaardigheid van de betrokken studenten zo betrouwbaar en valide mogelijk vast te stellen, tegen kosten die zo laag mogelijk zijn. In die specifieke situatie blijken objectieve tests goed te werken, waardoor de keuze voor objectieve tests in eerste instantie begrijpelijk lijkt.

In de situatie van docenten die schrijfvaardigheidsonderwijs verzorgen, gaat het er primair om studenten te laten schrijven. Het doel is niet in de eerste plaats vast te stellen wat het precieze niveau is, maar het doel is in de eerste plaats te zorgen dat de student een hoger schrijfvaardigheidsniveau bereikt. Om dat te bereiken is het nodig de student te laten schrijven. In deze situatie is het in beginsel niet nodig dat de meting perfect betrouwbaar is, maar is er wel een bepaalde check nodig op het werk van de student. In de praktijk is dat vaak holistische beoordeling.

Tussen deze twee uitersten bevindt zich een soort tussensituatie. De docent wil schrijfonderwijs geven en wil zijn studenten beter leren schrijven, maar moet op een bepaald mo-

ment ook een cijferlijst inleveren. De docent moet het schrijfvaardigheidsniveau van zijn studenten op een verantwoorde manier beoordelen. Het is niet voldoende de kamer te kunnen verwarmen, we willen ook de temperatuur weten.

Ook een selectiesituatie zoals een toelatingsexamen of een afsluitend examen maakt echter deel uit van een groter onderwijssysteem. Het doel is niet alleen te meten, maar het doel is ook te zorgen dat studenten een bepaald niveau bereiken. In eerste instantie wil men een meetmethode, maar vervolgens heeft dat weinig zin zonder effectieve onderwijsmethode. Ook nu willen we niet alleen de temperatuur in de kamer meten, maar ook kunnen verwarmen. De maat moet dus trainbaar zijn en een eenvoudige oplossing is de maat zelf daarvoor te gebruiken. Het criterium dat de toetsingsmethode bij voorkeur ook bruikbaar moet zijn als trainingsmethode (zie 8.1), komt daarmee terug.

8.3 Meten schrijfvaardigheid kost veel en levert weinig op

Kloppen de vier argumenten van de aanhangers van de 'meten is niet nodig'-benadering? We gaan eerst uit van de situatie van voor 1940.

Tot ongeveer 1940 gebruikte de College Board een drie uur durende toets om de vaardigheid in schrijven te meten. In die toets moesten studenten vragen over teksten en boeken beantwoorden (twee uur) en een essay schrijven (één uur). Behalve dat de toets veel studenttijd kostte, kostte de toets ook veel nakijktijd. De meting was duur en arbeidsintensief. Voor het meten van schrijfvaardigheid toonde Hopkins al in 1921 aan dat de beoordeling meer afhing van de beoordelaar en het jaar waarin het examen werd afgelegd, dan van wat er daadwerkelijk geschreven was (Godshalk et al., 1966, p. 2). De meting was dus ook onbetrouwbaar. Doordat de meting onbetrouwbaar was, was de meting ook niet erg valide. Leverde de meting nuttige informatie? Ook dit is zeer de vraag. Wanneer een student hoorde dat hij onvoldoende scoorde, was het allereerst de vraag of dat aan de student lag, aan het examen of de beoordelaar. Vervolgens was het niet erg duidelijk hoe een student zichzelf kon bijspijkeren of door een docent bijgespijkerd kon worden. Alle vier kritiekpunten waren daarmee in deze periode terecht

Na de Tweede Wereldoorlog werd de lengte van het College Board examen teruggebracht tot 1 uur en ging men geleidelijk over op objectieve tests (meerkeuzevragen) om de schrijfvaardigheid te testen. Het voordeel van de objectieve tests was dat ze betrouwbaar waren en dat het nakijken amper tijd kostte. Van de vier kritiekpunten vervielen er daarmee twee: onbetrouwbaar en tijdrovend. De vragen die overbleven waren of de informatie van de test valide was en in hoeverre die informatie bruikbaar was. In de ophef die door het gebruik van een volledig objectieve test ontstond, ging het vooral over de eerste vraag. Hoe kon schrijfvaardigheid goed gemeten worden zonder dat de student daadwerkelijk schreef? Objectieve tests waren niet valide om schrijfvaardigheid te meten, was de overtuiging van de docenten. Noyes schrijft hierover in zijn 'Introduction' van *The Measurement of Writing Ability (*in Godshalk et al., 1966, p. IV-V):

> *Gradually and reluctantly the examiners (all English teachers and predisposed to essay questions) turned to objective items. These, at least, could be scored with complete accuracy; in time, the English Composition Test became . . . wholly objective. Before long, the outcry began. By what right could a test which involved no writing whatever be called a test of composition? Further, what would happen to the teaching of writing in the schools, when teachers and students alike knew that writing had vanished from the College Board's Admissions Testing Program?*

Om deze kritiek te weerleggen, kregen Godshalk et al. van de College Board opdracht de validiteit van de objectieve tests aan te tonen. Dit reeds eerder aangehaalde onderzoek (Godshalk et al., 1966) vormde een doorbraak op het gebied van het meten van schrijfvaardigheid. Allereerst lieten Godshalk et al. zien dat het holistische oordeel met veel moeite (vijf topics, vijf beoordelaars per topic) betrouwbaar kon worden gemeten. Ten tweede definieerden ze validiteit objectief als de correlatie met het criterium: de totaalscore van de vijf beoordelingen op de vijf essays (voor die tijd werd het oordeel over de validiteit van schrijfvaardigheidsmaten vooral gebaseerd op de mening van de betrokkene over de inhoud van de maat en niet empirisch bepaald, de zogenaamde inhoudsvaliditeit). Ten derde lieten ze zien dat objectieve tests om schrijfvaardigheid te meten, valide waren in de zin dat ze zeer hoog correleerden met het totale holistische oordeel. Ten vierde volgde hieruit ook dat het holistische oordeel valide was. Het holistische oordeel bleek immers, mits betrouwbaar gemeten, hoog te correleren met andere tests (de objectieve tests) die bedoeld waren schrijfvaardigheid te meten en dat op een totaal andere manier deden.

Hoe zit het met de vier punten van kritiek op het meten van schrijfvaardigheid als de uitkomsten van Godshalk et al. (1966) daarin worden betrokken?

1. De meting is betrouwbaar.
2. De meting is valide.
3. Objectieve tests vergen relatief veel tijd om te construeren, maar weinig tijd om af te nemen en te scoren.
4. De uitkomst van de meting is geschikt om het niveau van een student te bepalen, maar een duidelijke trainingsprocedure ontbreekt nog.

Wanneer we tevreden zijn met het alleen meten van de schrijfvaardigheid, wat ook inderdaad het doel van de College Board was, lijken objectieve tests in beginsel goede mogelijkheden te bieden schrijfvaardigheid betrouwbaar, valide en snel te meten.

Voor de onderwijspraktijk lag dit echter anders. Door het zware accent dat de College Board legde op het objectief toetsen van schrijfvaardigheid nam de motivatie bij docenten en studenten om daadwerkelijk te (laten) schrijven vermoedelijk niet echt toe. Het eerder aangehaalde citaat van Noyes wijst hier op. Verder had men voor de onderwijspraktijk niets aan de objectieve tests, omdat ze de studenten niet daadwerkelijk lieten schrijven, waardoor ze niet bruikbaar waren als instructiemethode. De opkomst van de objectieve tests verbeterde daardoor de situatie in de onderwijspraktijk niet, maar verslechterde die eerder. De 'doorbraak' van Godshalk et al. was dus vooral een doorbraak op het gebied van het meten van schrijfvaardigheid en niet een doorbraak op het gebied van het onderwijs in schrijfvaardigheid, doordat de objectieve tests in de onderwijspraktijk niet konden worden ingepast. Alle vier punten van kritiek op het meten van schrijfvaardigheid bleven voor de onderwijspraktijk van toepassing: 1. het holistische oordeel was hinderlijk onbetrouwbaar; 2. door de lage betrouwbaarheid was het holistische oordeel niet erg valide; 3. het meten van de schrijfvaardigheid door holistisch te beoordelen kostte veel docenttijd en was vermoeiend; 4. het nuttig effect van de holistische beoordeling was zeer beperkt. Meten was misschien wel nodig, maar kostte veel tijd en leverde weinig op.

Helemaal correct is bovenstaande schets onder bepaalde omstandigheden niet. Wanneer een docent veel schrijfopdrachten geeft, neemt het aantal 'topics' en het aantal 'beoordelingen' toe. Hierdoor wordt het totale holistische oordeel belangrijk betrouwbaarder. Wanneer het holistische oordeel betrouwbaarder wordt, is de validiteit niet langer een probleem. De totaal benodigde beoordelingstijd wordt door veel schrijfopdrachten te geven, wel groter. De holistische beoordeling zorgt dan wel dat studenten schrijfervaring opdoen, maar zal verder qua feedback vermoedelijk niet bijzonder effectief zijn.

8.4 Wel een aangetoond verband, geen verklaring

Hoewel het onderzoek van Godshalk et al. (1966) niet tot duidelijke veranderingen in de onderwijspraktijk leidde, kunnen hun empirische resultaten wel bijdragen aan een beter begrip van het construct Schrijfvaardigheid. In werkelijkheid gebeurde dat tot nu toe echter niet of amper. De empirische resultaten werden niet goed begrepen, pasten niet goed in de ontstane discussie en raakten uit het zicht. Men vroeg zich niet af, hoe het merkwaardige door Godshalk et al. aangetoonde verband tussen objectieve tests en het holistische oordeel verklaard kon worden, maar de discussie focuste vooral op het wel of niet valide zijn van objectieve tests (en soms ook het holistische oordeel) als de discussie tenminste zo specifiek werd.

Zo stelde Steinmann (1967) in een bespreking van *The Measurement of Writing Ability* van Godshalk et al. (1966) dat het boek een tekort had aan 'conceptuele validiteit' en fout was. "Writing ability is evidently so complex that probably no test of finite length constructed by persons with finite knowledge could measure it" (Steinmann, 1967, p. 80). De tests van de College Board waren onvoldoende gebaseerd op kennis van de relevante disciplines, vond hij, waarmee hij vermoedelijk vooral Engels en 'composition' bedoelde. Er was verder geen enkel bewijs dat de gebruikte beoordelaars bekwaam waren, vond hij, dus dan was er ook geen enkel bewijs dat de objectieve tests valide waren.

Macrorie (1971) gebruikte, vijf jaar na het verschijnen van *The Measurement of Writing Ability* , bijna tien bladzijden om in zijn boekbespreking aan te geven, hoe verkeerd hij de reductionistische en metende benadering van de College Board vond. Na enkele pittige uitspraken gedaan te hebben, stelt hij zichzelf de retorische vraag of zijn commentaar niet wat te heftig gesteld is. Hij vindt van niet, want: "This is part of the penalty for reducing human communications to numerical responses" (Macrorie, 1971, p. 289). Over de zinnen van een toetsvraag die als voorbeeld in het rapport was opgenomen, schrijft hij: "The words have no more life and tongue in them than the list of numbers on page 77 of the monograph: .710, . 772, .747 " (p. 290). Pagina 77 was een bladzijde in de bijlagen met een correlatiematrix. Even verderop schrijft Macrorie: ". . . it is these very grounds that must be abandoned, entirely. The whole enterprise is a colossal reduction and drying out of life" (p. 290). Er moest op een totaal andere manier lesgegeven gaan worden, vond hij, daarna moest de College Board zichzelf opheffen. "I see that the first step is to teach teachers Other Ways, and once large numbers of their students are producing work that counts, recommend that such orga-

nizations as the College Entrance Examination Board abolish themselves and encourage their members to enter the new kind of teaching" (Macrorie, 1971, p. 290). Hoewel duidelijk lijkt dat Macrorie *The Measurement of Writing Ability* en de uitgangspunten van Godshalk et al. helemaal fout vindt, bevatte zijn bespreking geen concrete punten die de resultaten van Godshalk et al. weerlegden of invalideerden.

Deze voorbeelden van reacties zijn alleen bedoeld om te laten zien dat er tussen de droge, psychometrische benadering van Godshalk et al. waarbij de behaalde criteriumvaliditeiten in duizendsten werden gemeten en de docenten die in de klas Engels en 'composition' gaven, een wereld van verschil lag. De boodschap van Godshalk et al. dat objectieve tests aantoonbaar valide waren en eigenlijk nog valider dan het doorgaans nogal onbetrouwbare holistische oordeel van 'expert-beoordelaars' (docenten), was niet wat docenten graag wilden horen.

Dat het gevonden verband niet leidde tot een beter begrip van schrijfvaardigheid lijkt achteraf gezien begrijpelijk. Docenten vonden het bestaan van het verband moeilijk te accepteren. Godshalk et al. hadden van hun kant een beperkte doelstelling. Ze wilden aantonen dat de eigen objectieve tests van de College Board om schrijfvaardigheid te bepalen valide waren om op deze manier de kritiek uit het onderwijsveld op de College Board te pareren. De publicatie fungeerde niet alleen als onderzoeksverslag, maar ook als verweerschrift. De redenering die Godshalk et al. daarbij volgden, was samengevat als volgt. 'Het holistische oordeel wordt door docenten gezien als een valide maat om schrijfvaardigheid te meten. Wij laten zien dat dat holistische oordeel betrouwbaar gemeten kan worden. Wij laten verder empirisch zien dat onze objectieve tests zeer hoog correleren met dit holistische oordeel. Daaruit volgt dat onze objectieve tests valide zijn.'

De docenten redeneerden echter op basis van inhoudsvaliditeit. De beste manier om schrijfvaardigheid te meten was het holistische oordeel. Objectieve tests om schrijfvaardigheid te meten deden dat duidelijk op een totaal verkeerde manier. Objectieve tests om schrijfvaardigheid te meten waren dus duidelijk niet valide. Dat objectieve tests hoog correleerden met het holistische oordeel, was iets dat eigenlijk niet hoorde.

Godshalk et al. definieerden validiteit als criteriumvaliditeit en concludeerden vervolgens dat objectieve tests valide waren. De docenten zagen validiteit als inhoudsvaliditeit en concludeerden dat objectieve tests totaal anders werkten dan holistische beoordeling en dus

niet valide konden zijn. Iedere partij zag daarmee zijn eigen gelijk bevestigd, terwijl de discussie over de 'validiteit' door de verschillende invullingen van dit begrip, weinig concreets opleverde.

Dat het aangetoonde verband, objectieve tests correleren onverwacht hoog met het holistische oordeel, voor een belangrijk deel uit de belangstelling verdween, lag gedeeltelijk ook aan Godshalk et al. zelf. Allereerst beperkten ze zich niet tot het gevonden resultaat, maar trokken daar onmiddellijk de wat discutabele conclusie uit dat objectieve tests dus valide waren, dat wil zeggen: prima geschikt om schrijfvaardigheid te meten. Maar wanneer schrijfvaardigheid in het onderwijs met volledig objectieve toetsen wordt gemeten, heeft dat gevolgen voor het onderwijssysteem. Studenten en docenten gaan zich op die toetsingsmethode instellen. De 'logica' van een toetsmethode moet voor de betrokkenen duidelijk en inzichtelijk zijn. Onderwijs en toetsingsmethode moeten op elkaar aansluiten. Wanneer het onderwijs uitgaat van schrijfopdrachten, is het vreemd om de toetsing vervolgens op meerkeuzevragen te baseren.

Godshalk et al. hadden een bijdrage kunnen leveren, door in te gaan op de consequenties van het gevonden verband voor het begrip schrijfvaardigheid, maar deden dat niet. Vermoedelijk vonden ze als testconstructeurs dit niet een bijzonder interessante vraag. Verder is 'inhoud' psychometrisch gezien een vaag begrip dat snel in de buurt komt van 'face validity'. De laatste term is bedoeld om aan te geven dat iemand denkt dat een test valide is, maar dat daar verder geen enkel empirisch bewijs voor is. Ook gold in dit geval dat de objectieve tests speciaal geconstrueerd waren om schrijfvaardigheid te voorspellen, dat ze dat vervolgens inderdaad deden, hoefde dan geen verwondering te wekken. Vanuit de optiek van Godshalk et al. ging het vooral om de criteriumvaliditeit, dat wil zeggen de correlatie met het criterium, en viel er verder weinig te verklaren.

Tegelijkertijd realiseerden de auteurs zich dat deze opvatting niet helemaal klopte en dat het gevonden verband wel van belang was voor een beter begrip van schrijfvaardigheid. Aan de verklaring van het gevonden verband besteedden Godshalk et al. echter slechts een terloopse opmerking halverwege het rapport: "a finding that strongly suggests that the characteristics of single sentences play an important part in the overall impression made on the reader" (Godshalk et al. 1966, p. 19). In het slothoofdstuk merken ze wat cryptisch op: "But of even greater importance than this practical outcome has been the development of a clearer understanding of the nature of writing skill" (p. 39). Waaruit dat grotere begrip precies zou bestaan, werd echter niet aangegeven.

Wanneer twee totaal verschillende methodes tot soortgelijke uitkomsten leiden, kan dat wetenschappelijk en praktisch gezien van groot belang zijn. Vanuit wetenschappelijk oogpunt lijkt het merkwaardig om wel een verband aan te tonen, maar vervolgens niet te proberen dat verband te verklaren. In de psychometrie is het echter gebruikelijk correlaties als een gegeven te beschouwen. Bepaalde tests correleren nu eenmaal en vaak valt dat moeilijk te verklaren.

In feite wordt in het rapport van Godshalk et al. (1966) merkwaardig genoeg zelfs de sterkte van het verband niet duidelijk gerapporteerd. In hun conclusies schrijven ze: "When objective questions specifically designed to measure writing skills are evaluated against a reliable criterion of writing skills, they prove to be highly valid" (p. 40). Een voor de hand liggende vraag is dan, hoe valide die objectieve vragen dan precies waren? Men zou verwachten dat het rapport daar een duidelijk antwoord op geeft, maar het rapport vermeldt alleen de correlaties met het criterium. Voor de twee best presterende tests qua criteriumvaliditeit, Usage en Sentence Construction, bedroegen de correlaties met het criterium respectievelijk 0.707 en 0.705. Om deze correlaties zuiver te kunnen interpreteren, moet echter rekening worden gehouden met de onbetrouwbaarheid van de betrokken variabelen. De betrouwbaarheid van het criterium was bekend (0.841) en werd vermeld (Godshalk et al., 1966, p. 12, Table 1). De betrouwbaarheid van de gebruikte tests werd echter door Godshalk et al. niet vermeld. Ook de voor onbetrouwbaarheid gecorrigeerde correlaties met het criterium werden niet vermeld. Hoe sterk het verband tussen de objectieve tests en het totale holistische oordeel was na correctie voor onbetrouwbaarheid, viel daardoor voor lezers van het rapport niet eenvoudig te achterhalen.

Waarom de auteurs van het rapport de betrouwbaarheden van de zes objectieve tests en van de twee gebruikte 'interlinears'[1] nagelaten hebben te vermelden, wordt niet in het rapport aangegeven. Bij lezing van het rapport ontstaat wel de indruk dat de auteurs psychometrisch uitermate goed onderlegd waren en wisten dat de betrouwbaarheid van een variabele als een belangrijk en basaal gegeven geldt. Het lijkt daarmee moeilijk voorstelbaar dat het niet vermelden van deze basale en relevante informatie per abuis is gebeurd. Deze indruk wordt versterkt doordat het rapport op andere punten bepaald niet terughoudend was met het rapporteren van grote aantallen getallen. Zo werden meerdere grote correlatiematrixen volledig afgedrukt.

[1] Bij deze tests kregen de studenten een tekst met fouten te corrigeren. Interlinears worden in 8.5 uitgebreider besproken.

De meest aannemelijke verklaring, die ook in overeenstemming lijkt met de teneur uit nog te bespreken onderzoek (zie 8.6), lijkt dat Godshalk et al. (en de College Board) wel belang hadden bij het verdedigen van hun objectieve tests tegen aanvallen uit het onderwijsveld, maar geen belang hadden dit onderwijsveld (de eigen leden van de College Board) onnodig te bruuskeren. De docenten vonden het idee dat objectieve tests het holistische oordeel konden voorspellen moeilijk om te accepteren, het was niet nodig uit te spellen hoe enorm goed de objectieve tests dat - gecorrigeerd voor onbetrouwbaarheid - precies konden. Om dezelfde reden vonden ze het vermoedelijk ook verstandiger niet in te gaan op de verklaring van het verband. Het verband lag al moeilijk, een verklaring zou het 'foute' verband nog meer accentueren.

8.5 Sterkte en verklaring van het verband

Een lezer die op zoek gaat naar informatie over de betrouwbaarheden van de tests zal zich na enige tijd realiseren dat de betrouwbaarheden van de twee qua criteriumvaliditeit best presterende objectieve tests geschat kunnen worden uit hun onderlinge correlatie (0.775), omdat beide tests in hoge mate als zeer gelijk werden gezien (Godshalk et al., 1966, p. 17-18). Voor beide interlinears gold hetzelfde argument. Deze tests correleerden onderling 0.702.

Op basis van deze waarden vielen vervolgens de voor onbetrouwbaarheid gecorrigeerde correlaties te berekenen van de twee qua criteriumvaliditeit best presterende objectieve tests, Usage en Sentence Construction, en de qua criteriumvaliditeit best presterende interlinear, Valley Forge. De waargenomen correlaties met het totale holistische oordeel waren respectievelijk: 0.707, 0.705 en 0.668. Dit leverde voor onbetrouwbaarheid gecorrigeerde correlaties met het totale holistische oordeel op van respectievelijk: 0.876, 0.873 en 0.869. Gecorrigeerd voor onbetrouwbaarheid bleken de qua criteriumvaliditeit beste objectieve tests om schrijfvaardigheid te meten, daarmee allemaal (afgerond) 0.87 of hoger met het criterium te correleren. Deze waarde geldt als uitermate hoog.

Dit betekent dat de objectieve tests in beginsel (bij perfecte beoordelaarsbetrouwbaarheid) 76% van de betrouwbare variantie in de holistische totaalscore konden verklaren en voorspellen. (Het is misschien verleidelijk om te denken dat als iedere test ongeveer 76% vari-

antie verklaart, de drie tests gezamenlijk zeker 100% zouden moeten verklaren. Dat is echter niet het geval. Omdat de drie tests onderling zeer hoog correleren, verklaren ze steeds ongeveer dezelfde 76% van de variantie in de holistische totaalscore.) Deze objectieve tests waren daarmee niet alleen (criterium)valide, ze waren zelfs zo valide dat het totale holistische oordeel daar nog slechts in zeer beperkte mate van kon afwijken. Kennelijk zijn deze drie tests uiterst belangrijk om schrijfvaardigheid gemeten via het holistisch oordeel te begrijpen.

Van de twee objectieve tests Usage en Sentence Construction was al bekend dat ze onderling hoog correleerden en gezien werden als inwisselbaar. Deze twee tests vormden daarmee de beste representanten van de 'objectieve test'-factor. De interlinear die het hoogst met het totale holistische oordeel correleerde, Valley Forge (Godshalk et al., 1966, p. 52), correleerde ook hoog met deze twee objectieve tests. De gemiddelde onderlinge correlatie tussen de drie maten bedroeg 0.728, zodat ze als één factor beschouwd lijken te kunnen worden.

Dit valt ook anders te zien. Usage en Sentence Construction werden beschouwd als inwisselbaar en vormden daardoor en ook door hun relatief hoge onderlinge correlatie, een enkele factor. De vraag is dan wat de correlatie tussen deze factor en de interlinear Valley Forge zou zijn na correctie voor onbetrouwbaarheid. De gemiddelde correlatie van Valley Forge met deze twee objectieve tests bedroeg 0.704. De geschatte betrouwbaarheden van Usage en Sentence Construction aan de ene kant en Valley Forge aan de andere kant waren respectievelijk: 0.775 en 0.702. De gemiddelde voor onbetrouwbaarheid gecorrigeerde correlatie bedroeg dan 0.954. Dit is dermate hoog dat gesteld kan worden dat de interlinear Valley Forge en beide objectieve tests inderdaad vrijwel volledig dezelfde factor maten. Dit laat zien dat niet het meerkeuze-karakter bepalend is voor de 'objectieve test'-factor, maar de inhoud van de test.

Wanneer we ervan uitgaan dat de 'objectieve test'-factor goed wordt gerepresenteerd door de drie genoemde tests (Usage, Sentence Construction, Valley Forge) is de gemiddelde voor onbetrouwbaarheid gecorrigeerde correlatie van deze factor met het holistische oordeel (0.876+0.873+0.869)/3=0.87. Dit komt overeen met 76% gemeenschappelijke variantie.

Wanneer de 'objectieve test'-factor zo belangrijk is voor holistisch gemeten schrijfvaardigheid dat deze de schrijfvaardigheid voor ruim drie vierde bepaalt, rijst de vraag wat de 'objectieve test'-factor precies inhoudt. Een bijkomende vraag is of de TAVAN-score qua in-

houd ook gezien kan worden als behorend tot deze 'objectieve test'-factor. Om een indruk te krijgen van de inhoud van deze factor kijken we naar twee voorbeeld-items van de tests die het hoogst correleerden met de totale essayscore: Usage en Sentence Construction.
De Usage test van Godshalk et al. bevatte vooral voorbeelden van 'faulty usage' die door de student als fout herkend moesten worden. Een voorbeeld van een (faulty) Usage item van Godshalk et al. (1966, p. 6) was bijvoorbeeld:

He spoke bluntly and angrily to we spectators.
A. bluntly
B. angrily
C. we
D. spectators
E. No error

De student moest aangeven waar de fout zat of als er geen fout in de zin zat, E kiezen. In dit geval was het goede antwoord: C. In TAVAN zou dezelfde zin gepresenteerd worden als:

He spoke bluntly and angrily to we spectators.

Wanneer de student zou antwoorden met:

He spoke bluntly and angrily to spectators.

zou dit antwoord een match opleveren (goed gerekend worden). Andere antwoorden zouden geen match opleveren.

De Sentence Correction test ging ook uit van items met een zin die (meestal) een fout bevatte. De passage met de mogelijke fout was onderstreept. De student moest voor deze passage het best passende alternatief zoeken. Een voorbeeld van een item (Godshalk et al., 1966, p. 6):

While waving goodbye to our friends, the airplane took off, and we watched it disappear in the sky.
A. While waving
B. Upon waving
C. Having waved
D. Waving
E. While we waved

In TAVAN zou hetzelfde item gepresenteerd worden als:

While waving goodbye to our friends, the airplane took off, and we watched it disappear in the sky.

Of in het Nederlands:

> Terwijl vaarwel wuivend naar onze vrienden, steeg het vliegtuig op en zagen we het verdwijnen in de wolken.

Naast deze twee objectieve tests gebruikten Godshalk et al. nog vier andere objectieve tests om schrijfvaardigheid te meten: Paragraph Organization, Prose Groups, Error Recognition en Construction Shift. In de Paragraph Organization test kreeg de student steeds zes zinnen (A, B, C, D, E, F) die in de juiste volgorde moesten worden geplaatst om een betekenisvolle alinea te vormen. Deze test die niet meer op zinsniveau lag, maar op alineaniveau, correleerde beduidend lager dan de overige vijf objectieve tests met het totale holistische oordeel (0.459 tegen 0.643). De Prose Groups test bestond uit items waarin de student een kort stuk tekst kreeg aangeboden waarin één zin was weggelaten. Dit werd aangegegeven met '...' . De student moest vervolgens uit 4 zinnen (A, B, C, D) de zin kiezen die het beste paste in het fragment. Na Paragraph Organization correleerde Prose Groups van de objectieve tests het laagst met het totale holistische oordeel (0.568 tegen 0.662 voor de vier overige tests). Kennelijk spelen fouten op zinsniveau een grotere rol bij de totstandkoming van het holistische oordeel dan fouten op alineaniveau. Helemaal onverwacht is dit niet. Doordat het holistisch oordeel snel gegeven wordt, is het moeilijker fouten in de zinsvolgorde te zien dan duidelijke vormfouten in de zin.

Godshalk et al. gebruikten, zoals reeds opgemerkt in 8.4, naast de zes objectieve tests ook twee interlinears (1966, p. 8). Interlinears bestaan uit slecht geschreven passages met veel fouten en slordigheden met steeds een lege tussenregel voor het aangeven van de correcties. De fouten in de interlinears waren niet onderstreept en moesten door de student opgespoord en gecorrigeerd worden. Bij de beantwoording moest de student dus daadwerkelijk schrijven, althans herschrijven. Een bezwaar van de interlinears was dat de antwoorden door een beoordelaar moesten worden beoordeeld, wat het gebruik ervan arbeidsintensief maakte en enige beoordelaarsonbetrouwbaarheid introduceerde. In de praktijk bleek de overeenstemming tussen beoordelaars echter hoog.

Een voorbeeld van een deel van een passage uit een interlinear:

> Never had the fortunes of England dived to a lower ebb than at the moment
> when Elizabeth ascends the throne.

Na correct herschrijven werd dit:

> Never had the fortunes of England fallen to a lower ebb than at the moment
> when Elizabeth ascended the throne.

Deze passage zou zonder aanpassingen in TAVAN gepresenteerd kunnen worden. Een verschil met TAVAN is dat bij de interlinears de tekst als één geheel werd aangeboden. Hoewel dit bij TAVAN in beginsel wel mogelijk is, zal er normaal naar gestreefd worden de tekst in korte passages van één of enkele zinnen te presenteren. Verder zal getracht worden het aantal fouten per item te beperken tot één of twee om de opdracht eenduidiger te maken en de beoordeling (en ook de feedback) te vereenvoudigen. De structurering in TAVAN is daarmee normaal groter, waardoor de taak voor de student eenvoudiger wordt en ook de beoordeling van de antwoorden.

Een ander verschil tussen TAVAN en de interlinears is dat de student in de interlinears herschrijft door in de ruimte tussen de gedrukte regels de wijzigingen in de tekst aan te geven. Hierdoor wordt de herschreven tekst snel een moeilijk te lezen geheel van pijlen en doorhalingen. Bij TAVAN daarentegen zal de student de zin bewerken in de ingebouwde editor van de antwoordbox, waarbij de zin ook na veel bewerkingen nog steeds goed leesbaar is. Tegelijkertijd kan in de opdrachtbox altijd de originele zin nog geraadpleegd worden.

Het grootste verschil tussen TAVAN en de interlinears is echter dat in TAVAN het antwoord onmiddellijk door het TAVAN-programma gecheckt wordt aan de hand van door de docent of itemconstructeur opgegeven goede antwoorden. Het nakijken gebeurt meteen door de computer(server) waarbij een positief resultaat (online) aan de student wordt meegedeeld en een negatief resultaat blijkt door het antwoord van de docent te vermelden. Door deze automatisering is de testafname qua benodigde docenttijd zeer efficiënt, terwijl door de snelle feedback voor de student een leereffect mogelijk wordt.

De inhoud van de objectieve-testfactor kan nu preciezer omschreven worden. Bij alle drie tests (Usage, Sentence Correction en de interlinear Valley Forge) ging het om het herkennen van fouten in een zin of passage en het corrigeren van die fouten. In feite: vaardigheid in het redigeren van gebrekkige en foute zinnen. De benaming objectieve-testfactor is dan wat misleidend, de benaming 'Fouten Opsporen en Corrigeren' (FOC) lijkt beter aan te sluiten bij de inhoud van deze factor, die we daarom hierna zullen aanduiden als de FOC-factor.

Dit is echter ook precies de vaardigheid die de TAVAN-score meet. Op basis van de inhoud valt daarmee te verwachten dat de TAVAN-score hoog zal correleren met interlinears en objectieve tests om schrijfvaardigheid te meten. Het verschil met de objectieve tests van Godshalk et al. (1966) is dat er in TAVAN geen onderstreept deel is, zodat de fout overal in

de passage kan zitten en dat er in TAVAN geen verbetervoorstellen worden gedaan waaruit de student kan kiezen. De student bepaalt zelf hoe hij de zin herschrijft. Dit maakt TAVAN belangrijk realistischer dan een objectieve test waarbij slechts het beste alternatief gekozen kan worden. Uiteindelijk lijken beide maten echter volledig gebaseerd op de FOC-factor. De student die slecht is in het opsporen van fouten en het corrigeren daarvan zal op beide maten laag scoren.

Het resultaat van Godshalk et al. (1966) met betrekking tot de (hoge) criteriumvaliditeit van de objectieve tests om schrijfvaardigheid te meten, laat zich nu eenvoudig herinterpreteren. Men vond een (zeer sterk) verband (na correctie voor onbetrouwbaarheid meer dan 0.87) tussen het vermogen van studenten om fouten in teksten te herkennen en te corrigeren (de FOC-factor) en het holistische oordeel.

Onze verklaring voor het verband is dan als volgt. De holistische beoordelaars maten in feite niet rechtstreeks de schrijfvaardigheid van de studenten, maar gaven een evaluatief oordeel op basis van de, door de studenten geschreven, teksten. In die teksten kwam via het aantal fouten per honderd woorden de FOC-factor van de betrokken studenten tot uiting. De holistische beoordelaars lieten zich in belangrijke mate door het aantal fouten per honderd woorden in de teksten beïnvloeden, met als resultaat een hoge correlatie tussen het holistische oordeel en de via objectieve tests gemeten FOC-factor. Hoewel beide maten ogenschijnlijk op totaal verschillende manieren tot stand kwamen, maten beide in hoge mate de FOC-factor.

Een andere, complexere, verklaring zou kunnen zijn dat holistische beoordelaars zich in belangrijke mate ook laten beïnvloeden door andere factoren als de inhoud en de lengte van de essays. Van lengte is bekend dat die vaak sterk gecorreleerd is aan het holistische oordeel. Wanneer die overige factoren echter sterk gecorreleerd zijn met het aantal fouten per honderd woorden (de FOC-factor) is het resultaat dat de FOC-factor volstaat voor het voorspellen van het holistische oordeel.

Het lijkt goed mogelijk dat studenten die beter schrijven: a. met minder fouten schrijven per honderd woorden; b. in een gegeven tijdsbestek langere teksten produceren; c. een betere inhoud produceren (beter de boodschap communiceren) doordat ze minder afgeleid worden door problemen met de taal. Een holistische beoordelaar kan zich vervolgens door alle drie factoren laten beïnvloeden. Doordat deze drie variabelen (de FOC-factor, de tekstlengte, de

waardering van de inhoud/boodschap) echter vermoedelijk sterk gecorreleerd zullen zijn (taalgebruik en inhoud zijn in de praktijk bijvoorbeeld vaak niet goed te scheiden), is het voor de voorspelling van het holistische oordeel voldoende naar één of twee van deze variabelen te kijken. De holistische beoordelaar combineert dan wel meerdere soorten informatie tot een totaaloordeel, maar methodes die slechts één of twee van deze soorten informatie gebruiken, kunnen het resultaat van de holistische beoordeling dan goed voorspellen.

Het is in de praktijk moeilijk de inhoud (de boodschap) van een essay te waarderen los van het taalgebruik, daarom lijkt een betere optie de inhoud indirect te meten via een objectieve test die de leesvaardigheid van de student meet. De leesvaardigheid is een maat die aangeeft hoe goed iemand in staat is informatie uit teksten te halen (en vermoedelijk ook uit andere verbale informatie te halen zoals mondelinge antwoorden, mondelinge uitleg). Iemand met een goede leesvaardigheid beschikt daardoor in beginsel over meer informatie (inhoud) dan personen met een slechte leesvaardigheid. Informatie vormt de basis bij het schrijven van een essay. Men mag dus verwachten dat studenten met een goede leesvaardigheid essays zullen schrijven met een betere inhoud.

Men zou dan verwachten dat de score op een objectieve test om schrijfvaardigheid te meten (de FOC-factor), de tekstlengte en de score op een objectieve test die leesvaardigheid meet, alle drie sterk gecorreleerd zullen zijn met het holistische oordeel. In het onderzoek van Breland en Jones (1982, p. 12-13) werden deze variabelen gecorreleerd met twee holistische beoordelingen steeds uitgevoerd door twee beoordelaars: de 'ECT-beoordeling' en de 'PWS-beoordeling' (Perceptions of Writing Skill). De ECT-beoordeling was het holistische deel van de English Composition Test. De gemiddelde correlaties van het objectieve deel van de ECT, van tekstlengte en van de score op het leesdeel van de SAT (Scholistic Aptitude Test) bedroegen respectievelijk: 0.52, 0.51, 0.50. De betrouwbaarheid van beide holistische beoordelingen (de onderlinge correlatie) bedroeg slechts 0.58. De betrouwbaarheden van de overige variabelen vielen niet te achterhalen. Deze gemiddelde correlaties laten zien dat alle drie variabelen hoog en ook ongeveer alle drie even hoog correleren met het holistische oordeel. Het lijkt daarmee goed mogelijk dat alle vier betrokken variabelen (holistische oordeel, FOC-factor, tekstlengte en leesvaardigheid) één gemeenschappelijke factor vormen.

Doordat de FOC-factor 0.87 correleerde met het totale holistische oordeel, hebben de FOC-factor en het holistische oordeel 0.87x0.87x100=75.7% variantie gemeenschappelijk. In-

houd/leesvaardigheid en tekstlengte kunnen dan nog maximaal 24.3% overblijvende variantie verklaren van het holistische oordeel. Hieruit kan echter niet geconcludeerd worden dat ze matig zullen correleren met het holistische oordeel. Waarschijnlijker lijkt dat ze voor een groot deel zullen overlappen met de FOC-factor.

8.6 Eerder onderzoek naar het verband

Het resultaat dat de FOC-factor ongeveer drie vierde van de variantie in het totale holistische oordeel verklaarde, was niet alleen afleidbaar uit Godshalk et al. (1966). Breland en Gaynor (1979) vonden bij een soort replicatie van het onderzoek van Godshalk et al. (1966) een voor onbetrouwbaarheid gecorrigeerde correlatie van 0.90.

Verschillende onderzoeken leveren op dit punt kennelijk soortgelijke uitkomsten. Na correctie voor onbetrouwbaarheid ligt de correlatie tussen objectieve tests om schrijfvaardigheid te meten (de FOC-factor) en het totaal van een aantal beoordeelde essays in het gebied van ongeveer 0.87 tot 0.90. De objectieve tests verklaren daarmee minimaal drie vierde van de betrouwbare variantie in het totale holistische oordeel. Niet-FOC factoren zullen daarom slechts beperkt kunnen bijdragen aan een betere verklaring van het holistische oordeel.

Verrassend genoeg concludeerden Breland en Gaynor (1979, p. 127) uit de gevonden hoge waarde van 0.90 niet, dat de score op de objectieve tests het holistische oordeel vrijwel volledig bepaalde. Zij concludeerden slechts:

> *This study strongly suggests, as have previous studies, that direct and indirect assessments of writing skill . . . tend to tap similar skills. However, indirect measures lack face validity and credibility among members of the English profession and educators generally, and they tend to deliver a message to the students that writing is not important. . . . experience in direct assessment appears to be yielding improvements.*

Breland en Gaynor gingen door het vermelden van deze hoge waarde een stap verder dan Godshalk et al., maar probeerden daarna ijlings de hoge gevonden waarde zo te verpakken dat het onderwijsveld en de eigen leden van de College Board niet onnodig verontrust werden.

De essentie van een voor onbetrouwbaarheid gecorrigeerde correlatie van 0.90 is echter niet dat beide variabelen enigszins gecorreleerd zijn ('tend to tap similar skills'), maar dat ze in feite vrijwel volledig dezelfde factor meten (81% gemeenschappelijke variantie) en dat de hoeveelheid betrouwbare niet verklaarde variantie in het totale holistische oordeel naar verhouding zeer klein is geworden (19%).

Breland en Jones (1982) gaven, ondanks het resultaat dat Breland samen met Gaynor in 1979 gepubliceerd had (de zeer hoge voor onbetrouwbaarheid gecorrigeerde correlatie van 0.90 tussen de objectieve tests en het holistische oordeel) en ondanks het feit dat dit laatste onderzoek in grote lijnen een replicatie vormde van het onderzoek van Godshalk et al. uit 1966, de volgende samenvatting van dit laatste onderzoek (Breland & Jones, 1982, p. 2):

> *Correlations were obtained between students' scores on both the direct (essay) and the indirect (multiple-choice) measures and were found to be sufficiently high (.7) to conclude that the validity of the multiple-choice items was established (Godshalk, Swineford, and Coffman 1966).*

Over de hoge waarde van het verband na correctie voor onbetrouwbaarheid werd niets vermeld.

Ook Cooper (1984) die de door Breland en Gaynor gevonden waarde van 0.90 citeerde, reageerde nogal merkwaardig op deze hoge waarde: "These extreme high estimates raise the question of whether the tests are measuring essentially the same skills" (Cooper, 1984, p. 17). In werkelijkheid was de volgorde andersom. De vraag die men wilde beantwoorden was in hoeverre de twee maten hetzelfde maten, daartoe bepaalde men de correlatie en vervolgens de voor onbetrouwbaarheid gecorrigeerde correlatie. De gevonden waarde vormde het kwantitatieve antwoord op de vraag. Omdat die waarde in dit geval uitermate hoog was, was het antwoord in dit geval relatief eenduidig. Kennelijk was de voor de hand liggende conclusie echter niet de conclusie die Cooper wenselijk vond en 'redeneerde' hij daarom - mogelijk zonder zich dit te realiseren - naar de oorspronkelijke vraag terug.

Om de vraag opnieuw te beantwoorden, het gevonden antwoord werd immers geherinterpreteerd als de aanleiding voor de vraag, verwees Cooper naar een onderzoek van Thompson uit 1967. Thompson probeerde meer grip te krijgen op het holistische oordeel, niet door naar de inhoud van objectieve tests te kijken, maar door essays holistisch en analytisch (per schaal) te laten beoordelen. De beste voorspellers van de holistische scores op 45 papers van studenten waren dan "ratings on three criteria not amenable to multiple-choice testing: 'unsupported statement', 'independent judgement error', and 'lack of unity'" (geciteerd in

Cooper, 1984, p. 17). Thompson concludeerde dat essay-scores in de eerste plaats bepaald werden door hogere orde vaardigheden die niet konden worden gemeten door meerkeuze-vragen (Cooper, 1984, p. 18).

Uit het gegeven dat 'analytische' beoordelaars het holistische oordeel konden voorspellen, valt echter niet af te leiden zoals Thompson deed, dat dit analytische oordeel niet te voorspellen zou zijn via objectieve tests. Ook van het holistische oordeel werd dat immers verondersteld, maar vervolgens bleken objectieve tests het holistische oordeel juist zeer goed te kunnen voorspellen. Het gegeven dat de analytische beoordelaars correleerden met het holistische oordeel, lijkt eerder een aanwijzing (gegeven de minimaal 75% gemeenschappelijk variantie) dat ook het analytische oordeel vrijwel zeker correleerde met de scores op de objectieve tests.

Een tweede probleem dat zich bij het onderzoek van Thompson lijkt voor te doen en dat verderop uitgebreider aan de orde komt, is dat de volgorde van de analytische schalen een sterke uitwerking blijkt te hebben op de correlatie met de holistische beoordeling. In de praktijk lezen de beoordelaars het essay eerst in zijn geheel en vullen daarna de verschillende schalen in. Het gevolg is dat men eerst een holistisch oordeel vormt en dat vervolgens gebruikt om de verschillende schalen in te vullen. Doordat de schalen een bepaalde volgorde bezitten, hebben de schalen in de lijst die het eerst komen, in beginsel het sterkste verband met de holistische beoordeling en wordt dit verband zwakker naarmate men verder van het holistische oordeel verwijderd raakt. Dit zal later in deze paragraaf worden aangetoond op basis van een onderzoek van Breland en Jones (1982) dat een soortgelijke opzet had.

Een derde probleem dat zich bij dit soort onderzoek voordoet, is dat de schalen niet onafhankelijk van elkaar worden beoordeeld, maar afhankelijk van elkaar. Een normale beoordelaar is niet in staat een essay twintig keer opnieuw te lezen, zonder zich te herinneren wat hij eerder als beoordeling heeft gegeven. In de praktijk wordt het essay eenmaal gelezen en het holistische oordeel dat de beoordelaar dan heeft, wordt vervolgens geprojecteerd op de verschillende schalen.

In beginsel is het ook mogelijk dat beoordelaars een essay holistisch beoordelen op meerdere dimensies. In dat geval zouden de analytische schalen een aantal duidelijke factoren moeten opleveren die enerzijds betrouwbaar te meten zouden moeten zijn bij verschillende

beoordelaars en die anderzijds relatief onafhankelijk (ongecorreleerd) van elkaar zouden moeten zijn. Een dergelijke factorstructuur kan echter pas aangetoond worden, wanneer het eerder vermelde volgorde-probleem opgelost is, omdat dit de gevonden correlaties ernstig kan verstoren.

Als tweede bron voor een antwoord op de vraag in hoeverre objectieve tests en het holistische oordeel dezelfde vaardigheden meten, verwees Cooper (1984, p. 19) naar een onderzoek van Breland en Jones (1982) waarin analytische beoordeling werd gebruikt om meer licht te werpen op het holistische oordeel. Breland en Jones gebruikten een aselecte steekproef van 806 essays geschreven in 20 minuten in december 1979 in het kader van de English Composition Test (ECT) van de College Board. Deze essays waren beoordeeld door twee beoordelaars. Verder hadden de studenten ook het objectieve testdeel van de ECT (40 minuten) gemaakt en waren nog een aantal gegevens van hen beschikbaar. De essays werden opnieuw beoordeeld (de PWS-beoordeling) door twee beoordelaars per essay die de essays eerst holistisch en daarna analytisch op twintig verschillende punten beoordeelden. Breland en Jones probeerden hierna de holistische ECT-beoordeling te voorspellen uit alle beschikbare variabelen via lineaire regressie. Cooper merkte op basis van 'Table 9' van Breland en Jones op: "When direct and indirect assessments are combined, the multiple correlation jumps to .70, suggesting that for this population the instruments tap closely related but distinct skills" (Cooper, 1984, p. 19). De onderstreping is van Cooper. De holistische score zou naast de objectieve test factor nog een extra factor bevatten.

Wie echter het onderzoek van Breland en Jones (1982, p. 15, Table 9) bestudeert, ziet dat de bijdrage aan de voorspelling kwam van beoordelaars die beschikten over dezelfde essays (hetzelfde topic) als waarop het te voorspellen criterium gebaseerd was. Dit was daarmee een nogal oneigenlijke vergelijking, omdat topics specifieke variantie bevatten die niet gerelateerd is aan de schrijfvaardigheid van de student (gemeten volgens het totale holistische oordeel), terwijl de objectieve tests natuurlijk niet over die specifieke informatie konden beschikken. Vervolgens bleek het objectieve testdeel van de ECT ondanks deze informatie-handicap toch even goed het criterium te voorspellen als de holistische PWS-beoordelaars. Beide correlaties bedroegen: 0.58 (Breland & Jones, 1982, p. 13, Table 6). Dit wijst er daarmee niet op dat het holistische oordeel een tweede factor bevatte, maar wijst er juist op dat het holistische oordeel die niet bevatte. Het objectieve testdeel voorspelde het holistische criterium precies even goed als de nieuwe holistische beoordeling door twee beoordelaars.

Wanneer het objectieve testdeel als voorspeller werd gebruikt in combinatie met de lengte van het essay (Breland & Jones, 1982, p. 15, Table 9), bleek de belangrijkste analytische voorspeller van de PWS-beoordelaars (de schaal Overall Organization) de multiple correlatie amper te verhogen (van 0.69 naar 0.72). Wanneer men beschikte over de score op het objectieve testdeel en verder over de lengte van het geschreven essay, bevatte de belangrijkste samenvatting van het analytische oordeel amper nieuwe informatie. Extra informatie door een tweede beoordeling van hetzelfde essay verbeterde de voorspelling slechts zeer beperkt (van 48% naar 52% verklaarde variantie). Kennelijk levert de beoordeling van een serie teksten dus weinig nieuwe informatie op wanneer de scores op het objectieve testdeel (40 minuten) bekend zijn, alsmede de lengte van de geschreven teksten.

Breland en Jones (1982, p. 1) zelf kwamen echter tot een conclusie die volledig tegenovergesteld was aan de eerder gevonden relatie tussen de FOC-factor en het holistisch oordeel. Omdat de relatie tussen de score op objectieve tests en het holistische oordeel vele malen empirisch is aangetoond, lijkt hun conclusie meer 'politiek correct' dan empirisch juist.

> *The results showed that certain characteristics of discourse, in contrast to syntactic and lexical characteristics, influenced judgments the most. The characteristics of discourse included organization, transition, use of supporting evidence, and the originality of ideas presented. In the sample examined, traditional syntactic emphases-such as subject-verb agreement, punctuation, and pronoun usage-had less influence on scores assigned. The results suggest that instruction in English composition courses should emphasize discourse skills.*

Het probleem met hun onderzoek is dat ze het holistische oordeel over een serie teksten (afkomstig van één topic) probeerden te verklaren door dezelfde serie teksten voor te leggen aan andere beoordelaars die de teksten opnieuw moesten beoordelen. De nieuwe beoordelaars moesten eerst hun holistische oordeel geven en daarna het analytische deel van het formulier invullen (Breland & Jones, 1982, p. 6, Figure 1). Op deze manier werd in feite tweemaal holistisch beoordeeld waarbij de beoordelaars in de tweede ronde hun oordeel moesten motiveren door de twintig analytische schalen in te vullen. Dit leidde er allereerst toe dat het holistische oordeel door de beoordelaars in de tweede ronde vertaald werd in twintig schaalwaarden. De twintig schaalwaarden vormden daardoor geen nieuwe informatie, maar alleen een soort motivering van het gegeven holistische oordeel. Er is daardoor weinig garantie dat de uitgebrachte analytische beoordelingen inderdaad betrekking hadden op specifieke aspecten van de beoordeelde teksten.

Wanneer de analytische beoordeling inderdaad onafhankelijk van de holistische beoordeling was uitgevoerd en wanneer ook alle schalen onafhankelijk van elkaar waren beoordeeld (in plaats van achter elkaar per tekst), zou men verwachten dat de analytische beoordeling belangrijk meer informatie zou bevatten dan de eerste of tweede holistische beoordeling. Door twintig keer te meten zal men normaal immers meer informatie verzamelen dan door slechts eenmaal te meten. Uit 'Table 6' (Breland & Jones, 1982, p. 13) bleek echter dat de correlatie van de analytische somscore 0.57 correleerde met het ECT holistische oordeel en 0.86 met het PWS holistische oordeel. De eerste correlatie was iets lager dan de betrouwbaarheid (de onderlinge correlatie) van de holistische oordelen (0.58). De totale analytische beoordeling gebaseerd op twintig schalen bevatte dus iets minder informatie dan het holistische (ECT) oordeel. Dat de tweede correlatie veel hoger was dan de eerste laat zien dat de analytische beoordelingen en het PWS holistische oordeel inderdaad sterk gecorreleerd waren en niet het resultaat waren van onafhankelijk beoordelingen.

Verder viel te zien in 'Table 4' (Breland & Jones, 1982, p. 12) dat alle negen 'Discourse Characteristics' minimaal 0.30 en maximaal 0.52 correleerden met het holistische ECT-oordeel. Men zou verwachten dat sommige kenmerken vrij hoog zouden correleren en andere juist vrij laag, wanneer de beoordelingen inderdaad betrekking hadden op verschillende aspecten van de teksten. In werkelijkheid correleerden alle kenmerken zwak positief met het holistische ECT-oordeel. Dit wijst erop dat alle schalen dezelfde factor maten en onderling positief correleerden. Via 'Appendix E' (de correlatiematrix) was dit eenvoudig te controleren geweest, maar deze bijlage bleek in de pdf van de College Board waarmee het rapport online was gezet, om onbekende redenen niet opgenomen te zijn (op 27-08-2013, 23.00 uur).

De twintig dimensies op het formulier stonden in een volgorde van 1 tot 20. De volgorde waarin de dimensies op het beoordelingsformulier voorkwamen bleek bij narekenen significant en zeer sterk te correleren ($r=0.77$, $p=0.000$, $N=20$) met de mate waarin die dimensie correleerde met de holistische PWS-beoordeling (Breland & Jones, 1982, p. 12, Table 4 en Table 5). De beoordelaars hadden de begrijpelijke neiging de eerste dimensies overeenkomstig hun holistische oordeel in te vullen en tendeerden naarmate ze verder kwamen met het invullen van het formulier meer naar het gemiddelde. Omdat de dimensies gegroepeerd waren in drie groepen in de volgorde: Discourse Characteristics, Syntactic Characteristics, Lexical Characteristics en de 'Discourse Characteristics' voorop stonden, was het resultaat

dat deze categorie als belangrijkste uit het onderzoek kwam. De uitkomsten vallen dus te verklaren als een volgorde-effect dat door de onderzoekers vooraf (onbewust) is aangebracht. Niet het belang van de verschillende schalen werd gemeten, maar de volgorde op het beoordelingsformulier.

Het onderzoek naar de sterkte en de verklaring van het verband tussen objectieve tests en holistische beoordeling lijkt daarmee niet erg voorspoedig verlopen te zijn. Allereerst was de grote sterkte van het - voor onbetrouwbaarheid gecorrigeerde - verband een beladen onderwerp, waardoor het lang duurde voor de hoge gevonden waarde duidelijk geformuleerd werd. Toen de sterkte van het verband ten slotte wel duidelijk gesteld werd, was de conclusie zo beladen dat men die niet trok. In plaats daarvan spande men zich in de hoge gevonden waarde als niet belangrijk voor te stellen of zelfs weer terug te gaan naar de oorspronkelijke vraag. Verder richtte men de aandacht niet op het conflict tussen inhoudsvaliditeit en criteriumvaliditeit (de inhoud was sterk verschillend, de uitkomsten niet), maar op de inhoud van het holistische oordeel door dit te koppelen aan een tweede holistische beoordeling. De manier waarop men dat deed, was dat men beoordelaars holistisch liet beoordelen en daarna een aantal analytische schalen liet invullen. Het resultaat was vooral dat men opnieuw het holistische oordeel mat als verklaring van hetzelfde holistische oordeel, maar nu geprojecteerd op een een groot aantal schalen. Dat de volgorde van de schalen sterk doorwerkte in de resultaten, realiseerde men zich daarbij niet.

Dat het onderzoek naar de sterkte en de verklaring van het verband niet erg opschoot, mag daarmee duidelijk zijn. De vraag, waarom het onderzoek op dit punt weinig opschoot, is daarmee nog niet beantwoord. Een recent onderzoeksrapport van de College Board (Shaw & Kobrin, 2012) suggereert bij nauwkeurige lezing een verklaring. In dit rapport wordt getracht na te gaan wat de bijdrage is van het SAT-essaydeel aan de validiteit van de SAT (oorspronkelijk: Scholastic Aptitude Test). De SAT is in de VS de door universiteiten meest gebruikte toelatingstest. Het doel van de SAT is het functioneren van de toekomstige student aan de universiteit zo goed mogelijk te voorspellen; de correlatie met het FYGPA (First Year Grade Point Average) moet daarom zo hoog mogelijk zijn. In maart 2005 werd in de SAT een essaydeel opgenomen. Dit deel bestaat uit een vraag die de student in 25 minuten tijd zo goed mogelijk moet beantwoorden door een kort essay te schrijven. Het geschreven essay wordt vervolgens holistisch beoordeeld door twee beoordelaars op een schaal van 1 tot 6 zodat een student minimaal twee punten krijgt en maximaal twaalf punten. Het voordeel van het essaydeel is dat studenten weten dat ze moeten schrijven, het na-

deel is dat de essayscore relatief onbetrouwbaar is, waardoor de bijdrage aan de voorspelling uiteindelijk gering of zelfs afwezig is. Shaw en Kobrin (2012) merken over dit punt op (p. 4):

> *When the predictive value of a test is measured by the correlation of students' scores on the test with college outcome measures such as first-year grade point average and grades in college courses, the indirect writing assessments will always prevail.*

Het zou echter kunnen dat een combinatie van beide maten, het holistische oordeel valt immers niet helemaal samen met de objectieve testfactor, de beste voorspelling levert van het FYGPA.

Deze veronderstelling lijkt simpel te toetsen. De eerste voorspeller is de SAT-score zonder het essaydeel, de tweede voorspeller is de SAT-score met het essaydeel. De voorspeller die het hoogst correleert met FYGPA heeft gewonnen. Verder zijn we niet beperkt tot vaststellen welke voorspeller het het beste doet, maar kunnen we via het verschil tussen beide correlaties vaststellen, hoeveel het essaydeel precies aan de voorspelling toevoegt, de zogenaamde incrementele validiteit.

Over deze gebruikelijke benadering merkten Shaw en Kobrin (2012, p. 9) echter terloops op:

> *Because the essay score scale is much smaller, and scores are less variable than scores on the other SAT sections, using a traditional regression approach to determine the contribution of essay scores to the prediction of college grades would most likely lead one to incorrectly surmise that essay scores do not provide any increment to the prediction.*

Ze gaan er dus vanuit dat de normale benadering zou laten zien dat het essaydeel vermoedelijk helemaal niets bijdraagt aan de voorspelling. De kritische lezer zal zich op dit punt afvragen, waar de auteurs dit stellige vermoeden ('most likely') op baseren. Kennelijk hebben de auteurs op dit punt meer informatie dan de lezer: heeft men misschien eerst de gebruikelijke benadering gevolgd?

Dat de 'essay score scale' veel kleiner is en dat de scores misschien 'less variable' zijn (wanneer men daarmee bedoelt dat de SD relatief klein is, is dat in feite een herhaling van het eerste punt, de schaal is vrij kort), maakt echter voor het toepassen van een traditionele regressie benadering helemaal niets uit, omdat juist uitgegaan wordt van interval-variabelen.

De SD van een schaal wordt dus niet gezien als een absoluut gegeven, maar als iets dat per schaal verschilt en daarom voor de correlatie-berekening gestandaardiseerd wordt. Het argument om niet de gebruikelijke benadering te volgen, lijkt daarmee nogal uit de lucht gegrepen. Kennelijk is het probleem niet dat de essayscore-schaal relatief kort is, maar dat die niets toevoegt aan de voorspelling, aan de correlatie met het criterium.

Shaw en Kobrin vervolgen hierna (p. 9): "As will be shown in our analysis, this conclusion would be misinformed. The *straightforward approach*, or graphical presentation of the data, depicts a different story" (cursivering door Shaw en Kobrin). Met andere woorden: de gebruikelijke benadering leidt tot een verkeerde uitkomst, de rechttoe-rechtaan benadering van Shaw en Kobrin vertelt een heel ander verhaal.

Dat verhaal wordt verteld in Figuur 2 van hun rapport en in hun Tabel 3. In hun tekst merken ze daarover op: "Figure 2 shows the clear relationship between SAT essay scores and FYGPA, after essentially controlling for SAT scores and HSGPA. Table 3 includes the values that Figure 2 is based on" (Shaw & Kobrin, 2012, p. 9). De lezer die hun Tabel 3 bestudeert, vindt daarin tachtig getallen in vier kolommen voor in totaal twintig verschillende categorieën. Volgens de titel betreft het: "The Relationship between SAT Essay Score and Academic Performance". In de sociale wetenschappen is het echter gebruikelijk de relatie tussen twee variabelen uit te drukken in een enkel getal (meestal de productmoment correlatie) en desgewenst kan men vervolgens dat verband nog eenvoudig grafisch weergeven via het scatterdiagram. Men kan zich dus afvragen of een benadering waarbij een relatie tussen twee variabelen wordt weergegeven via tachtig getallen wel helemaal zo 'straightforward' is als de auteurs beweren.

De lezer die vervolgens op zoek gaat naar uitleg om de tabel beter te begrijpen, stuit op de merkwaardige passage in de tekst (die hiervoor geciteerd werd) die luidt: "after essentially controlling for SAT scores and HSGP" (p. 9). Bij sociaal wetenschappelijk onderzoek kan een variabele als de SAT-score op twee manieren gecontroleerd worden. De eerste manier is dat men alleen studenten gebruikt die precies dezelfde SAT-score hadden. De tweede manier is dat men statistisch corrigeert voor verschillen in SAT-score. Maar wat wordt bedoeld met 'essentially controlling'?

Na uitgebreide bestudering van de tabel blijkt men hiermee te bedoelen dat men de studenten op basis van de SAT-score zonder essaydeel, heeft ingedeeld in vijf niveaugroepen. De

hoger scorende groepen blijken het dan vervolgens op het essaydeel beter te doen dan de lager scorende groepen.

Wat Shaw en Kobrin daarmee dus in feite op een zeer ingewikkelde manier aantonen, is dat de SAT-score zonder het essaydeel, positief correleert met de score op dat essaydeel (zonder dat we daarbij overigens de precieze correlatie meegedeeld krijgen). Niemand heeft dit echter ooit betwijfeld. Uiteraard correleert het essaydeel positief met het overige deel van de SAT. Dit was niet het probleem waar het om begonnen was. De vraag was of het essaydeel iets toevoegde aan de voorspelling. Het knappe van Shaw en Kobrin is echter dat ze dit nogal nietszeggende resultaat op zo'n manier weten te brengen, dat het toch al snel enkele uren lezen en zoeken kosten, voordat de lezer precies kan aangeven wat er niet klopt.

Vermoedelijk levert het essaydeel in de SAT geen duidelijke bijdrage aan de criteriumvaliditeit, maar was deze conclusie niet de conclusie die de auteurs eigenlijk wilden. In dit verband moet men zich realiseren dat de auteurs in dienst waren van de College Board. Hoewel de College Board een non-profit organisatie is, realiseerde ze in 2011 een omzet van 0,721 miljard US-dollar (Google Search, College Board) met onder andere de verkoop van toelatingstests aan haar leden: de aangesloten onderwijsinstellingen. De auteurs hebben daarmee belang bij een uitkomst die de afnemers van de College Board positief zullen evalueren. Het punt dat bij Godshalk et al. (1966) al een rol speelde en in 9.4 als mogelijke verklaring werd vermeld, namelijk dat men het onderwijsveld niet onnodig wilde bruuskeren, bleef kennelijk ook daarna de uitkomsten van het College-Board-onderzoek sturen.

Hoewel de College Board op het gebied van holistische beoordeling de belangrijkste onderzoeksinstelling is, heeft iets soortgelijks vermoedelijk ook voor veel ander onderzoek op dit gebied gegolden. De onderzoekers hadden nauwe banden met het onderwijsveld of waren daar zelf uit afkomstig. Het gevolg was dat de politieke correctheid van de conclusies in veel gevallen een grote rol in het onderzoek ging spelen. Men 'ontdekte' via het onderzoek niet hoe het werkelijk was, maar bevestigde vooral hoe men dacht dat het behoorde te zijn.

8.7 Constructvaliditeit basale schrijfvaardigheid

In eerste instantie werd schrijfvaardigheid gemeten via het holistische oordeel. Een docent ziet, als 'expert-beoordelaar', of een student goed of slecht schrijft, was het idee. Het holis-

tische oordeel werd beschouwd als inhoudsvalide. Godshalk et al. (1966) lieten vervolgens zien dat objectieve tests om schrijfvaardigheid te meten, hoog met de holistische totaalscore correleerden. Objectieve tests waren niet inhoudsvalide (de studenten schreven niet), maar wel criteriumvalide. Hierna waren er twee verschillende manieren bekend om schrijfvaardigheid te meten: de directe en de indirecte methodes. Schrijfvaardigheid was, hoewel niet iedereen dat wilde accepteren, een construct geworden. Een derde manier van meten moet dan bij voorkeur overeenstemmen met de twee reeds bestaande manieren. De drie verschillende maten moeten zo mogelijk in belangrijk mate soortgelijke uitkomsten geven. Ook voor een eventuele vierde manier, en zo verder, zou dit moeten gelden.

In het fouteneffect-onderzoek (deelstudie 4) schreven 48 studenten teksten over hun ervaringen met TAVAN die vervolgens holistisch beoordeeld werden door beide onderzoekers, door andere studenten en die beoordeeld werden op aantal fouten per honderd woorden door beide onderzoekers. Doordat de studenten die deze teksten schreven, hadden meegedaan aan het TAVAN2-programma, was ook de TAVAN-score bekend (het percentage 'matchende' antwoorden). Door de resultaten van het fouteneffect-onderzoek te koppelen aan de resultaten van TAVAN2 werd het daardoor mogelijk vier verschillende variabelen om schrijfvaardigheid te meten, aan elkaar te relateren: 1. het holistische oordeel over de door de student geschreven tekst, 2. de TAVAN-score van de student, 3. het aantal fouten per honderd woorden in de door de student geschreven tekst en 4. de kwaliteit van de door de student uitgebrachte holistische oordelen.

Godshalk et al. (1966) lieten zien dat objectieve tests om schrijfvaardigheid te meten hoog correleerden met het holistische oordeel. In paragraaf 8.5 lieten we zien dat de hoogst correlerende objectieve tests qua inhoud de FOC-factor maten, die ook gemeten werd door de TAVAN-score. Daarom was de verwachting dat de TAVAN-score en het holistische oordeel sterk zouden correleren.

Doordat het aantal fouten per honderd woorden ook een index vormde voor de FOC-factor, maar dan direct gemeten (via een geschreven tekst), leek het zelfs mogelijk dat het holistische oordeel, de TAVAN-score en het aantal fouten per honderd woorden in belangrijke mate onderling zouden kunnen correleren en alle drie in belangrijke mate dezelfde factor zouden meten. Wanneer dit inderdaad het geval zou zijn, zou dit een belangrijk resultaat betekenen, omdat (basale) schrijfvaardigheid vervolgens op drie verschillende manieren kan worden vastgesteld. Dit zou een belangrijke uitbreiding betekenen op de twee manieren die na Godshalk et al. bekend waren.

In dat geval zou het in beginsel zelfs gaan om vier verschillende manieren. Tussen de TAVAN-score en de gebruikelijke objectieve tests bestaat immers een belangrijk onderscheid: de TAVAN-score meet direct, een objectieve test meet indirect. Voor de TAVAN-score moet daadwerkelijk geschreven, althans herschreven worden, terwijl het bij een objectieve test voldoende is het juiste alternatief te kiezen.

Strikt genomen zou het dan nog mogelijk zijn dat de TAVAN-score qua inhoud wel de FOC-factor zou meten, maar qua correlatie niet of niet erg zou overeenstemmen met het resultaat van een objectieve test. Wanneer de TAVAN-score echter, na correctie voor onbetrouwbaarheid, belangrijk overeenstemt met het holistische oordeel komt dit argument grotendeels te vervallen. Het is immers niet mogelijk dat objectieve tests en de TAVAN-score beiden 75% gemeenschappelijke variantie hebben met het holistische oordeel zonder elkaar grotendeels (voor tenminste 50% van de variantie) te overlappen.

Daar de TAVAN-score qua inhoud vooral de FOC-factor lijkt te meten en het aantal fouten per honderd woorden ook, kan men zich afvragen of dit daadwerkelijk twee verschillende soorten maten zijn. Qua inhoud is dit inderdaad het geval. Het verschil zit in de openheid van de opdracht. Bij het aantal fouten per honderd woorden bepaalt de student zelf hoe hij de zin formuleert en welke zin hij wil formuleren. In beginsel is het daarmee mogelijk ieder lastige constructie te omzeilen. Bij de TAVAN-score gaat het om het herschrijven van zinnen met één of meer fouten, waarbij de zin zo beperkt mogelijk wordt aangepast. De TAVAN-score kan men vermoedelijk het beste opvatten als een geautomatiseerde interlinear, terwijl het aantal fouten per honderd woorden in beginsel op ieder willekeurige tekst kan worden toegepast. De ene maat gaat uit van een door de student geschreven tekst, de andere maat gaat meer uit van een testsituatie.

In het fouteneffect-onderzoek werd nog een 'vijfde' manier toegepast (de methode van objectieve tests werd in het fouteneffect-onderzoek niet gebruikt): de kwaliteit van de student als holistische beoordelaar. Iemand die goed kan schrijven, bleek beter in staat teksten op kwaliteit te rangordenen. Wanneer we iemand een stel qua schrijfniveau verschillende teksten geven, valt uit de rangordening van de teksten in beginsel af te leiden hoe goed iemand schrijft.

Hoewel de betrouwbaarheid van deze methode in het geval van het fouteneffect-onderzoek beperkt was (0.42), zou dit door met bijvoorbeeld meerdere beoordelingsrondes te werken

eenvoudig op te lossen zijn. Een interessant punt voor deze vijfde methode, is dat de tweede, de derde en de vierde methodes allemaal vooral de FOC-factor lijken te meten. Voor deze vijfde methode lijkt dat niet automatisch noodzakelijk te zijn. Qua inhoud focust de methode immers niet bij voorbaat op fouten. Doordat het holistische totaaloordeel ongeveer 25% niet aan het aantal fouten (de FOC-factor) gebonden variantie bevat, zou deze methode wellicht kunnen helpen licht te werpen op deze resterende 25%.

Een zesde methode is in beginsel ook mogelijk, maar is in het fouteneffect-onderzoek niet toegepast. Laat een student de fouten in een tekst signaleren. Naarmate een student meer bevestigde fouten signaleert, kan hij beter schrijven. De bevestigde fouten in de tekst fungeren als items die wel (goed) of niet (fout) gesignaleerd worden. Voor de beoordelaars van het foutenonderzoek waren op dit punt wel gegevens beschikbaar, waaruit bleek dat beoordelaars op dit punt inderdaad sterk verschilden, maar waren geen andere schrijfvaardigheidsmaten bekend om de validiteit van deze methode te bevestigen.

Deze zesde en laatste methode lijkt een vereenvoudigde en open variant van een objectieve test die Godshalk et al. (1966, p. 7)) gebruikten: Error Recognition. Bij deze test moest de student voor een zin aangeven of die wel of niet een fout bevatte en om welke soort fout het dan precies ging (van in totaal vier verschillende soorten fouten). Deze Error Recognition test correleerde wel hoog met het totaal van de holistische beoordelingen (0.592), maar toch belangrijk lager dan Usage en Sentence Correction (respectievelijk 0.707 en 0.705). Mogelijk werd deze belangrijk lagere criteriumvaliditeit veroorzaakt door de ingewikkelde formulering van de items en de te specifieke vraagstelling.

In totaal zouden er dan volgens het voorgaande zes qua inhoudsvaliditeit verschillende manieren bestaan om schrijfvaardigheid vast te stellen:

1. het holistische oordeel over door de student geschreven tekst;
2. een objectieve test afgenomen bij de student;
3. de TAVAN-score op basis van door de student herschreven zinnen in TAVAN;
4. het aantal fouten per honderd woorden in door de student geschreven tekst;
5. de kwaliteit van het holistische oordeel uitgebracht door de student;
6. het percentage van alle bevestigde fouten die een student signaleert in een tekst.

Hierbij moet variabele 4 omgedraaid worden door te vermenigvuldigen met -1 om negatieve correlaties te vermijden. Een student die veel fouten maakt, zal immers normaal een slechte schrijfvaardigheid bezitten in plaats van een goede.

Wanneer al deze zes maten het construct 'basale schrijfvaardigheid' meten, zouden ze onderling qua resultaten in beginsel moeten overeenstemmen. Door de gegevens van het fouteneffect-onderzoek en TAVAN2 te combineren, was het mogelijk voor de manieren 1, 3, 4 en 5 na te gaan of deze veronderstelling klopte.

Mate van overeenstemming en aantal factoren

Wanneer stemmen verschillende maten qua resultaten overeen? Met 'overeenstemmen' wordt in dit geval bedoeld 'correleren': de hoogte van de productmoment-correlaties tussen de verschillende maten. Wanneer twee maten hoog correleren, betekent dit nog niet dat beide maten in de praktijk ongeveer dezelfde waarden zullen opleveren, doordat de correlatie betrekking heeft op gestandaardiseerde variabelen (gemiddelde 0 en SD 1). Voor een gegeven steekproef van studenten/teksten zijn het gemiddelde en de SD van een meetmethode niet informatief, omdat ze methode-specifiek zijn. Dit punt wordt hier vermeld, omdat een soortgelijk 'probleem' zich voordoet bij holistische beoordeling en onderzoekers in dat geval vaak tot ingewikkelde constructies besluiten om de numeriek verschillende uitkomsten vergelijkbaar te maken, terwijl er statistisch gezien in feite geen probleem was, maar na die ingrepen vaak wel.

Zo vermeldden Prenger en De Glopper (2011, p. 80):

> *Nadat alle teksten door* [moet vermoedelijk zijn: 'van'] *de studenten beoordeeld waren, zijn de oordelen van de beoordelaars per tekst bekeken en geanalyseerd. Als criterium bij het vergelijken van de tekstoordelen hebben we een verschil van minder dan 30 punten in tekstoordeel als aanvaardbaar beschouwd. Bij een verschil van meer dan 30 punten . . . moest de betreffende tekst opnieuw beoordeeld worden door een vierde (nieuwe) beoordelaar. De extreemste score van de vier werd dan verwijderd en vervolgens werd het gemiddelde bepaald van de drie overgebleven scores.*

Na deze herbeoordeling in de meeste extreme gevallen berekenden de auteurs de gemiddelde onderlinge correlatie tussen de beoordelaars als 0.65. Indien men die herbeoordeling achterwege had gelaten, was de gemiddelde onderlinge correlatie een zinvol gegeven geweest, maar door de gevolgde procedure is de betekenis van deze waarde nu onduidelijk. Een eenvoudige check die men wel op de beoordelaars had moeten en kunnen uitvoeren, namelijk of iedere beoordelaar een positieve gecorrigeerde itemtotaal-correlatie had, liet

men echter achterwege of werd althans niet vermeld. Dit voorbeeld is alleen bedoeld als illustratie, soortgelijke voorbeelden zijn op grote schaal te vinden.

Bij het interpreteren van de gevonden correlaties moet in beginsel rekening worden gehouden met de onbetrouwbaarheid van de desbetreffende variabelen doordat een variabele nooit hoger kan correleren dan de vierkantswortel van zijn betrouwbaarheid. Een variabele die met een betrouwbaarheid van 0.36 gemeten is, kan daarmee in beginsel nooit hoger dan 0.60 met een andere variabele correleren. Dit probleem speelde vooral een rol bij de meting van hoe goed een student holistisch kon beoordelen. Deze meting was relatief onbetrouwbaar, doordat ze slechts gebaseerd was op zes waarnemingen. Daarom zijn ook de voor onbetrouwbaarheid gecorrigeerde correlaties berekend. Een nadeel van deze correctie is dat speciaal bij kleine aantallen en bij geschatte betrouwbaarheden die mogelijk afwijken van de juiste betrouwbaarheid, de berekende waarde een grote foutenmarge kan vertonen.

Ten slotte kan men zich afvragen, hoe sterk de verschillende maten onderling (gemiddeld) moeten correleren na correctie voor onbetrouwbaarheid, om te kunnen stellen dat er inderdaad sprake is van een gemeenschappelijke factor. Bij testconstructie kunnen items die bijvoorbeeld gemiddeld 0.15 onderling correleren, toch een uiterst homogene en betrouwbare test opleveren. Dit voorbeeld is echter misleidend, omdat de betrouwbaarheid van de items (de onderlinge correlatie) ook 0.15 is. De gemiddelde voor onbetrouwbaarheid gecorrigeerde correlatie komt daardoor op precies 1 uit. Hoewel de items onderling relatief laag correleren, meten ze allemaal perfect dezelfde gemeenschappelijke factor. Bij methodes die qua inhoud sterk verschillen, zal men echter doorgaans ook na correctie voor onbetrouwbaarheid lang geen perfecte correlaties vinden. Op deze vraag valt dus moeilijk bij voorbaat een prescriptief antwoord te geven.

Een laatste belangrijk punt is of de correlatiematrix verklaard kan worden door één factor of dat meerdere factoren vereist zijn. Vaak zal in de praktijk coëfficiënt alfa voor dit doel berekend worden. Strikt genomen bewijst alfa echter niet automatisch dat een serie maten unidimensionaal is (verklaard kan worden door een enkele factor) of multidimensionaal (alleen verklaard kan worden door meerdere factoren). Alfa is gebaseerd op de gemiddelde onderlinge correlatie en het aantal items of subschalen. Logischer lijkt het dan om te kijken naar de gemiddelde onderlinge correlatie zodat het aantal maten niet langer een rol speelt. Een volgende voor de hand liggende stap is om te corrigeren voor de onbetrouwbaarheid zoals hiervoor werd geïllustreerd, omdat de hoogte van de onderlinge correlaties sterk af-

hangt van de betrouwbaarheid waarmee de variabelen gemeten zijn. Wanneer nu na correctie voor onbetrouwbaarheid een hoge gemiddelde onderlinge correlatie gevonden wordt, betekent dit in ieder geval dat de verschillende maten op zijn minst één factor gemeenschappelijk moeten hebben. Kan hier echter ook uit afgeleid worden dat er geen tweede factor benodigd is om de gevonden correlaties te verklaren?

Wanneer basale schrijfvaardigheid opgebouwd zou zijn uit twee onafhankelijke factoren, zou de ideale situatie zijn dat, uitgaande van vier variabelen (A, B, C en D), twee variabelen de eerste factor zouden meten en twee variabelen de tweede. In dat geval zouden de vier correlaties tussen enerzijds A en B en anderzijds C en D laag uitvallen. terwijl de twee correlaties tussen A en B en tussen C en D hoog zouden uitvallen. De resulterende gemiddelde onderlinge correlatie zou dan laag uitvallen. Wanneer de onderlinge gemiddelde correlatie hoog uitvalt, is dit duidelijke twee factormodel niet plausibel.

Een minder duidelijke mogelijkheid voor een tweede factor zou kunnen zijn dat de twee variabelen A en B onderling duidelijk hoger correleren evenals de twee variabelen C en D. In dat geval zouden alle correlaties duidelijk positief zijn, maar zouden sommige correlaties belangrijk hoger zijn dan andere. Wanneer alle correlaties ongeveer even hoog zijn, is ook deze mogelijkheid niet plausibel.

Het resultaat van Godshalk et al. (1966) na herinterpretatie via de correctie voor onbetrouwbaarheid voor de overeenstemming tussen objectieve tests en het holistische oordeel, gaf aan dat beide soorten maten ongeveer 75% variantie gemeenschappelijk hadden. Het holistische oordeel lijkt daarmee mogelijk nog een tweede factor te bevatten die niet verklaard wordt door objectieve tests (de FOC-factor). Het probleem hierbij is echter dat zo lang er geen derde maat is, die duidelijk correleert met deze tweede factor, het moeilijk is die tweede factor ondubbelzinnig en lost van het holistische oordeel aan te tonen. Iedere meetmethode bevat in beginsel een deel unieke variantie. De discussie wordt daarom doorgaans beperkt tot het verklaren van de correlaties tussen de verschillende meetmethodes, dat wil zeggen: tot de gemeenschappelijke variantie.

Hoewel er een groot aantal indices voorgesteld zijn om het onderscheid tussen één factor en twee factoren te kwantificeren, is geen enkele index echt ingeburgerd en geaccepteerd, vervolgens zal vaak ook de interpretatie een probleem zijn. Een eenvoudige en pragmatische oplossing lijkt dan het schatten van de eerste factor te zijn op grond van de beschikbare va-

riabelen en vervolgens voor deze eerste factor te controleren door de matrix met partiële correlaties te gebruiken. Correlaties die door de eerste factor niet verklaard worden, zullen als duidelijke niet-nul correlaties in deze matrix overblijven. Om de hoogte van deze partiële correlaties aan te geven, kan vervolgens weer de gemiddelde onderlinge partiële correlatie gebruikt worden. Een andere mogelijkheid is het uitvoeren van een factoranalyse.

Gebruikte schrijfvaardigheidsmaten

Hierna worden de vier gebruikte schrijfvaardigheidsmaten uit TAVAN2 en het fouteneffect-onderzoek kort besproken. Het holistische oordeel over de door de studenten geschreven teksten was gebaseerd op beide onderzoekers (als expert-beoordelaars) en op een wisselend aantal studenten. In de ene versie van deze maat werd alle 48 teksten gebruikt ongeacht het aantal studentbeoordelaars. Voor deze maat werden de gestandaardiseerde scores opgeteld van drie subschalen: onderzoeker A, onderzoeker B en het totaal van de studenten die de tekst hadden beoordeeld. Doordat het aantal studenten dat een bepaalde tekst beoordeelde soms klein was, was de betrouwbaarheid van deze totale score relatief laag met een alfa van 0.73. In de andere versie van deze maat werden alleen die teksten gebruikt die door tenminste zes studenten waren beoordeeld. De betrouwbaarheid (alfa) van deze maat bedroeg 0.83.

Deze schattingen van de betrouwbaarheid van de holistische beoordelingen zijn vermoedelijk wat te hoog, doordat het in feite beoordelaarsbetrouwbaarheden zijn. De schattingen zijn slechts gebaseerd op één enkel topic. Wanneer de studenten ook een tekst over een ander topic hadden geschreven, zou de overeenstemming tussen de gezamenlijke beoordelaars voor beide topics vermoedelijk belangrijk lager uitvallen (de score-betrouwbaarheid).

De reden om het oordeel van beide onderzoekers afzonderlijk op te nemen, was dat één onderzoeker dezelfde betrouwbaarheid bleek te hebben als het totaal van ten minste zes studenten. De reden om steeds te standaardiseren was dat anders door verschillen in de spreiding de ene (groep) beoordelaar(s) veel meer invloed kon krijgen dan de andere (groep) beoordelaar(s), wat niet de bedoeling was.

De TAVAN-score was gebaseerd op de resultaten die de studenten behaald hadden op de drie deellessen van les 2 bij de tweede keer dat het TAVAN-programma werd gegeven (TAVAN2). Les 2 werd gekozen, omdat de studenten op dat moment inmiddels wel gewend

waren aan TAVAN, terwijl er naar verwachting nog geen groot leereffect ontstaan kon zijn. Les 2 fungeerde daardoor als een soort nulmeting. De score werd berekend door via lineaire regressie de score op het peiltoetsdeel van les 2 te voorspellen vanuit de twee overige delen en vervolgens het gemiddelde van deze drie variabelen (de twee voorspellingen en de score op het peiltoets-deel) te berekenen. De alfa-betrouwbaarheid van les 2 bleek op deze wijze berekend 0.97 te zijn (met drie 'items').

Het aantal fouten per honderd woorden in de door de student geschreven tekst werd vastgesteld per teksthelft voor iedere beoordelaar en vervolgens gestandaardiseerd. Vervolgens werd per teksthelft het gemiddelde van beide beoordelaars gebruikt voor de berekening van de alfa-betrouwbaarheid van beide helften gezamenlijk. De gevonden alfa bedroeg 0.90 op basis van twee 'items' (subscores).

Als maat voor hoe goed een student beoordeelt, de kwaliteit van het uitgebrachte holistische oordeel, zal men doorgaans geneigd zijn een correlatie-coëfficiënt te berekenen. Doordat echter slechts zes waarnemingen beschikbaar waren per student en bij het berekenen van een correlatie een aantal vrijheidsgraden verloren gaan, leek dat in dit geval niet de beste oplossing. Uiteindelijk is daarom gekozen voor een afstandsmaat. Wanneer bijvoorbeeld volgens de gemiddelde holistische beoordeling van beide onderzoekers de juiste rangorde van de zes teksten van boven naar beneden respectievelijk '3, 1, 6, 5, 4, 2' was en de beoordeling van de student was '3, 1, 5, 6, 2, 4', dan waren de respectievelijke afstanden '0, 0, 1, 1, 2, 2'. De som van de afstanden was dan 6. Op deze wijze was het via een split-half methode ook mogelijk de betrouwbaarheid te bereken. Deze bedroeg 0.42 voor beide helften samengenomen.

Het grootste bezwaar van een afstandsmaat is dat een student die het qua oordeel erg goed doet, zeer laag scoort. De afstand tussen zijn beoordeling en het gecombineerde oordeel van alle anderen is dan immers minimaal. Dit bezwaar kon echter eenvoudig opgelost worden door uit te gaan van de maximale afstand en daar de gevonden afstand vanaf te trekken (de omgedraaide afstand). Een student met een perfect oordeel scoorde dan maximaal.

Relaties tussen de schrijfvaardigheidsmaten

In Tabel 8.1 is de correlatiematrix weergegeven met de correlaties tussen de vier variabelen om de basale schrijfvaardigheid te meten. Onder de diagonaal zijn de gevonden correlaties

weergegeven, boven de diagonaal de voor onbetrouwbaarheid gecorrigeerde correlaties. Op de diagonaal zijn de betrouwbaarheden vermeld. De voor onbetrouwbaarheid gecorrigeerde correlatie tussen het holistische oordeel en het aantal fouten PHW is volgens de formule groter dan 1, maar kan in werkelijkheid uiteraard hoogstens 1 worden. Deze afwijking komt door steekproef-onnauwkeurigheid en onnauwkeurigheid in de schattingen van de betrouwbaarheden. Wanneer deze correlatie op 1 wordt gesteld, is de gemiddelde voor onbetrouwbaarheid gecorrigeerde correlatie van de variabelen 1 tot en met 4 met de andere drie variabelen respectievelijk: 0.91, 0.79, 0.89, 0.88.

Men kan deze waarden opvatten als een maat voor de constructvaliditeit van de desbetreffende maten. In plaats van de voor onbetrouwbaarheid gecorrigeerde correlatie met een enkel criterium te gebruiken, gebruiken we nu de gemiddelde correlatie, na correctie voor onbetrouwbaarheid, met alle beschikbare criteria. Hierbij moet wel opgemerkt worden dat het een relatief kleine steekproef betreft en dat door de correctie voor onbetrouwbaarheid te gebruiken de gevonden waarden nog verder kunnen afwijken van de 'echte' waarde. Aan de verschillen tussen de maten kan daardoor niet al te veel gewicht worden toegekend. Wel lijkt duidelijk te zijn, dat de gevonden waarden rond of boven de 0.80 liggen en daarmee zeer hoog zijn. De gemiddelde voor onbetrouwbaarheid gecorrigeerde correlatie tussen de vier variabelen bedroeg 0.87. Dit komt overeen met gemiddeld 76% gemeenschappelijke variantie tussen de verschillende variabelen.

Tabel 8.1 Correlaties (linksonder) en voor onbetrouwbaarheid gecorrigeerde correlaties (rechtsboven) tussen de vier variabelen bedoeld basale schrijfvaardigheid te meten. Op de diagonaal de schattingen van de betrouwbaarheid.

	1	2	3	4
1. Holistische Oordeel	(0.73)	0.84	1.11	0.90
2. TAVAN-score	0.71	(0.97)	0.74	0.80
3. Aantal Fouten PHW	0.90	0.69	(0.90)	0.93
4. Kwaliteit (uitgebrachte) Holistische Oordeel	0.50	0.51	0.57	(0.42)

Coëfficiënt alfa voor de eerste drie variabelen samengenomen (en gebaseerd op de waargenomen, ongecorrigeerde correlaties) bedroeg 0.91. De alfa voor alle vier variabelen samengenomen bedroeg 0.87. De alfa voor de vier variabelen wanneer ze perfect betrouwbaar gemeten zouden zijn, bedroeg 0.96. Het lijkt daarmee duidelijk dat alle vier variabelen in hoge mate dezelfde factor maten.

Visuele inspectie van de correlatiematrix leert dat de voor onbetrouwbaarheid gecorrigeerde correlaties allemaal uitermate hoog zijn en vlak bij elkaar liggen qua hoogte, zodat de verschillende variabelen elkaar in zeer hoge mate lijken te overlappen en er weinig ruimte lijkt voor het bestaan van een tweede factor.

Om te checken of er mogelijk een tweede factor benodigd zou kunnen zijn voor de verklaring van de correlaties is de basale schrijfvaardigheid geoperationaliseerd als de somscore van de vier gestandaardiseerde schrijfvaardigheidsvariabelen. Vervolgens is de partiële correlatiematrix berekend waarbij statistisch gecontroleerd werd op de somscore. Dit resulteerde in een matrix met vijf negatieve correlaties en één positieve correlatie. Deze overwegend negatieve partiële correlaties wijzen er niet op dat de vier schrijfvaardigheidsmaten gezamenlijk nog een tweede factor meten.

Om deze uitkomst te checken is een factoranalyse (principale componenten) uitgevoerd op de correlatiematrix van de vier variabelen om basale schrijfvaardigheid te meten. Deze factoranalyse leverde één hoofdcomponent op met een eigenwaarde groter dan 1 (de eigenwaarde was: 2.86) die 74.0% van de gemeenschappelijke variantie verklaarde. De daarop volgende factor verklaarde slechts 14.7% van de variantie (met een eigenwaarde van 0.56). Beide analysemethodes leveren daarmee geen steun voor de veronderstelling dat basale schrijfvaardigheid zou bestaan uit meer dan één factor.

Bij de interpretatie van deze uitkomst dient men zich echter te realiseren dat dit een negatief resultaat is. Er is gezocht naar een tweede factor, maar die is niet gevonden. Dat men na zoeken zijn sleutelbos niet vindt, toont niet aan dat die sleutelbos niet kan bestaan. Van een eventuele tweede schrijfvaardigheidsfactor is bekend dat die vermoedelijk hoogstens 20% of misschien nog minder van de gemeenschappelijke variantie zal verklaren. In verhouding tot de eerste factor (de FOC-factor) is dit een relatief zwakke factor die daardoor lastig aan te tonen valt. Om die factor via factoranalyse of een soortgelijke techniek aan te tonen, moeten tenminste twee variabelen die factor meten en daarbij liefst ook nog betrouwbaar zijn. Van de twee tests die deze factor mogelijk zouden kunnen meten (het holistische oor-

deel en de kwaliteit van de uitgebrachte holistische oordelen), was één echter buitengewoon onbetrouwbaar. Tenslotte is de resulterende correlatiematrix gebaseerd op een relatief klein aantal gevallen waardoor steekproeffluctuaties een grote rol konden spelen. Aan het resultaat dat er geen aanwijzingen werden gevonden voor een tweede factor mag daardoor in dit geval niet al te veel gewicht worden gehecht.

Het positieve resultaat was dat de vier qua inhoud sterk verschillende maten om basale schrijfvaardigheid vast te stellen, in hoge mate dezelfde uitkomsten leverden. Dit betekent dat de TAVAN-score, het holistisch oordeel, het aantal fouten per honderd woorden en de kwaliteit van het uitgebrachte holistisch oordeel in beginsel alle vier geschikte (valide) maten zijn voor het vaststellen van basale schrijfvaardigheid.

De relevantie van deze uitkomst kan gemakkelijk over het hoofd gezien worden, tenzij men zich realiseert dat er tot nu toe in feite maar één geaccepteerde methode was om schrijfvaardigheid rechtstreeks (door de student te laten schrijven) vast te stellen. Dat was de geschreven tekst holistisch te beoordelen. Deze procedure had echter een aantal belangrijke bezwaren. Beoordelaars stemden vaak slecht overeen. De ene beoordelaar kon qua kwaliteit (de gemiddelde correlatie met andere beoordelaars) en qua gemiddelde en SD sterk afwijken van andere beoordelaars. Voor een betrouwbare beoordeling waren daarmee meerdere beoordelaars nodig. Tenslotte was de beoordeling vermoeiend en arbeidsintensief. Voor de onderwijspraktijk betekende dit dat schrijfvaardigheid niet eenvoudig meetbaar was.

Wanneer er echter een cluster van variabelen bestaat, waarvan iedere variabele afzonderlijk gebruikt kan worden om basale schrijfvaardigheid vast te stellen, lijkt het meetprobleem daarmee op zijn minst belangrijk vereenvoudigd te worden, doordat men vervolgens uit verschillende methodes kan kiezen. Het gevonden resultaat lijkt daarmee voor de onderwijspraktijk een grote relevantie te kunnen hebben.

8.8 Effectief schrijfonderwijs en automatisch meten

In 8.3 werd gesteld dat in de situatie tot nu toe (alleen het holistische oordeel en objectieve tests waren tot nu toe als meetmethodes beschikbaar) het meten van schrijfvaardigheid veel kosten met zich mee bracht, maar in de onderwijspraktijk vaak weinig opleverde. Geldt

deze uitspraak ook nog na de resultaten gepresenteerd in de vorige paragraaf? In Tabel 8.2 wordt een overzicht gegeven van de verschillende nu beschikbare methodes om schrijfvaardigheid te meten uitgaande van de vier punten van kritiek die in 8.2 werden vermeld: 1. onbetrouwbaarheid van de maat; 2. niet valide zijn van de maat; 3. duur, arbeidsintensief en tijdrovend zijn van de meting (kosten) 4. de meting levert weinig bruikbare informatie.

Het vierde punt, de bruikbaarheid van de informatie die de meting oplevert, is daarbij geherformuleerd tot (bruikbaar als) 'trainingsmethode' en voorop gezet. Bij het meten van schrijfvaardigheid ligt de focus vaak primair op de betrouwbaarheid, vervolgens op de validiteit. De kosten van de meting duiken soms in de discussie op, maar het doel, het nut van de meting komt zelden ter sprake. Logischer lijkt het, het nut van de meting voorop te stellen. Wat doet men voor nuttigs met de informatie uit de meting? Op welke manier wordt die informatie zinvol gebruikt? Het alleen kunnen vaststellen van het niveau van schrijfvaardigheid is voor selectiedoeleinden wel zinvol, maar binnen het kader van een volledig onderwijssysteem niet, tenzij er een duidelijke methode voorhanden is een tekortschietend niveau gericht te verbeteren. Uitgaande van dit principe en dit doortrekkend, moeten meetmethode en onderwijsmethode op zijn minst voor een deel samenvallen en in elkaar over kunnen gaan. De manier waarop gemeten wordt, moet ook bruikbaar zijn als trainingsmethode, wil de meetmethode in het kader van onderwijs zinvol bruikbaar zijn.

Tabel 8.2 Overzicht kritiekpunten voor verschillende methodes om schrijfvaardigheid te meten (-- = slecht, --/++ = matig, ++ = goed, *=verwacht).

	holistische beoordeling	objectieve test	aantal fouten	TAVAN-score	KHO-methode (kwaliteit uitgebrachte holistische oordeel)
1. trainingsmethode	--/++	--	--/++	++	++*
2. tijd/kosten	--	++	--	++	++*
3. validiteit	--/++	++	++	++	++
4. betrouwbaarheid	--/++	++	++	++	++

Holistische beoordeling per essay kost relatief weinig tijd, maar wanneer een docent het werk van enkele groepen studenten moet beoordelen, betekent dit snel enkele uren vermoeiend en geestdodend werk. Als trainingsmiddel is het holistische oordeel matig geschikt, want de feedback die het levert aan de student komt te laat en is te weinig specifiek. De betrouwbaarheid van het holistische oordeel kan voor onderzoeksdoeleinden groot gemaakt worden, maar is uitgaande van één beoordelaar en één schrijfopdracht doorgaans minimaal. Wanneer de betrouwbaarheid minimaal is, is de validiteit dat ook. In het geval dat een docent veel teksten laat schrijven en beoordeelt, vervallen de bezwaren betreffende de geringe betrouwbaarheid en validiteit grotendeels. Daarom is hier in de tabel 'matig' ingevuld. Uitgaande van het holistische oordeel lijkt de kritiek 'het kost veel en levert weinig op' overwegend terecht.

Objectieve schrijfvaardigheidstests kosten wel tijd om te construeren, maar kosten eenmaal gemaakt weinig tijd om af te nemen en te scoren. In de tabel zijn de kosten daarom aangegeven met '++'. Voor training en onderwijs lijken objectieve tests niet geschikt. De betrouwbaarheid en de validiteit kunnen prima zijn (Godshalk et al., 1966).

Of objectieve tests overigens inderdaad altijd ongeschikt zijn voor trainingsdoeleinden zoals doorgaans wordt aangenomen, lijkt twijfelachtig. TAVAN laat de studenten herschrijven en beoordeelt onmiddellijk de gegeven antwoorden. De items worden objectief gescoord, maar zijn ondanks dat wel bruikbaar voor trainingsdoeleinden. De Delftse Methode (Montens & Sciarone, 1992) bestaat voor een belangrijk deel uit oefeningen waarbij in een zin het juiste woord moet worden ingevuld. Kennelijk is de stelling dat objectieve items niet geschikt zijn voor trainingsdoeleinden in zijn algemeenheid niet juist. Wel lijkt plausibel dat gangbare meerkeuzevragen bestaande uit bijvoorbeeld vier antwoord-alternatieven minder geschikt zullen zijn voor trainingsdoeleinden doordat de items complex zijn om te lezen en juist de zwakke studenten mede daardoor de neiging zullen hebben te gaan raden.

Nakijken op aantal fouten per honderd woorden kost veel tijd en is vermoeiend. Qua tijd en kosten scoort deze methode dus slecht. Als trainingsmethode lijkt deze methode beperkt geschikt, doordat de student vooral te horen krijgt wat allemaal niet goed was. Aan de andere kant levert de methode wel specifieke feedback op basis waarvan een stuk kan worden bijgesteld. Beoordeling op aantal fouten per honderd woorden bleek vaak betrouwbaar te zijn en bleek ook hoog te correleren met het holistische oordeel. De methode is daarmee ook valide.

Een TAVAN-les is relatief snel te construeren en wordt daarna automatisch afgenomen en nagekeken. Qua benodigde docenttijd scoort TAVAN daarmee positief. Een TAVAN-les is door de snelle en gerichte feedback en de automatische aanbieding van de zinnen een effectief trainingsmiddel. Dat bleek uit het onderzoek naar de effectiviteit van het TAVAN-programma (deelstudie 3, hoofdstuk 6). De TAVAN-score bleek verder zeer betrouwbaar te zijn en bleek hoog te correleren met het aantal fouten per honderd woorden en het holistische oordeel. De TAVAN-score is daarmee ook constructvalide.

Studenten teksten laten beoordelen, het meten van de kwaliteit van door de student uitgebrachte holistische oordelen (KHO-methode), is een nieuwe in het kader van het fouteneffect-onderzoek ontwikkelde meetmethode. Bij het fouteneffect-onderzoek bracht het organiseren van de beoordeling door de studenten nogal wat hoofdbrekens met zich mee, waardoor de procedure in eerste instantie vrij bewerkelijk was. Door stroomlijning en vereenvoudiging van de beoordelingstaak lijkt dit probleem oplosbaar. De antwoorden bij deze testmethode zijn objectief scoorbaar en daarmee ook door de computer te beoordelen.

Uitgaande van zes teksten ter lengte van één A4 lijkt de test te open en te ongestructureerd om effectief te kunnen zijn als trainingsmethode. In plaats van een zestal teksten zou het aantal teksten per keer (per item) teruggebracht kunnen worden naar twee, terwijl de tekstlengte gereduceerd zou kunnen worden tot bijvoorbeeld een enkele zin. In dat geval lijken de items door de vereenvoudiging en grotere structurering als indirecte trainingsmethode bruikbaar (het idee achter een indirecte trainingsmethode is dat een student het verschil moet kunnen zien tussen een goede en een slechte zin; de student leert een discriminatie.)

Een item zou er dan bijvoorbeeld als volgt uit kunnen zien. De opdracht voor de student is het beste alternatief te kiezen.

A. Hij hoort van zijn kinderen dat het regelmatig voorkomt, dat er niet goed op hen gelet wordt.
B. Hij hoort van zijn kinderen dat er regelmatig niet goed op hen gelet wordt.

Bijzonder aan deze objectieve testmethode is dat het oordeel gevraagd wordt over de zinnen, zodat het ook mogelijk is twee correcte zinnen voor te leggen waardoor de items niet beperkt zijn tot het opsporen van fouten en het corrigeren daarvan (de FOC-factor). Deze nieuwe maat zou daarmee qua inhoud voor een deel iets anders kunnen meten dan de FOC-factor. Door voldoende items te gebruiken, kan de methode zeer betrouwbaar worden. Af-

gaande op de eerder gerapporteerde, voor onbetrouwbaarheid gecorrigeerde correlatie met de overige schrijfvaardigheidsmaten is de methode ook valide. In Tabel 8.2 is via '*' aangegeven dat het oordeel over de bruikbaarheid als trainingsmethode en het oordeel over de benodigde tijd verwachtingen zijn; op dit moment zijn deze punten nog niet in de praktijk gerealiseerd.

Deze tabel overziend valt vooral de TAVAN-score op met op alle vier punten een positieve waardering. Deze methode lijkt daarmee de eerste schrijfvaardigheidsmaat die ook goed bruikbaar is als onderwijsmethode waarbij daadwerkelijk geschreven wordt, althans herschreven. Verder is belangrijk voor de toepassing in de praktijk dat de kosten laag zijn. Het schrijfonderwijs met TAVAN is qua benodigde docenttijd niet kostbaar, doordat er geen nakijktijd benodigd is en ook amper voorbereidingstijd. Wel is er (op dit moment nog) tijd benodigd voor de coördinatie (het checken en klaarzetten van de lessen en het verwerken van de uitkomsten). Het argument dat meten kostbaar is en weinig oplevert, gaat daardoor voor deze methode niet op. Doordat het TAVAN-onderwijs in vergelijking met traditioneel onderwijs vele malen effectiever blijkt te zijn (zie deelstudie 3, paragraaf 6.3), levert de gebruikelijke docenttijd vele malen meer leerwinst op dan traditioneel schrijfvaardigheidsonderwijs. De kosten van deze methode zijn dus beperkt, terwijl de opbrengst groot is. Het meetresultaat wordt tijdens het onderwijsproces als het ware automatisch meegeleverd en is bovendien ook nog valide en betrouwbaar.

In de tweede plaats vallen de verwachte mogelijkheden van de KHO-methode op die net als de TAVAN-score op alle vier punten positief scoort. Ook deze methode kan zonder veel problemen via een programma als TAVAN geautomatiseerd aangeboden en verwerkt worden. Een ogenschijnlijk bezwaar is dat er bij deze methode niet daadwerkelijk geschreven wordt. Schrijfvaardigheid bevat echter ook een indirecte component. Een goede schrijver is in staat te zien wat de betere zin, fragment of tekst is. Zonder die vaardigheid is het immers niet mogelijk de meest optimale constructies te kiezen. Men kan dit vergelijken met fotografie. Een goede fotograaf weet wat een effectief plaatje is en wat minder effectief is. Pas daarna kan hij proberen dat plaatje te construeren. Een bijzonderheid van de KHO-methode zou kunnen zijn dat deze methode het wellicht mogelijk maakt een eventuele tweede (non-FOC) factor te meten.

8.9 Tien manieren om schrijfvaardigheid te meten

Men kan zich afvragen, welke manieren er nog meer zijn om (basale) schrijfvaardigheid te meten. In Tabel 8.3 wordt een overzicht gegeven van een tiental methodes die gebruikt kunnen worden om basale schrijfvaardigheid te meten. De eerste vijf daarvan zijn reeds besproken in paragraaf 8.8, de tweede vijf nog niet.

Tabel 8.3 Tien manieren waarop (basale) schrijfvaardigheid gemeten kan worden (*= gebruikt in tenminste één deelonderzoek; += aangetoond valide; +/- = soms valide; d=directe maat dat wil zeggen gebaseerd op geschreven of herschreven tekst; i= indirecte maat waarbij niet geschreven wordt; FOC=meet overwegend Fouten-Opsporen-en-Corrigeren-factor).

1	*	d	+	FOC Holistische Oordeel (HO) (over tekst/teksten van student)
2		i	+	FOC Objectieve schrijfvaardigheidstests
3	*	d	+	FOC Aantal Fouten Per Honderd Woorden (AF-PHW)
4	*	d	+	FOC TAVAN-score
5	*	i	+	FOC Kwaliteit Holistische Oordeel (KHO) (uitgebracht door student)
6	*	d	+	--- Lengte van de tekst (aantal woorden) bij een beperkte schrijftijd
7	*	i	?	--- Inschatting eigen schrijfvaardigheid
8	*	d	+	FOC Score op 'linears' (open antwoord test met te verbeteren foute zinnen)
9		i	+	FOC Vocabulaire-omvang (test) / Verbale Intelligentietests
10		d	?	--- Gebruikte woorden/woordkeuze (in door student geschreven tekst)

Het argument dat schrijfvaardigheid eigenlijk niet te meten valt, lijkt gezien het aantal vermelde maten, wat overdreven. Bij iedere maat kan men uiteraard volhouden dat het niet de ideale manier is om schrijfvaardigheid te meten. Hierbij is de ideale manier die men zich voorstelt, dermate complex, dat die ideale manier eigenlijk niet te operationaliseren valt. Verder heeft men doorgaans ook geen pogingen ondernomen op dit gebied. Het eerder in ander verband aangehaalde citaat van Steinmann (1967, p. 80) verwoordt deze opvatting: "Writing ability is evidently so complex that probably no test of finite length constructed by persons with finite knowledge could measure it."

De lengte van het essay (het aantal woorden) is vaak een goede indicator voor de schrijfvaardigheid, maar op het moment dat dit de enige indicator is die gebruikt wordt en dit is bij studenten bekend, dan is het voor studenten mogelijk gebruik te maken van deze kennis en zou de maat niet langer valide kunnen zijn. In het algemeen produceren goede schrijvers echter in dezelfde tijd belangrijk meer tekst dan slechte schrijvers.

Breland, Bonner en Kubota (1995, p. 9, Table 10) vonden een correlatie van gemiddeld 0.72 tussen de lengte van het essay en het holistische oordeel. Deze correlatie was hoger dan van ieder andere variabele met het holistische oordeel en ook hoger dan de correlatie met de SAT-verbal (deze was gemiddeld 0.54), de TSWE (deze was gemiddeld 0.46) en het objectieve deel van de ECT (gemiddeld 0.47) met het holistische oordeel. De TSWE (Test of Standard Written English) dient om studenten in te delen in niveaugroepen.

Breland en Jones (1982, p. 13-14) vonden een correlatie van 0.58 tussen tekstlengte en het holistische ECT-oordeel. De beoordelaarsbetrouwbaarheid van dit holistische oordeel (twee beoordelaars) voor één topic werd vermeld en was 0.58. De scorebetrouwbaarheid van het holistische oordeel (de correlatie met een tweede topic) ligt normaal belangrijk lager (de scorebetrouwbaarheid is het produkt van beoordelaarsbetrouwbaarheid en de topicbetrouwbaarheid). Coffman (1966, p. 154) vermeldde voor de topicbetrouwbaarheid van de ECT een waarde van 0.68. Dit levert een scorebetrouwbaarheid van 0.39. Wanneer we ervan uitgaan dat de tekstlengte volledig betrouwbaar gemeten werd, is de voor onbetrouwbaarheid gecorrigeerde correlatie met het holistische oordeel 0.83. Doordat ook de tekstlengte in werkelijkheid lang niet perfect betrouwbaar is (de tekstlengte van een ander topic correleert wel, maar niet perfect), is dit een minimum-schatting. In werkelijkheid zal de voor onbetrouwbaarheid gecorrigeerde correlatie nog hoger liggen. Dit laat zien dat tekstlengte in de praktijk een zeer valide indicator van schrijfvaardigheid kan zijn.

Het verband tussen tekstlengte en het holistische oordeel wordt normaal gebaseerd op dezelfde topictekst. Dit betekent dat de holistische beoordelaar zich bij zijn oordeel (onbewust) sterk kan baseren op de tekstlengte. In dat geval zou tekstlengte gezien worden als een valide voorspeller van schrijfvaardigheid, maar dat in werkelijkheid (gemeten met een andere maten dan het holistische oordeel) mogelijk niet zijn. In het foutenonderzoek (zie deelstudie 1, paragraaf 4.3) werd echter een sterk negatief verband (-0.65) gevonden tussen tekstlengte en het aantal fouten per honderd woorden, terwijl in langere teksten juist duidelijk meer fouten werden gesignaleerd (0.58). Dit duidt erop dat het taalgebruik in kortere

teksten verhoudingsgewijs inderdaad meer te wensen overlaat: betere schrijvers produceren in dezelfde tijd langere teksten.

Een mogelijk probleem met tekstlengte is dat er niet onder alle omstandigheden een duidelijk verband gevonden wordt met andere variabelen om schrijfvaardigheid te meten. Zo bleek in het fouteneffect-onderzoek (zie deelstudie 4, paragraaf 7.3) de tekstlengte wel in een schaal te passen met het holistische oordeel, het aantal fouten per honderd woorden, de inschatting van de eigen schrijfvaardigheid en de TAVAN2 score op les 2 met een positieve gecorrigeerde item-totaal correlatie van 0.36, maar bleek deze voor de andere 'items' van de schaal veel hoger te liggen (tussen 0.67 en 0.76). In dit geval was de voor het schrijven beschikbare tijd echter niet gebonden aan een strikte limiet, terwijl de studenten vaak wel zeer bij het onderwerp betrokken waren, zodat ook de minder gemakkelijk schrijvende studenten toch gemotiveerd waren hun mening te geven.

Studenten hebben doorgaans wel een idee van hoe goed ze schrijven in verhouding tot hun medestudenten. Het oordeel over de eigen schrijfvaardigheid kan daardoor bruikbare informatie geven over de schrijfvaardigheid. In deelstudie 3 (paragraaf 6.3) naar de effectiviteit van het TAVAN-programma werd een correlatie van -0.67 gevonden tussen het oordeel over de eigen schrijfvaardigheid en het aantal gestandaardiseerde fouten PHW in de begintekst. Deze correlatie laat zien dat de inschatting van de eigen schrijfvaardigheid soms valide informatie over het schrijfvaardigheidsniveau zal leveren, hoewel de docent hier moeilijk een oordeel voldoende/onvoldoende op kan baseren. In deze deelstudie naar de effectiviteit van TAVAN bleek echter ook dat de groep overige studenten positiever was gaan denken over de eigen schrijfvaardigheid, terwijl het traditionele programma bij de controlegroep niet tot daadwerkelijke verbetering in de schrijfvaardigheid had geleid. De TAVAN-groep daarentegen verbeterde wel aantoonbaar, maar ging zichzelf gemiddeld genomen niet positiever inschatten. Het oordeel van studenten over de eigen schrijfvaardigheid is daarmee niet goed bruikbaar (niet valide) om studenten en programma's te evalueren.

De linear is een tekst met fouten en gebreken die door de student gecorrigeerd moet worden op de tussenliggende, blanco regels en die vervolgens door beoordelaars wordt nagekeken. Godshalk et al. (1966) vonden dat de linears vrijwel even hoog correleerden met de totale essayscore als de qua criteriumvaliditeit best presterende objectieve schrijfvaardigheidstests. Qua opzet lijken de linears te vergelijken met de begin- en eindtekst die gebruikt werden in deelstudie 3 naar het effect van het TAVAN-programma (zie 6.2). In dat geval moest een tekst met fouten en gebreken door de student op de computer bewerkt worden tot een

correcte tekst. Hoewel een goede linear qua criteriumvaliditeit vergelijkbaar bleek te zijn met een objectieve test, is het voordeel van een linear boven een objectieve test dat de student daadwerkelijk schrijft, althans herschrijft. Het voordeel van een linear boven een essaytest is dat de inhoud gegeven is en in beginsel vastligt. De student hoeft geen literatuur door te werken, aantekeningen te ordenen, maar kan zich volledig richten op het probleem van het zo goed mogelijk verwoorden. Eventuele verschillen op inhoudelijk gebied doordat de ene student veel meer weet van een onderwerp dan de andere, worden daardoor beperkt. Dit voordeel vormt echter mogelijk ook een nadeel. Een essay-opdracht stelt een student voor een complexere taak.

Verbale intelligentietests blijken vrijwel even goed holistisch beoordeelde schrijfvaardigheid te voorspellen als objectieve tests, speciaal ontwikkeld om schrijfvaardigheid te meten. Zo vermelden Breland en Jones (1982, p. 13, Table 6) een correlatie tussen de SAT-verbal en de ECT holistische beoordeling van 0.56. Het objectieve schrijfvaardigheidsdeel van de ECT correleerde slechts marginaal hoger met 0.58. De betrouwbaarheid van het ECT holistisch oordeel was 0.58. Verbale intelligentietests lijken daarmee in de praktijk (vrijwel) even criteriumvalide te kunnen zijn als objectieve tests die speciaal ontwikkeld werden om schrijfvaardigheid te meten en ook dezelfde FOC-factor te meten.

Interessant in dit verband is dat van de twee delen waaruit de SAT-verbal was opgebouwd, het leesvaardigheidsdeel marginaal hoger correleerde met het holistische oordeel dan het schrijfvaardigheidsdeel. Daar in de prakijk de twee delen van een objectieve test doorgaans min of meer dezelfde betrouwbaarheid bezitten, lijkt dit te betekenen dat leesvaardigheid op zijn minst net zo belangrijk is voor schrijfvaardigheid als schrijfvaardigheid. Dit is een vreemde conclusie. De juiste conclusie moet kennelijk zijn dat leesvaardigheid en schrijfvaardigheid elkaar vrijwel volledig overlappen. Dit wordt ook bevestigd door de door Breland en Jones (1982, p. 13, Table 6) vermelde gegevens. Hoewel de SAT-reading en de SAT-writing hoog positief correleren met het holistische oordeel (respectievelijk 0.54 en 0.52) is de correlatie van beide tests gezamenlijk (SAT-verbal) slechts marginaal hoger (0.56). Beide tests meten dus in hoge mate dezelfde factor. Het strikte onderscheid dat op inhoudelijke gronden gemaakt wordt tussen leesvaardigheid en schrijfvaardigheid blijkt empirisch gezien niet bevestigd te worden door duidelijk verschillende uitkomsten.

Voor de als tiende vermelde methode, de (zeldzaamheid van de) in een essay gebruikte woorden, konden wij geen empirische bevestiging vinden. Algemeen wordt verondersteld

dat betere schrijvers een grotere woordenschat hebben. Tests die bedoeld zijn de omvang van de woordenschat te meten, zijn echter niet gebaseerd op de woorden die daadwerkelijk in het essay gebruikt zijn. De vermoedelijke reden dat er moeilijk onderzoek te vinden is dat de frequentie van de daadwerkelijk gebruikte woorden koppelt aan bijvoorbeeld het holistische oordeel, is dat dit onderzoek zonder speciale programmatuur en zonder een database met woordfrequenties lastig uit te voeren is. Voor ieder essay moet immers in beginsel van ieder woord de frequentie vastgesteld worden. Essays moeten dus digitaal beschikbaar zijn of gemaakt worden alsmede de woordfrequenties. Vervolgens is speciale programmatuur vereist om per essay een index te berekenen voor het gebruik van zeldzame woorden. Verder doet zich mogelijk een complicatie voor. Zwakke studenten willen nog wel eens moeilijke termen gebruiken op plaatsen waar die niet terecht zijn. In hoeverre deze tiende maat dus inderdaad valide is, zal nog onderzocht moeten worden.

In Tabel 8.3 is 'analytisch beoordelen' van teksten niet opgenomen als afzonderlijke methode om basale schrijfvaardigheid te beoordelen. Een eerste reden is dat in de praktijk de beoordelaar vaak eerst voor zichzelf een holistisch oordeel zal formuleren en vervolgens vanuit dit holistische oordeel de verschillende schalen voor de analytische beoordeling zal invullen. De meerwaarde van een analytische beoordeling is daarmee niet aangetoond. Een tweede probleem is dat door de beoordeling aan regels en voorschriften te binden, niet langer zeker is dat de beoordeling valide is. Meer specifiek: per analytisch beoordelingsschema zal men minimaal moeten aantonen dat de beoordeling redelijk hoog correleert met bijvoorbeeld een vrij holistisch oordeel. Dat een beoordelingsformulier er indrukwekkend uitziet, geeft wel 'face validity', maar geen garantie dat het resulterende oordeel ook daadwerkelijk hoog correleert met andere criteria om schrijfvaardigheid te meten.

In Tabel 8.3 valt te zien dat er vier indirecte methodes zijn, waarbij de student niet hoeft te schrijven. Bij twee van deze vier methodes (2 en 9) moeten objectieve testitems beantwoord worden. Ook de KHO-methode (5) kan opgevat worden als een vorm van een objectieve test. Van deze methodes blijken de specifiek op correct taalgebruik gerichte objectieve tests (2) uitzonderlijk goed in staat het holistische oordeel te voorspellen. Ook voor de KHO-methode rapporteerden we hiervoor (zie 9.7) een zeer hoge voor onbetrouwbaarheid gecorrigeerde correlatie met de overige schrijfvaardigheidsmaten. Van de indirecte methodes betreft verder één de subjectieve inschatting van de eigen schrijfvaardigheid door de student zelf.

Van de zes directe methodes (methodes waarbij geschreven moet worden) zijn drie relatief simpel en eenvoudig toe te passen: het holistische oordeel, beoordeling op aantal fouten per honderd woorden en tekstlengte. Tekstlengte is een simpele maat die vaak goed blijkt te correleren met het holistische oordeel. Beoordeling op fouten levert specifieke informatie voor de student, maar is didactisch gezien mogelijk niet altijd even effectief en kan mogelijk zelfs averechts werken. Het holistische oordeel levert weinig specifieke informatie en blijkt hoog te correleren met aantal fouten en tekstlengte. Het voordeel van een holistische oordeel is dat het relatief snel kan worden toegekend, dat het rekening houdt met tekstlengte, fouten en de inhoud van de tekst, terwijl het niet de didactische nadelen heeft van beoordeling op fouten per honderd woorden.

Van de drie overblijvende directe methodes (linears, TAVAN-score, gebruikte woorden) gaan de linears uit van een testsituatie. Dit is echter op zich niet echt noodzakelijk. Men zou ook foute zinnen kunnen geven in een oefensituatie. Het probleem is dan echter het nakijken, wat ook bij gebruik als test een probleem vormt. Linears zijn dus arbeidsintensief, waarbij in een onderwijssituatie de feedback te laat zal komen om heel effectief te zijn.

Door gebruik te maken van de computer en TAVAN vervallen deze problemen, doordat TAVAN de verbeterde zin onmiddellijk checkt. De feedback is hierdoor snel en gericht. Verder houdt TAVAN de score bij, zodat tegelijkertijd ook informatie wordt verkregen over het schrijfniveau van de student.

De laatste directe methode (4, gebruikte woorden/woordkeuze) lijkt in beginsel bij via de computer geschreven teksten relatief eenvoudig toe te passen via een computerroutine. Studenten die beter schrijven, hebben vaak ook een grotere woordenschat die terugkomt in de tekst. Op dit moment is echter nog onduidelijk wat voor validiteitscoëfficiënten dit oplevert: wat is bijvoorbeeld de correlatie met het holistische oordeel? Verder onderzoek op dit punt is geboden.

8.10 Geautomatiseerde holistische beoordeling

Het meten van schrijfvaardigheid is duur, doorgaans onbetrouwbaar en leverde tot nu toe weinig op. Een methode zoals TAVAN lost dat probleem op door aan de ene kant de beoor-

delingstaak te vereenvoudigen en aan de andere kant de beoordeling te automatiseren. Voor een methode gebaseerd op de kwaliteit van het holistische oordeel (KHO) zal hetzelfde gelden, omdat in dat geval alleen aan de student gevraagd wordt de beste van twee zinnen of twee passages te kiezen. Is het echter niet mogelijk nog een stap verder te gaan en de computer in te zetten als holistische beoordelaar? Moderne computersystemen kunnen complexe taken aan als schaken, autorijden en jeopardy spelen. Het lijkt daarmee aannemelijk dat computers op een bepaald moment in staat zullen zijn teksten van studenten op kwaliteit te beoordelen. In feite is dat wat TAVAN inmiddels al doet met herschreven zinnen en herschreven korte passages.

Strikt genomen blijken computers echter ook al langere teksten te kunnen beoordelen. De lengte van een tekst is namelijk vaak een betrouwbare voorspeller van de kwaliteit, die het even goed doet (qua hoogte van de correlatie met het criterium) als een goede menselijke beoordelaar of zelfs beter. In werkelijkheid gebruikt men dan echter tekstlengte om de kwaliteit te beoordelen en voegt de computer op deze manier geen wezenlijke informatie aan de meting toe. De voor de beoordelaar al aanwezige informatie wordt gekwantificeerd en gepresenteerd als het oordeel van een digitaal expertsysteem.

Veronderstel dat we het tekstlengte-probleem zouden kunnen omzeilen zodat de computer op een manier vergelijkbaar met een menselijke beoordelaar naar de tekst zou kijken. Het zou dan mogelijk zijn de computer in te zetten als holistische beoordelaar. Voor selectiedoeleinden heeft dit weinig nut, omdat objectieve tests voor dat doel uitstekend werken. Voor onderwijs- en toetsdoeleinden lijkt dit misschien mogelijkheden te bieden.

Een eerste vraag is in dit verband, hoe men zich precies de onderwijs- en toetssituatie voorstelt. Onderwijs- en toetssituatie moeten vergelijkbaar zijn. Wanneer studenten de opdracht krijgen een essay over een bepaald onderwerp te schrijven, zou de geschreven tekst door het computerprogramma kunnen worden beoordeeld. In de toetssituatie zou dan een vergelijkbare opdracht gegeven moeten worden, die vervolgens ook weer door hetzelfde computerprogramma zou moeten worden beoordeeld. Vaak legt men echter voor de toetsing strengere kwaliteitsnormen aan dan voor het onderwijs (Williamson et al., 2010, p. 7). In de praktijk gebeurt daardoor eerder het omgekeerde: door het gebruik van de geautomatiseerde tekstbeoordeling ontstaan verschillen tussen de onderwijs- en de toetssituatie .

Een tweede probleem is de betrouwbaarheid. Wanneer het computerprogramma even slecht beoordeelt als een normale menselijke beoordelaar, bevat het oordeel zoveel ruis dat het

voor studenten amper bruikbare informatie bevat. Om echt bruikbaar te zijn, zou het programma daarom ten minste het werk van een aantal expert-beoordelaars moeten kunnen overnemen zodat het veel betrouwbaarder zou zijn. Het moet dus niet evenveel kunnen als een menselijke beoordelaar, maar belangrijk meer.

Een derde probleem heeft te maken met de doelstelling. Wat probeert men de studenten precies te leren? De bedoeling lijkt duidelijk: de studenten moeten beter leren schrijven. Maar men beoordeelt de geschreven tekst via (geautomatiseerde) holistische beoordeling. Na correctie voor onbetrouwbaarheid bleek in het voorgaande dat het holistische oordeel vrijwel volledig samenvalt met de FOC-factor. In feite is de oefensituatie daarmee volledig gericht op een betere beheersing van de FOC-factor oftewel basale schrijfvaardigheid.

Voor dat doel beschikken we echter inmiddels over een effectieve trainingsopzet, namelijk TAVAN. De vraag is dan of oefenen in het schrijven van een (lang) essay dat pas daarna beoordeeld wordt op kwaliteit, een efficiënte opzet is in vergelijking met TAVAN (of eventueel een programma met KHO-items). Voorlopig is er, voor zover bekend, geen empirische evidentie waaruit blijkt dat oefenen met het schrijven van essays studenten op de FOC-factor sneller vooruit doet gaan dan oefenen met TAVAN.

Een begrijpelijke argument tegen het gebruik van TAVAN in dit geval is dat het niet de bedoeling is studenten te trainen op de FOC-factor, maar dat het juist de bedoeling is studenten te leren een samenhangend en logisch betoog op te zetten. Dat lijkt een prima doelstelling, maar het holistische oordeel valt vrijwel volledig samen met de FOC-factor en bevat daarmee geen inhoudelijke component (of alleen een inhoudelijke component die volledig samenvalt met de FOC-factor). Uitgaande van deze doelstelling moet men dan niet (geautomatiseerd) holistisch beoordelen, maar inhoudelijk. Of dat (betrouwbaar) mogelijk is en hoe dat zou moeten, is op dit moment niet duidelijk.

Een derde probleem is de verwerking van de feedback. Wanneer studenten alleen een kort holistisch oordeel over de kwaliteit van hun tekst krijgen, krijgen ze in feite te horen of ze met relatief veel of met relatief weinig fouten per honderd woorden schrijven. Het holistische oordeel bleek immers na correctie voor onbetrouwbaarheid vrijwel perfect te correleren met het aantal fouten per honderd woorden. Om die boodschap te communiceren is de TAVAN-score echter veel betrouwbaarder, sneller en gemakkelijker vast te stellen en voor studenten ook nog instructiever.

Wanneer de feedback heel gedetailleerd wordt gerapporteerd, ontstaat een ander probleem. Studenten die in TAVAN langere stukken tekst aangeboden kregen met meerdere fouten, bleken niet meer in staat de informatie over hoe het idealiter wel had gemoeten, te verwerken.

Men moet zich op dit punt realiseren dat een student na een fout antwoord wordt geconfronteerd met drie stukken tekst: de foute zin (A), het gegeven antwoord (B) en het goede antwoord (C). In totaal kunnen er daardoor drie verschillende vergelijkingen gemaakt worden om de aangeboden informatie te analyseren: A-B, A-C, B-C. Wanneer in een zin één fout verbeterd moet worden, valt met enig studeren nog wel na te gaan, wat men verkeerd deed. Bij twee fouten per zin wordt dit al aanmerkelijk lastiger. Bij vijf fouten per alinea zijn er 15 verschillende vergelijkingen mogelijk. In de praktijk blijkt dit voor studenten niet meer goed te doen.

Interessant in dit verband is wellicht dat het nakijken en beoordelen van langere antwoorden voor het TAVAN-programma geen problemen blijkt te leveren. De voor dit doel ontwikkelde software-routine en de hardware kunnen dit probleemloos aan, maar voor de studenten ligt dit belangrijk anders.

De vraag is daarmee: wat levert geautomatiseerd holistisch beoordelen nu precies op? Welke nadelen van holistisch beoordelen neemt het weg en welke voordelen voegt het idealiter toe? In Tabel 8.2 zijn vier criteria gebruikt om meetmethodes voor het vaststellen van basale schrijfvaardigheid te beoordelen: geschiktheid als trainingsmethode, tijd en kosten, validiteit, betrouwbaarheid. Holistische beoordeling scoort op deze vier criteria respectievelijk: matig, slecht, matig, matig. Geautomatiseerde holistische beoordeling zou dan naar verwachting als volgt scoren: matig, goed, matig, matig. De hoeveelheid tijd en kosten verbeteren, maar verder blijven alle bezwaren van holistische beoordeling gelden.

Stel nu dat men alle problemen zou overwinnen zodat men de student een tekst zou kunnen laten schrijven en de computer zou feilloos het exacte aantal fouten per honderd woorden (een andere en preciezere omschrijving van het holistische oordeel) bepalen. (In feite lijkt dit sterk op de al eerder beschreven situatie die met TAVAN is uitgeprobeerd met een tekst bestaande uit een aantal zinnen. De schrijfopdracht gaat dan echter uit van herschrijven.) De validiteit en de betrouwbaarheid zouden in dat geval perfect zijn, de kosten zouden prima zijn, maar ook in dat geval blijft de geschiktheid als trainingsmiddel matig, doordat de feedback te laat komt en, of te simpel is, of te complex is om bruikbaar te zijn.

Ondertussen is het nog lang niet zo ver, terwijl er al sinds 1966 aan dit soort programma's gewerkt wordt, toen Page op verzoek van de College Board een eerste versie van Project Essay Grader (PEG) ontwikkelde (Dikli, 2006). Een programma als PEG blijkt als holistisch beoordelaar heel goed te werken, maar baseert zich in feite vooral op de tekstlengte (p. 25).

Dikli (2006) merkt in een overzicht van zeven programma's op dit gebied op: "The results of several AES [Automated Essay Scoring] studies reported high agreement rates between AES systems and human raters" (p. 4). Het lijkt goed mogelijk dat dit klopt, maar doordat de firma's die deze programma's op de markt brengen, doorgaans ook de bron van deze resultaten zijn, is de grens tussen reclame en onderzoek in dit geval niet altijd duidelijk.

Onduidelijk is ook wat Dikli precies bedoelt met 'high agreement rates' in dit verband, terwijl de hoogst mogelijke correlatie sterk beperkt wordt door de lage betrouwbaarheid van menselijke holistische beoordelaars. Een punt dat hij zich niet lijkt te realiseren. Hij rapporteert 'agreement rates' tussen menselijke beoordelaars van 0.94, 0.96 en 0.97 (p. 27) die ontleend zijn aan onderzoek van anderen. In de literatuurlijst staat de desbetreffende bron alleen vermeld met een link naar een pdf-document, die niet (meer) blijkt te werken. Ook Google blijkt niet in staat de bewuste pdf te lokaliseren. Bij verder zoeken blijken de hoge waarden afkomstig uit onderzoek uitgevoerd door het bedrijf, Vantage Learning, dat het desbetreffende programma commercieel exploiteert. Ten slotte blijkt de gebruikte 'agreement rate' bij een korte beoordelingsschaal (goed/slecht) altijd te resulteren in een overeenstemmingspercentage van honderd procent, ongeacht de werkelijke overeenstemming tussen de beoordelaars.

Gebruikelijk bij software-reviews is dat de auteur of de testende instantie zelf de programma's uitprobeert en onderwerpt aan een reeks praktijktests. In dit geval is daarvan bij Dikli geen sprake, waardoor ook niet duidelijk wordt wat het best werkende programma is en hoe goed dat in de praktijk werkt.

Hoewel de door hem gerapporteerde agreement rates vrijwel perfect werkende programma's suggereren, stelt hij ten slotte: "An effective way of using AES technology to score essays is to incorporate the AES system into the writing evaluation process as a second or third rater" (Dikli, 2006, p. 27-28). Kennelijk is het nog niet verstandig deze programma's als zelfstandige beoordelaar te gebruiken. Maar dat lijkt te betekenen dat ze het in de praktijk nog

slechter doen, dan een menselijk beoordelaar. Een uitkomst die gezien de beschikbaarheid van de tekstlengte en de doorgaans hoge correlatie daarvan met het holistische oordeel, moeilijk te begrijpen valt.

Williamson et al. (2010) geven vanuit de College Board, de ETS (Educational Testing Service) en Pearson (zelf een belangrijke uitgever op dit gebied) een overzicht van de punten die van belang zijn bij de aanschaf van software voor geautomatiseerde holistische beoordeling. Zij merken op: "There is much that is not yet known about the performance of these systems" (p. 5).

Een eerste eis die Williamson et al. stellen, is dat de software soortgelijke resultaten moet opleveren als menselijke beoordelaars en dat dit uitgebreid gecontroleerd moet worden. Als maat voor de overeenstemming tussen de machine en de menselijke beoordelaars moet de correlatie berekend worden of een gewogen kappa, maar niet een overeenstemmingspercentage, omdat dit tot zeer misleidende resultaten kan leiden.

Een tweede eis is dat de manier waarop het programma tot een bepaalde score komt, helder en duidelijk moet zijn. Een eis die gezien de beschikbaarheid van tekstlengte zeer begrijpelijk lijkt.

Een derde eis is dat onderzocht moet worden of het programma niet (onbedoeld) verschillende bevolkingsgroepen verschillend beoordeelt. Speciaal in de VS is de angst groot dat een beoordelingssysteem door een ingebouwde culturele bias onbedoeld achterstandsgroepen op nog grotere afstand zet.

Een vierde eis is dat de validiteit van de automatische scoring eerst onderzocht en aangetoond moet worden door de correlatie te berekenen met een criteriumvariabele. Een vijfde eis is dat er geen systematische verschillen mogen zijn in de hoogte van de door het programma toegekende scores en de scores van menselijke beoordelaars.

Voor de toepassing van een beoordelingsprogramma zien ze twee opties. De eerste is wanneer de beoordeling er niet echt toe doet. De tweede mogelijkheid is het programma te gebruiken in combinatie met een menselijke beoordelaar (p. 7).

Een dergelijke selectieve inzet leidt ons inziens echter gemakkelijk tot ongewenste verschillen tussen de trainingssituatie en de testsituatie. Ook de eerste optie lijkt wat merkwaardig:

als de beoordeling er niet echt toe doet, waarom beoordeelt men dan? Het idee is kennelijk dat het goed is om studenten te laten schrijven en dat men aan de beoordeling van de tekst niet al te zwaar moet tillen. In dat geval lijkt een beoordeling op tekstlengte echter veel duidelijker en transparanter. Te meer omdat het aantal fouten per honderd woorden waarmee studenten schrijven betrouwbaar en eenvoudig vooraf te meten valt met TAVAN en in beginsel ook met TAVAN valt te remediëren.

Een bezwaar van geautomatiseerde holistische beoordeling dat nog niet aan de orde is geweest, is dat de programma's eerst getraind moet worden op een collectie soortgelijke essays die al beoordeeld zijn. De benodigde aantallen beginnen soms bij 100, maar meer gebruikelijk is dat enkele honderden tot duizend (liefst dubbel) beoordeelde essays benodigd zijn ((Dikli, 2006). De algemene inzetbaarheid van dit soort programma's wordt daardoor in feite beperkt tot zeer grootschalige toepassingen.

Voorlopig lijken de praktische mogelijkheden van geautomatiseerde holistische beoordeling daarmee nog zeer beperkt. Het lijkt echter ook duidelijk dat wanneer een programma inderdaad in staat zou zijn het aantal fouten per honderd woorden in teksten betrouwbaar vast te stellen, dit een belangrijke en interessante ontwikkeling zou zijn waarvan de toepassingsmogelijkheden op dit moment nog moeilijk te overzien zijn.

8.11 Samenvatting, conclusies en discussie

In dit proefschrift staan twee samenhangende problemen centraal: het meten en het optimaliseren van basale schrijfvaardigheid. Zolang we basale schrijfvaardigheid niet goed kunnen meten, is het ook niet mogelijk onderwijsprogramma's te onderzoeken op hun effectiviteit in het verhogen van basale schrijfvaardigheid. Omgekeerd heeft het meten van basale schrijfvaardigheid weinig zin zo lang we niet beschikken over methodes om basale schrijfvaardigheid gericht te vergroten. Voor onderwijssituaties geldt daarmee de eis: een goede meetmethode moet ook een goede trainingsmethode zijn. Meten en trainen moeten in onderwijssituaties samenvallen. Dit uitgangspunt vormde de reden om te zoeken naar betere en eventueel andere methodes om basale schrijfvaardigheid te meten.

Een vreemde en foute relatie

Wanneer we ons verdiepen in de geschiedenis van het meten van (basale) schrijfvaardigheid, zien we dat dit begint met het holistisch beoordelen van door studenten geschreven essays in reactie op een onderwerp of vraag (het topic). Wanneer je studenten beter wilt leren schrijven, zul je ze moeten laten schrijven. Om te zorgen dat ze inderdaad schrijven, zul je moeten beoordelen. Verder kost het het beoordelen van het werk van twee groepen studenten (zeg in totaal zestig studenten), ook al doet men dit zeer snel, toch al snel enkele uren. Holistische beoordeling is in deze situatie de enige mogelijke optie en ook zinvol, omdat het doel - studenten te laten schrijven - bereikt wordt.

Bij toelatingsexamens is het doel echter niet studenten iets te leren, maar de beste studenten zo eerlijk en zo goed mogelijk te selecteren. In eerste instantie werd ook voor dit doel holistische beoordeling gebruikt. Holistische beoordeling blijkt dan echter hinderlijk onbetrouwbaar. In reactie hierop werden ten slotte objectieve tests (bestaande uit multiple-choice items) ingevoerd om de schrijfvaardigheid te meten. Deze objectieve tests bleken betrouwbaar te meten. In eerste instantie werd verondersteld dat basale schrijfvaardigheid niet valide gemeten kan worden via multiple-choice tests, omdat de student in dat geval niet echt schrijft, maar Godshalk et al. (1966) en daarna ook anderen, lieten zien dat de uitkomsten van objectieve tests hoog correleren met de resultaten van holistische beoordeling. Objectieve tests bleken criteriumvalide te zijn.

Dit resultaat was voor docenten en veel anderen onverwacht, vreemd en fout. Hoe was het mogelijk dat een test waarbij een student niet hoefde te schrijven in staat was het holistische oordeel van 'expert-beoordelaars' (in de praktijk: docenten) over een door de student geschreven tekst te voorspellen? Behalve dat deze uitkomst onverwacht was, was deze uitkomst ook niet wat docenten graag wilden horen. Holistische beoordeling kost veel tijd en veronderstelt dat het oordeel van de docent niet ter discussie staat. Nu bleek niet alleen dat het docentenoordeel uitermate onbetrouwbaar te zijn (wat al langer bekend was), maar ook nog eens voorspeld te kunnen worden door iets simpels als een meerkeuze-toets.

Na de publicatie van Godshalk et al. (1966) bestonden er daarmee twee verschillende methodes om (basale) schrijfvaardigheid te meten. Beide methodes verschillenden sterk qua inhoud (procedure), maar leverden soortgelijke uitkomsten. De ene methode was ontwikkeld in de onderwijspraktijk, de andere methode was ontwikkeld in de examen- en selectie-

praktijk. Beiden methodes waren vooral geschikt voor hun eigen specifieke situatie en hadden hun eigen aanhangers. De docenten waren overtuigd van de validiteit van hun holistisch oordeel, de testconstructeurs waren overtuigd van de validiteit van hun objectieve tests.

In een volstrekt rationele wereld zou men misschien daarna het verband verder gekwantificeerd hebben en zou men zich afgevraagd hebben, hoe het mogelijk was dat twee inhoudelijk totaal verschillende maten qua resultaten soortgelijke uitkomsten leverden. In plaats daarvan ontstonden er uitgebreide discussies over betrouwbaarheid en validiteit, waarbij iedere partij het begrip validiteit op zijn eigen wijze invulde. In plaats van te zoeken naar een verklaring, polariseerde men of maakte men vooral omtrekkende bewegingen.

Verklaring van het verband: drie maal meten van de FOC-factor

Wanneer we wel proberen het verband tussen objectieve tests en holistische beoordeling te verklaren, vallen een aantal zaken op. Allereerst blijken de objectieve tests in hoge mate één factor te meten. Ten tweede blijkt het verband tussen deze objectieve testfactor en het holistische oordeel na correctie voor onbetrouwbaarheid zeer hoog (0.87 of hoger) te zijn, zodat het holistische oordeel en de objectieve testfactor vrijwel volledig overlappen. Ten derde blijkt de 'objectieve test' factor ook 'direct' gemeten te kunnen worden via een 'interlinear'. Dit is een test waarbij de student foute zinnen moet corrigeren. De objectieve test factor kan daarmee niet alleen 'indirect' gemeten worden via meerkeuze-items, maar ook 'direct' door de student te laten (her)schrijven. Bij een analyse van de inhoud van de tests bleek vervolgens een vierde punt. De items van alle drie betrokken tests hadden betrekking op de FOC-factor (Fouten Opsporen en Corrigeren). Dezelfde vaardigheid die ook centraal stond in het TAVAN-programmma en daar gemeten werd via de TAVAN-score (het matchingspercentage).

De verklaring voor het door Godshalk et al. (1966) gevonden verband tussen objectieve tests om schrijfvaardigheid te meten en het holistische oordeel is dan als volgt. 1. Via de objectieve tests wordt rechtstreeks bij de student de FOC-factor gemeten. 2. Het FOC-niveau van de student komt via het aantal fouten PHW (per honderd woorden) tot uiting in de door de student geschreven tekst. 3. De holistische beoordelaars laten zich door het aantal fouten PHW in de tekst (en mogelijk ook door andere gecorreleerde factoren als tekstlengte

en kwaliteit van de inhoud) sterk beïnvloeden. Het aantal fouten PHW in de tekst verzorgt daarmee de koppeling tussen de objectieve testscore en het holistische oordeel.

Wanneer deze verklaring juist is, moet het aantal fouten PHW in de tekst hoog correleren met enerzijds de objectieve testscore en anderzijds het holistische oordeel. Door de gegevens verkregen bij het fouteneffect-onderzoek te combineren met de gegevens van TAVAN2 (de tweede keer dat het TAVAN-programma uitgevoerd werd) was het mogelijk na te gaan of deze verklaring voor het verband tussen het holistische oordeel en de objectieve testscore bevestigd werd. Als objectieve testscore werd daarbij de TAVAN-score gebruikt. De correlatie van het aantal fouten PHW in de 48 teksten met de TAVAN-score bedroeg 0.69 en met het holistische oordeel 0.90 (zie paragraaf 8.7). Beide correlaties zijn daarmee hoog. Het voorgestelde verklaringsmodel wordt bevestigd.

Men kan ook redeneren dat alle drie betrokken variabelen in feite vooral dezelfde onderliggende FOC-factor meten. De gemiddelde onderlinge correlatie tussen de drie variabelen zou in dat geval hoog moeten zijn. De gemiddelde voor onbetrouwbaarheid gecorrigeerde correlatie tussen de drie variabelen (holistische oordeel, aantal fouten PHW, TAVAN-score) bedroeg 0.86 (zie paragraaf 8.7). Deze hoge waarde bevestigt dat de drie variabelen inderdaad in hoge mate dezelfde factor meten.

Het is gebruikelijk om objectieve tests voor het meten van schrijfvaardigheid aan te duiden als 'indirecte' maten, doordat ze niet uitgaan van door de student geschreven tekst. Holistische beoordeling wordt daarentegen aangeduid als een 'directe' maat, omdat nu wel wordt uitgegaan van door de student geschreven tekst. In werkelijkheid is er echter eerst het schrijfvaardigheidsniveau van de student, vervolgens schrijft de student de tekst waarbij fouten ontstaan zonder gecorrigeerd te worden en pas in laatste instantie beoordeelt de holistische beoordelaar de tekst, waardoor het aantal fouten PHW uiteindelijk een belangrijke rol gaat spelen in het holistische oordeel.

Men moet hierbij bedenken dat het holistische oordeel normaal het totaal is van een aantal beoordelaars. Het gevolg is dat alleen de factoren die bij alle beoordelaars hetzelfde werken (waar alle beoordelaars het over eens zijn), uiteindelijk overblijven. Wanneer iedere individuele beoordelaar ook kijkt naar iets als de 'communicatieve waarde' zal dat in het uiteindelijke totaal wegvallen wanneer de ene beoordelaar die waarde hoog inschat en de volgende laag.

Het geschetste model verklaart ook waarom holistische beoordeling in de praktijk relatief onbetrouwbaar is. De holistische beoordelaar beschikt in ieder geval over drie verschillende variabelen om zijn oordeel op te baseren en bepaalt zelf hoe hij deze variabelen weegt. Tekstlengte is een vrij objectieve variabele, het aantal fouten PHW blijkt dat in beginsel ook te zijn. Het oordeel over de kwaliteit van de inhoud zal echter vaak persoonsgebonden zijn, waardoor verschillende beoordelaars op dit punt niet overeenstemmen.

Een praktische consequentie van het gevonden resultaat is dat basale schrijfvaardigheid op drie totaal verschillende manieren vastgesteld kan worden die alle drie criteriumvalide zijn: via holistische beoordeling van de door de student geschreven tekst, via het aantal fouten PHW in de geschreven tekst en via de TAVAN-score (of eventueel een andere objectieve test bedoeld de FOC-factor te meten). Hoewel deze drie maten qua inhoud en procedure mogelijk zeer verschillend lijken, wordt in werkelijkheid steeds vooral de FOC-factor gemeten.

FOC-factor ook bepalend voor kwaliteit holistische oordeel

In het fouteneffect-onderzoek werd ook nog een vierde variabele gebruikt: de kwaliteit van het holistische oordeel uitgebracht door de student over een zestal teksten van medestudenten. Naarmate een student het ideale holistische oordeel dichter benaderde, scoorde hij beter. Deze maat bleek na correctie voor onbetrouwbaarheid gemiddeld 0.88 (zie paragraaf 8.7) te correleren met de drie overige maten om de FOC-factor vast te stellen en daarmee ook in zeer hoge mate de FOC-factor te meten. De consequentie van dit resultaat is dat hiermee een derde nieuwe methode (na de TAVAN-score en het aantal fouten PHW) is aangetoond die gebruikt kan worden om basale schrijfvaardigheid criteriumvalide te meten.

De verklaring van dit verband ligt voor de hand. Het holistische oordeel blijkt uiteindelijk vooral tot stand te komen via het aantal fouten PHW. Naarmate een student beter is in het signaleren van de fouten in de te beoordelen teksten, wordt zijn holistische beoordeling beter. Ook bij de kwaliteit van holistische beoordeling speelt de FOC-factor daarmee een doorslaggevende rol.

Constructvaliditeit, trainingsmethodes en aantal overige meetmethodes

In totaal zijn er daarmee vijf verschillende manieren bekend die alle vijf soortgelijke uitkomsten leveren (criteriumvalide zijn) om basale schrijfvaardigheid te meten: holistische oordeel, objectieve test, TAVAN-score, aantal fouten PHW en kwaliteit holistische oordeel. Basale schrijfvaardigheid is daarmee een 'construct' geworden dat op een groot aantal verschillende manieren gemeten kan worden.

Wil een meetmethode in een onderwijssituatie bruikbaar zijn, dan moet de methode ook bruikbaar zijn als trainingsmethode, werd eerder gesteld. Holistische beoordeling als trainingsmethode is niet erg effectief, omdat de feedback te laat komt en weinig specifiek is. Objectieve meerkeuze-tests lijken niet effectief, omdat de student niet daadwerkelijk schrijft. Voor de TAVAN-score gelden deze bezwaren niet. De student (her)schrijft daadwerkelijk, terwijl de feedback onmiddellijk komt en specifiek is. De TAVAN-score vormt daarmee de eerste schrijfvaardigheidsmaat die ook goed bruikbaar is als trainingsmethode.

Verder lijkt de maat 'kwaliteit van het uitgebrachte holistische oordeel' met enige aanpassingen als basis te kunnen dienen voor een trainingsmethode. Het is immers mogelijk het aantal te beoordelen teksten per item te reduceren tot twee en te werken met zeer korte vergelijkbare tekstfragmenten. De opdracht voor de student is dan het beste alternatief te kiezen. Onder deze omstandigheden zal de student naar verwachting (snel) leren het beste alternatief qua stijl en taal te selecteren. Hij leert te discrimineren tussen slecht taalgebruik en goed taalgebruik. Op dit moment is deze optie nog niet daadwerkelijk gerealiseerd en uitgeprobeerd, maar is dit een punt voor verder onderzoek.

Het theoretisch interessante van deze optie is dat de student receptief schrijven zou leren, dat wil zeggen: zonder daadwerkelijk te schrijven. Het praktische belang van deze optie is dat er naast de TAVAN-score een tweede methode beschikbaar zou komen om basale schrijfvaardigheid gericht te trainen en te verhogen. Verder zou deze methode door de eenvoud van de items gebruikt kunnen worden om basale schrijfvaardigheid 'indirect' op een efficiënte wijze vast te stellen.

Wanneer er vijf verschillende methodes bestaan om basale schrijfvaardigheid vast te stellen, bestaan er dan mogelijk niet meer? In totaal konden we nog vijf andere methodes vinden: de lengte van de geschreven tekst, de inschatting van de eigen schrijfvaardigheid, het herschrijven van een tekst vol fouten (de 'interlinear' kan men zien als een gestructureerde

vorm hiervan), verbale intelligentietests en de in de tekst gebruikte woorden. Dit laat zien dat er in beginsel een groot aantal manieren bestaat om schrijfvaardigheid vast te stellen. Voor de onderwijspraktijk lijkt vooral de lengte van de geschreven tekst mogelijk een bruikbare variabele. Het herschrijven van een tekst vol fouten is vermoedelijk een goede oefening voor de student, maar is lastig na te kijken. Interlinears vallen door hun gestructureerdere vorm beter na te kijken, maar qua effectiviteit en nakijken lijkt een oplossing als TAVAN dan te prefereren.

Bestaat er een tweede factor?

Op basis van de vier gebruikte maten voor het vaststellen van basale schrijfvaardigheid (holistische oordeel, TAVAN-score, aantal fouten PHW en kwaliteit van het uitgebrachte holistische oordeel) bleek het niet mogelijk het bestaan van een tweede factor (anders dan de FOC-factor) aan te tonen. Doordat de objectieve testscore (de FOC-factor) het holistische oordeel wel grotendeels, maar niet volledig overlapt (het percentage gemeenschappelijk variantie is na correctie voor onbetrouwbaarheid ruim 75%), wordt doorgaans verondersteld dat er een tweede factor zou moeten bestaan voor de verklaring van de overige 25% of een deel daarvan.

Uit het gegeven dat het in dit geval niet lukte het bestaan van een niet-FOC factor aan te tonen, mag echter niet afgeleid worden dat een dergelijke tweede factor niet bestaat. De eerder gegeven verklaring voor de relatief lage topicbetrouwbaarheid van het holistische oordeel gaat er juist vanuit dat beoordelaars zich door de inhoud van de tekst laten beïnvloeden. Deze veronderstelling lijkt ook in overeenstemming met de eerder besproken uitkomsten van EEG-onderzoek (zie deelstudie 4, paragraaf 7.1) en met het idee dat taal en teksten vorm en inhoud hebben. Het aantonen van een tweede factor zou het doel van verder onderzoek moeten zijn.

Conclusies en discussie

1. Wil een maat in een onderwijssetting zinvol zijn, dan moet die maat niet alleen geschikt zijn om te meten (te toetsen), maar ook om te trainen (te oefenen). Anders ontstaat de situatie dat meten wel kosten (aan bijvoorbeeld tijd) met zich mee brengt, maar niet resulteert in effectieve interventie.

Deze stelling is primair een uitgangspunt en ontwerpprincipe. De invoering van het essay-deel in de SAT door de College Board laat wel zien dat bij al te duidelijke afwijking van dit principe de politieke druk zo groot kan worden, dat de testende instantie op zijn minst voor een deel overstag moet gaan. De ervaringen met TAVAN tot nu toe wijzen erop dat het combineren van oefenen en meten zeer effectief kan zijn. Traditionele objectieve toetsen om schrijfvaardigheid te meten lijken op dit punt niet te voldoen. De stelling van Godshalk et al. (1966) dat objectieve tests valide zijn, is volgens dit uitgangspunt niet zonder meer juist. Deze objectieve tests zijn criteriumvalide (correleren met het holistisch oordeel), maar niet 'onderwijsvalide': ze zijn niet bruikbaar zijn als trainingsmethode.

2. Objectieve tests om schrijfvaardigheid te meten overlappen doorgaans voor meer dan 75% met de holistische totaalscore (na correctie voor onbetrouwbaarheid) zodat beide maten voor het grootste deel dezelfde factor meten.

Deze hoge waarden zijn in meerdere onderzoeken gevonden (zie paragraaf 8.5), maar zelden duidelijk voor het voetlicht gebracht. Wel moet hierbij aangetekend worden dat dit alleen geldt voor objectieve tests die overwegend de FOC-factor meten en niet automatisch voor iedere willekeurig geconstrueerde meerkeuze-taaltoets.

3. De verklaring voor het 'vreemde' verband tussen objectieve tests om schrijfvaardigheid te meten en het holistische oordeel dat Godshalk et al. (1966) aantoonden, is dat de objectieve tests ontwikkeld zijn om vast te stellen hoe goed studenten scoren op de FOC-factor. Het niveau van de FOC-factor bepaalt vervolgens hoeveel fouten PHW een student in zijn tekst maakt. Het aantal fouten PHW in de tekst bepaalt vervolgens in belangrijke mate het holistische oordeel.

De juistheid van deze verklaring kon worden aangetoond via de gegevens van het fouteneffect-onderzoek in combinatie met de resultaten van TAVAN2. Alle drie betrokken variabelen bleken onderling hoog te correleren (gemiddelde na correctie voor onbetrouwbaarheid: 0.86) en daarmee in hoge mate dezelfde (FOC-)factor te meten.

4. De TAVAN-score en het aantal fouten PHW vormen twee nieuwe manieren om basale schrijfvaardigheid vast te stellen. Beide maten blijken criteriumvalide te zijn (correleren hoog correleren met het holistische oordeel en onderling). Dit brengt het totaal aantal beschikbare maten met aangetoonde criteriumvaliditeit op vier. De twee reeds eerder bekende manieren waren: het holistische oordeel en de objectieve test.

5. Een derde nieuwe manier waarop basale schrijfvaardigheid op een criteriumvalide manier bleek te kunnen worden vastgesteld, was het meten van de kwaliteit van het holistische oordeel uitgebracht door de student over zes teksten geschreven door medestudenten. Naarmate een student het ideale holistische oordeel beter benaderde, bleek hij ook hoger te scoren op de overige drie schrijfvaardigheidsmaten. De gemiddelde onderlinge correlatie met de overige drie schrijfvaardigheidsmaten was na correctie voor onbetrouwbaarheid 0.88 en daarmee zeer hoog.

6. Alle vijf maten om basale schrijfvaardigheid te meten die aangetoond criteriumvalide zijn (holistische oordeel, objectieve test, TAVAN-score, aantal fouten PHW, kwaliteit van het uitgebrachte holistische oordeel) blijken overwegend de FOC-factor te meten.

Dat de objectieve tests voor het meten van schrijfvaardigheid vooral de FOC-factor meten, blijkt uit de inhoud van de items. De TAVAN-score gaat expliciet uit van foute zinnen die gecorrigeerd moeten worden. Het aantal fouten PHW richt zich uiteraard op 'fouten' (wat dat precies zijn, bepaalt de beoordelaar) in de tekst. Daar al deze maten onderling hoog correleren, meten kennelijk ook de overige twee maten overwegend de FOC-factor.

7. De nieuwe maat 'kwaliteit van het uitgebrachte holistische oordeel' valt in beginsel ook te gebruiken als trainingsmethode door uit te gaan van twee korte tekstfragmenten per item waaruit de beste gekozen moet worden. Daarmee zou naast de TAVAN-methode een tweede methode beschikbaar komen om basale schrijfvaardigheid te trainen. Op dit moment is deze methode nog niet gerealiseerd en vormt dit een punt voor verder onderzoek.

Het theoretisch belang van een dergelijke toepassing is, mits aangetoond als effectief, dat dit zou aantonen dat schrijven ook geleerd kan worden op een receptieve manier, namelijk door te laten lezen.

8. Van de verschillende manieren om basale schrijfvaardigheid te meten, is de TAVAN-score tot nu toe - voor zover ons bekend - de enige methode waarvan de effectiviteit voor trainingsdoeleinden is aangetoond.

Daar er tot nu slechts twee gangbare methodes waren om basale schrijfvaardigheid vast te stellen, waarvan de objectieve test alleen geschikt werd geacht voor selectiedoeleinden zou er voor het holistische oordeel dan duidelijke evidentie moeten zijn van een leereffect. Dat

leereffect is echter moeilijk betrouwbaar te kwantificeren, doordat het holistische oordeel bekend onbetrouwbaar is. Verder is het erg gevoelig voor beoordelaarseffecten die kunnen optreden als niet volledig blind wordt beoordeeld. In beginsel zou men echter ook van herhaald laten schrijven steeds gevolgd door holistische beoordeling wel een trainingseffect verwachten. Op dit moment is ons echter geen onderzoek bekend waaruit dit duidelijk blijkt. Dat hoeft echter nog niet te betekenen dat dergelijk onderzoek niet bestaat. Indien dergelijk onderzoek wel te vinden zou zijn, zou het interessant zijn om de effectiviteit van beide methodes te vergelijken.

De belangrijkste conclusie van dit deelonderzoek naar het meten van basale schrijfvaardigheid lijkt te moeten zijn dat de TAVAN-score, het aantal fouten PHW en het holistische oordeel onderling zeer sterk gecorreleerd zijn (gemiddelde correlatie: 0.77, na correctie voor onbetrouwbaarheid: 0.86, zie paragraaf 8.7) en dat alle drie in zeer hoge mate de FOC-factor meten: het kunnen opsporen van fouten in tekst en het corrigeren daarvan. Basale schrijfvaardigheid lijkt daarmee vooral te maken te hebben met het kunnen opsporen en corrigeren van fouten.

Hoewel er in dit onderzoek geen tweede (niet-FOC) taalfactor werd gevonden, lijkt het voorlopig plausibel te veronderstellen dat die wel bestaat. Toekomstig onderzoek zou moeten proberen deze factor aan te tonen. Op voorhand valt echter reeds te stellen dat een dergelijke tweede factor vele malen minder belangrijk zal zijn, in termen van het percentage verklaarde variantie in het holistische oordeel, dan de FOC-factor. Kennelijk spelen fouten in de tekst een doorslaggevende rol bij holistische beoordeling, ook al is dat in de opvatting van sommigen misschien onterecht.

9

Deelstudie 6

Betrouwbaarheidsproblemen

Inleiding

Holistische beoordeling en onbetrouwbaarheid lijken onafscheidelijk. In dit hoofdstuk wordt gezocht naar een mogelijke verklaring voor dit fenomeen. Als men schrijfvaardigheid alleen goed kan beoordelen door studenten te laten schrijven, zoals men lang heeft aangenomen, hoe is het dan mogelijk dat beoordelaars onderling zo weinig overeenstemmen? Het idee achter holistische beoordeling was juist dat het eenvoudig te zien was aan een geschreven tekst of iemand wel of niet goed kon schrijven. De lage betrouwbaarheid die vaak gevonden wordt bij holistische beoordeling lijkt op gespannen voet te staan met dit idee.

Allereerst gaan we in paragraaf 9.1 in op de verschillende manieren waarop betrouwbaarheid kan worden gekwantificeerd en de problemen die aan het berekenen van de betrouwbaarheid verbonden kunnen zijn. Onderzoekers zijn geneigd betrouwbaarheid als een feitelijk gegeven te zien. Er blijken echter op hetzelfde databestand een groot aantal verschillende intraclass-correlaties als maat voor betrouwbaarheid berekend te kunnen worden, die qua hoogte sterk verschillen. Vervolgens is in de praktijk niet altijd duidelijk wat er precies berekend is en gaat er bij de berekening gemakkelijk iets mis. Ten slotte levert een design met meerdere beoordelaars en meerdere topics een aantal specifieke problemen op bij het berekenen van de betrouwbaarheid.

De bedoeling van deze paragraaf is collega-onderzoekers te attenderen op mogelijke valkuilen en daarnaast levert deze paragraaf het begrippenkader voor de daarop volgende paragrafen.

In 9.2 laten we voor het aantal fouten PHW (per honderd woorden) zien door een opsplitsing van de teksten van het fouteneffect-onderzoek dat voor een betrouwbare beoordeling (>0.80) een tekst met een lengte van een halve A4 (250 woorden) al volstaat. De topicbetrouwbaarheid bij beoordeling op aantal fouten per honderd woorden blijkt zeer hoog te zijn, dat wil zeggen: het aantal fouten dat studenten maken blijkt een zeer betrouwbare indicator van basale schrijfvaardigheid te zijn.

Het belang van deze paragraaf is dat voor het vaststellen van basale schrijfvaardigheid een korte tekst, van een halve A4, al voldoende is. Het idee dat voor het vaststellen van basale schrijfvaardigheid een hele serie teksten nodig is, klopt kennelijk niet.

In 9.3 gaan we verder in op dit punt. Het resultaat dat bij beoordeling op aantal fouten PHW een halve A4 voldoende is, lijkt in duidelijk contrast te staan met wat bekend is over het aantal benodigde topics bij holistische beoordeling. Technischer geformuleerd: hoe is het mogelijk dat de topicbetrouwbaarheid bij holistische beoordeling zoveel lager ligt dan bij beoordeling op aantal fouten PHW?

Onze verklaring is ten slotte dat de lage topicbetrouwbaarheid niet ontstaat doordat de schrijfvaardigheid van de studenten per topic sterk fluctueert zoals gemakshalve altijd werd aangenomen, maar dat de holistische beoordelaars zich laten beïnvloeden ('meeslepen') door de inhoud van de teksten in plaats van door het taalgebruik.

9.1 Welke (on)betrouwbaarheid?

Het grootste probleem bij holistische beoordeling is de lage betrouwbaarheid. Breland (1983, p. 1) schreef: "the history of direct writing skill assessment is a bleak one. As far back as 1880 it was recognized that the essay examination was beset with the curse of unreliability." Het probleem signaleren is vermoedelijk de eerste stap naar een oplossing.

Een tweede belangrijke stap om grip te krijgen op het probleem is het kwantificeren van de betrouwbaarheid. In de psychologie en testleer is de gebruikelijke definitie van betrouwbaarheid de verwachte correlatie met een soortgelijke tweede meting op dezelfde personen. In de praktijk is die tweede meting niet nodig en wordt coëfficiënt alfa gebruikt om op basis van de correlaties tussen de items van de test de verwachte correlatie met een andere test te voorspellen die opgebouwd is uit hetzelfde aantal en soort items. De empirische basis voor die voorspelling is de gemiddelde onderlinge correlatie tussen de itemscores van de afgenomen test en het aantal items in de test (in de praktijk worden de scores niet altijd gestandaardiseerd en worden in dat geval niet de onderlinge correlaties gebruikt, maar de onderlinge covarianties, maar verder maakt dit geen wezenlijk verschil).

Bij de (holistische) beoordeling van de schrijfvaardigheid ligt het kwantificeren echter ingewikkelder, terwijl de onderzoeker zich dat vaak niet onmiddellijk realiseert. Onderzoekers vermelden vaak wel betrouwbaarheden, maar niet wat en hoe ze die precies berekend hebben. De literatuur op dit gebied biedt allerhande oplossingen en mogelijkheden, die echter

vaak belangrijk verschillende waarden opleveren. Shrout & Fleiss (1979, p. 420) behandelen zes voorbeelden van intraclass correlaties die bij meetfouten en beoordelingsproblemen gebruikt kunnen worden en schrijven hierover:

> *There are numerous versions of the intraclass correlation coefficient (ICC) that can give quite different results when applied to the same data. Unfortunately, many researchers are not aware of the differences between the forms, and those who are often fail to report which form they used. . . . Unfortunately, most textbooks . . . describe only one or two forms of the several possible. Making the plight of the researchers worse, some of the older references . . . contain mistakes.*

In hun artikel geven ze als voorbeeld zes verschillende correlaties berekend voor dezelfde dataset. De laagste waarde was .17, de hoogste .91. Welke correlatie precies wordt uitgerekend kan als schatting van de betrouwbaarheid dus veel verschil uitmaken. In de praktijk blijkt ook niet altijd duidelijk te zijn, wat nu precies nog wel een ICC is en wat niet meer.

> *A rigorous definition is [in this article] adopted for the ICC, namely, that the ICC is the correlation between one measurement (either a single rating or a mean of several ratings) on a target and another measurement obtained on that target. The ICC is thus a bona fide correlation coefficient.*
> (Shrout & Fleiss, 1979, p. 422)

Allerhande berekende en bedachte indexen worden daarmee door Shrout en Fleiss als het ware verboden: de betrouwbaarheidsindex moet een daadwerkelijk waargenomen correlatie zijn. Men kan zich afvragen of dit helemaal realistisch is. Is coëfficiënt alfa, die als voorbeeld door Shrout en Fleiss behandeld wordt, een waargenomen correlatie? De gemiddelde onderlinge correlatie tussen de items wordt waargenomen, maar de verwachte correlatie met die even lange test opgebouwd uit hetzelfde soort items, nog niet. De bedoeling lijkt echter duidelijk en terecht: gerapporteerde coëfficiënten moeten een duidelijke empirische basis hebben en niet stilzwijgend via slimme bewerkingen worden opgehoogd.

Bij een normale studietoets heeft men het over 'de betrouwbaarheid'. Gulliksen (1936) heeft het in dit verband over de 'total test reliability' (p. 189). Wanneer de betrouwbaarheid (in de praktijk meestal coëfficiënt alfa) laag is, is de kans groot dat een student die op de toets zeer onvoldoende scoorde, bij de herkansing een royale voldoende haalt zonder de stof opnieuw bestudeerd te hebben.

De problemen rond de betrouwbaarheid bij holistische beoordeling treden op, zodra men de resultaten van meerdere schrijfopdrachten ('topics') laat beoordelen door meerdere beoordelaars die niet volledig overeenstemmen gegeven dezelfde teksten. Ook wanneer men een enkele beoordelaar gebruikt die bijvoorbeeld een tijdje later hetzelfde topic opnieuw beoordeelt, maar dan niet volledig overeenstemt met de eerdere beoordeling, ontstaat een soortgelijk probleem.

Bij een normale studietoets is de betrouwbaarheid in eerste instantie een empirisch waarneembaar iets. Wanneer men twee versies maakt voor dezelfde toets en beide versies afneemt bij dezelfde studenten, zullen de uitkomsten niet perfect correleren, maar slechts beperkt. Omdat een studietoets normaal bestaat uit een groot aantal items, kan men ieder item opvatten als een korte toets en de gemiddelde onderlinge correlatie tussen de items kan dan gebruikt worden om de correlatie met een even lange nieuwe toets te voorspellen. Het model achter coëfficiënt alfa sluit dus dicht aan op bij de observaties en gebruikt geen veronderstellingen over de redenen waarom de scores tussen afnames fluctueren. Het model geldt daardoor als zeer robuust.

Bij holistische beoordeling van een aantal topicopdrachten door een aantal beoordelaars is de situatie echter gecompliceerder. De topicopdrachten komen overeen met de items bij een normale studietoets. De beoordelaars vormen nu echter een extra component in het geheel. Op deze complicatie werd op twee verschillende manieren gereageerd die in de praktijk soms ook nog door elkaar gingen lopen. De ene manier is gebaseerd op een correlationele benadering, de andere manier van reageren is gebaseerd op een variantie-analytische benadering. Qua berekende uitkomsten zullen beide benaderingen, indien goed toegepast, dezelfde resultaten opleveren, maar de getrokken conclusies zullen in de praktijk gemakkelijk verschillen.

Bij de correlationele aanpak berekenen we de betrouwbaarheid op basis van de gemiddelde onderlinge correlatie tussen de topicscores en blijft de berekening van de betrouwbaarheid van de totale test gelijk. Door te kijken naar de gemiddelde onderlinge correlatie van de beoordelaars die hetzelfde topic beoordelen, kan echter ook de beoordelaarsbetrouwbaarheid van de topicscore worden bepaald. Dit maakt het mogelijk de correlatie tussen de topicscores te corrigeren voor de onbetrouwbaarheid van de beoordeling. Op deze manier vindt men de betrouwbaarheid van de topicscore bij perfect betrouwbare beoordeling. Omdat deze

laatste waarde vastligt, kan men hierna uitrekenen wat er gebeurt met de totale testbetrouwbaarheid bij verandering van het aantal beoordelaars. De correlationele aanpak maakt geen assumpties over het aan de beoordelingen ten grondslag liggende model, is daardoor relatief robuust en maakt het eenvoudig mogelijk voorspellingen te doen.

Godshalk et al. (1966) gebruikten een variantie-analytische benadering, die verder niet werd toegelicht, om de totale testbetrouwbaarheid te berekenen. Hoewel deze benadering tot dezelfde totale testbetrouwbaarheid blijkt te leiden als de meer gebruikelijke benadering gebaseerd op correlaties, maakte deze variantie-analytische aanpak de wijze waarop de betrouwbaarheid werd berekend en kon worden berekend, niet voor iedereen inzichtelijker. Als extra leverde de methode ook de beoordelaarsbetrouwbaarheid van de totale test. Deze blijkt echter ook via correlaties afgeleid te kunnen worden, terwijl de praktische waarde in beide gevallen onduidelijk is. Dit laatste punt wordt verderop in deze paragraaf uitgebreider besproken.

Een nevenopbrengst van het variantie-analytische model was dat men op een bepaalde manier naar de holistische beoordelingen ging kijken. De basisassumptie van het variantie-analytische model is namelijk dat het holistische oordeel tot stand komt op basis van drie factoren: de student, het topic en de beoordelaar. In de analyse van Godshalk et al. (1966) werden voor alle drie factoren significante hoofdeffecten vastgesteld (p. 12). Beoordelaars en topics werden daardoor hierna als aangetoonde bronnen van variantie gezien.

Hierna volgen we de correlationele benadering zoals Coffman (1966) die in ander verband summier formuleerde. Stel dat studenten twee topics hebben geschreven en dat ieder topic beoordeeld is door twee beoordelaars. In de datafile krijgen we nu per student vier gegevens, dus vier variabelen met beoordelingen. Wanneer we de topics aangeven met A en B en de beoordelaars met 1, 2, 3 en 4 levert dit als variabelen: A1, A2, B3, B4.

Een eerste mogelijkheid is alle vier variabelen op te geven voor de gebruikelijke berekening van coëfficiënt alfa. De berekening van coëffiënt alfa is tegenwoordig vaak een standaardresponse bij data-analyses die vrij automatisch en zonder veel nadenken wordt uitgevoerd, in de verwachting dat dit de betrouwbaarheid geeft. In dit geval zijn er echter meerdere betrouwbaarheden (bronnen van variantie) en kan men zich afvragen of alfa in dit geval wel de juiste index vormt. De veronderstelling bij alfa is dat de items gelijksoortig zijn. Wie

naar de variabelen in de datafile kijkt, is geneigd te veronderstellen dat dit klopt. Wie echter de moeite neemt de correlatiematrix (zie Tabel 9.1) te inspecteren, ziet dat daar in feite twee verschillende soorten correlaties inzitten in plaats van één. Bij vier correlaties ('bb') is het topic hetzelfde, maar de beoordelaars verschillend. Bij acht correlaties ('sb') zijn niet alleen de beoordelaars verschillend, maar ook de topics.

Tabel 9.1 Correlatiematrix voor 4 beoordelaars/beoordelingen (1-4) en 2 topics (A , B).

	A1	A2	B3	B4
A1	--	bb	sb	sb
A2	bb	--	sb	sb
B3	sb	sb	--	bb
B4	sb	sb	bb	--

Op basis van deze vier variabelen vallen er in feite twee verschillende soorten betrouwbaarheid te onderscheiden. Bij de gemiddelde bb-correlatie gaat het om de beoordelaarsbetrouwbaarheid tussen twee beoordelaars. Dit is de overeenstemming tussen twee beoordelaars die hetzelfde topic beoordelen. De vraag die men in feite wil beantwoorden is of men soortgelijke uitkomsten zou krijgen met andere, soortgelijke beoordelaars. Bij de gemiddelde sb-correlaties gaat het om de scorebetrouwbaarheid tussen twee beoordelaars die verschillende topics beoordelen. Doordat alle aspecten van de test hierbij veranderd worden (topics en beoordelaars) werd deze correlatie door Coffman (1966) aangeduid als de 'score reliability' (p. 151) van één topic beoordeeld door één beoordelaar.

Het model van Coffman gaat er vanuit dat alle relevante informatie over de betrouwbaarheid kan worden samengevat in twee gemiddelde correlaties: de gemiddelde correlatie van beoordelaars die hetzelfde topic beoordeelden en de gemiddelde correlatie van beoordelaars die verschillende topics beoordeelden. Topics en ook beoordelaars worden daarmee als inwisselbaar gezien. In de praktijk zal dit niet altijd opgaan. De ene topic zal het soms beter doen dan een andere topic. Zo vonden wij in het onderzoek naar het effect van het TAVAN-programma (hoofdstuk 6) dat de ene tekst beter leek te werken dan de andere tekst. Ook Godshalk et al. (1966) vonden een soortgelijk verschil tussen de twee gebruikte linears.

Behalve dat topics kunnen verschillen, kunnen ook beoordelaars belangrijk verschillen. In het fouteneffect-onderzoek bleken studenten het bijvoorbeeld als holistische beoordelaars gemiddeld belangrijk slechter te doen dan de beide onderzoekers. Aan de andere kant lijkt het goed verdedigbaar om uiteindelijk simpelweg te werken met het gemiddelde. De situatie lijkt te vergelijken met de benadering bij coëfficiënt alfa. Ook daar wordt gewerkt met de gemiddelde correlatie tussen de items. Dat sommige items het beter doen dan andere items kan blijken bij de itemanalyse, maar speelt in de uiteindelijke berekening van alfa geen rol meer. Op dezelfde manier is het mogelijk dat sommige beoordelaars het belangrijk beter doen dan andere, maar kan de uiteindelijke selectie van beoordelaars gekarakteriseerd worden door deze twee gemiddelde correlaties. Doordat alle relevante informatie over de beoordeling geacht wordt opgenomen te zijn in deze twee gemiddelde correlaties, maakt het uitgaande van dit model ook niet echt uit wie er heeft beoordeeld. Of er dus inderdaad vier verschillende beoordelaars gebruikt zijn of in feite twee of zelfs één beoordelaar gebruikt is, maakt voor het model niet uit. Er wordt puur gekeken naar beide gemiddelde correlaties.

Ogenschijnlijk vormen de beoordelaarsbetrouwbaarheid en de scorebetrouwbaarheid alle betrouwbaarheden die men uit de correlatiematrix kan afleiden. De scorebetrouwbaarheid kan echter opgebouwd gedacht worden uit twee componenten: de topics veroorzaken een bepaalde onbetrouwbaarheid en de beoordelaars (Gulliksen, 1936). Stel dat we een scorebetrouwbaarheid van 0.24 gevonden hebben en een beoordelaarsbetrouwbaarheid van 0.36. Die topicscores zijn gebaseerd op een enkele beoordelaar die nogal onbetrouwbaar was. Hoe hoog ligt dan de 'werkelijke' correlatie tussen de topicscores als we perfect betrouwbaar zouden beoordelen? Daar de beoordelaarsonbetrouwbaarheid voorkomt in beide scores die de scorebetrouwbaarheid leveren, komt de correctie voor onbetrouwbaarheid in dit geval neer op de verhouding tussen beide correlaties: 0.24/0.36=0.67. Wanneer we heel veel beoordelaars zouden inzetten, zou de gemiddelde correlatie tussen de topicscores tenslotte vlak bij 0.67 uitkomen, maar nooit daarboven. Deze situatie valt te vergelijken met een meerkeuze item in een studietoets. De beoordeling is perfect betrouwbaar, maar ondanks die perfect betrouwbare beoordeling is de gemiddelde correlatie van het item met andere items uit de toets lang niet perfect.

Uit de scorebetrouwbaarheid (de gemiddelde correlatie tussen twee beoordelaars die verschillende topics beoordelen) van een topic beoordeeld door één beoordelaar en de beoordelaarsbetrouwbaarheid van één beoordelaar (de gemiddelde correlatie tussen twee beoor-

delaars die hetzelfde topic beoordelen) valt dus de topicbetrouwbaarheid bij perfect betrouwbare beoordeling af te leiden. Meer in het algemeen is de scorebetrouwbaarheid (sb) gelijk aan het product van de topicbetrouwbaarheid bij perfect betrouwbare beoordeling (tbp) en de beoordelaarsbetrouwbaarheid (bb). Er geldt dus:

$$sb = tbp \cdot bb$$

Dit is alleen de definitie van de topicbetrouwbaarheid bij perfect betrouwbare beoordeling, maar anders geschreven.

Wanneer we nu meerdere beoordelaars gebruiken, stijgt de beoordelaarsbetrouwbaarheid, maar verandert de topicbetrouwbaarheid niet. De scorebetrouwbaarheid is echter het product van beide en stijgt dus mee. Hoeveel de beoordelaarsbetrouwbaarheid stijgt, valt uit te rekenen via de formule voor testverlenging[1]. Twee beoordelaars vormen een test van dubbele lengte.

Om de totale scorebetrouwbaarheid van de test uit te rekenen, is het voldoende de feitelijke scorebetrouwbaarheid tussen de topics te bepalen bij het gegeven aantal beoordelaars (de onbetrouwbaarheid van de beoordeling zit daarin verwerkt) via de gemiddelde onderlinge correlatie tussen de topicscores en vervolgens via de formule voor testverlenging de betrouwbaarheid van de totale test te berekenen. Het is dus voldoende het oordeel A1 en A2 bij elkaar op te tellen, hetzelfde te doen voor B1 en B2 en vervolgens de correlatie tussen deze twee somvariabelen te bepalen. Vervolgens levert de formule voor testverlenging de totale scorebetrouwbaarheid van beide topics gezamenlijk en daarmee van de totale test. In SPSS kan dit uitgevoerd worden door 'alfa standardized' te berekenen voor de twee somvariabelen: A1+A2 en B1+B2 (door het aanvinken van de optie voor itemanalyse).

Maakt het in de praktijk uit of we de alfa berekenen op basis van A1, A2, B3, B4 of op basis van beide somvariabelen A1+A2, B3+B4? Op basis van een gesimuleerde dataset van 300 personen waarbij we voor de topicbetrouwbaarheid en de beoordelaarsbetrouwbaarheid ernaar streefden ruwweg de waarden van Godshalk et al. (1966) te benaderen (onze topicbetrouwbaarheid was 0.66 en de beoordelaarsbetrouwbaarheid gemiddeld 0.37 over vier topics) vonden we uitgaande van de variabelen A1, A2, B3, B4 een alfa van 0.65 en uitgaande van de somvariabelen A1+A2 en B1+B2 een alfa van 0.59. Het maakt dus inderdaad verschil.

[1] Zie bijlage 1 voor de formule voor testverlenging.

Verder is het verschil zoals men zou verwachten. De bb-correlaties zijn gemiddeld belangrijk hoger dan de sb-correlaties en trekken daardoor de schatting van de betrouwbaarheid omhoog. Doordat het een simulatie betrof, was het eenvoudig de gegevens van een tweede test met nieuwe beoordelaars aan te maken om de juistheid van de berekende alfa te checken. Voor de correlatie tussen de totalen van beide tests vonden we 0.58 wat goed overeenkomt met de eerder gevonden alfa van 0.59. De betrouwbaarheid kan dus op de gebruikelijke manier berekend worden, mits men net als anders uitgaat van de itemscores (de totalen per topic) in plaats van de scores per beoordelaar (voor deze simulatie werd getrokken uit normaalverdelingen met gemiddelde 0 en SD=10 voor het vaardigheidsniveau van de studenten, SD=7 voor de topicfout en SD=15 voor de beoordelaarsfout).

Door uit te gaan van de correlaties tussen A1, A2, B3 en B4 viel hetzelfde resultaat af te leiden. Het gemiddelde van de bb-correlaties was 0.382. Het gemiddelde van de sb-correlaties was 0.291. De topicbetrouwbaarheid is de verhouding van die twee, dus 0.291/0.382=0.762. De betrouwbaarheid van twee beoordelaars samen is via de formule voor testverlenging dan 0.553. De scorebetrouwbaarheid van twee beoordelaars samen wordt dan 0.762x0.553=0.421. De totale test omvatte twee topics, via de formule voor testverlenging vinden we dan 0.593. Dit kwam overeen met de eerder gevonden waarde voor alfa van 0.59.

De voorgaande berekening lijkt misschien wat overbodig. Waarom zouden we de betrouwbaarheid op een ingewikkelde manier berekenen als het op een eenvoudige manier kan? Het voordeel van deze berekeningsmethode is dat we nu ook de betrouwbaarheid kunnen berekenen als bijvoorbeeld het aantal beoordelaars verdubbeld wordt. De betrouwbaarheid van één beoordelaar was 0.382. De betrouwbaarheid van vier beoordelaars is dan volgens de formule voor testverlenging 4x0.382/(1+(4-1)x0.382)=0.712. De scorebetrouwbaarheid wordt dan: 0.712x0.762=0.543. Voor twee topics beoordeeld door vier beoordelaars per topic wordt dan de verwachte betrouwbaarheid via de formule voor testverlenging 0.704. De simulatie levert een alfa van 0.734 en een correlatie met een tweede test bestaande uit twee topics met vier beoordelaars per topic van 0.718. Beide methodes leveren daarmee soortgelijke uitkomsten.

Een andere methode voor de berekening van de betrouwbaarheid bij holistische beoordeling die men mogelijk zou kunnen overwegen en die belangrijk afwijkt van de voorgaande, is het gebruik van alfa-stratified (Nunnally, 1967, p. 229). Wanneer de scores op een aantal

tests worden samengevoegd die een bekende betrouwbaarheid hebben, speelt de betrouwbaarheid van de afzonderlijke tests een rol in de uiteindelijke betrouwbaarheid van het totaal. Wanneer de afzonderlijke tests bijvoorbeeld perfect betrouwbaar zijn, maar onderling niet correleren (totaal verschillende factoren meten) zal de resulterende score toch perfect betrouwbaar zijn. Op deze wijze is het mogelijk de beoordelaarsbetrouwbaarheid in de berekening van de uiteindelijke betrouwbaarheid te betrekken. Op basis van de door Godshalk et al. (1966) vermelde correlatiematrix was het mogelijk alfa-stratified te berekenen (uitgaande van gestandaardiseerde variabelen). Dit leverde een waarde van 0.932 uitgaande van deze gegevens. Coffman (1966) vond echter - in grote lijnen via de eerder beschreven methode - een waarde van 0.841 voor de totale scorebetrouwbaarheid. Alfa-stratified leverde hier dus een belangrijk hogere schatting van de totale betrouwbaarheid op.

De door Coffman gerapporteerde waarde bleek bij narekenen overeen te komen met de betrouwbaarheid van de topicscores uitgaande van hun onderlinge gemiddelde correlatie, dus zonder rekening te houden met de (beoordelaars)betrouwbaarheden van deze subschalen. Deze misschien wat voorzichtige benadering lijkt goed verdedigbaar. Voor zover de beoordelaars het onderling oneens zijn, zit dat al verwerkt in de topicscores. Bij een objectieve studietoets is het ook niet gebruikelijk meerdere beoordelaars in te zetten, vervolgens te concluderen dat die het onderling perfect eens zijn en de betrouwbaarheid van de toets dan te berekenen via alfa stratified. De subschalen zouden dan perfect betrouwbaar zijn en daarom zou de score op iedere objectieve toets, ongeacht de correlaties tussen de items, ook perfect betrouwbaar zijn, wat een weinig realistisch resultaat zou zijn.

Een voordeel van de benadering van Coffman is dat de berekening van de betrouwbaarheid in feite op precies dezelfde manier blijft gaan als bij een normale studietoets: per topic worden de scores van de beoordelaars samengenomen door te middelen of op te tellen. De betrouwbaarheid berekenen we vervolgens op basis van de gemiddelde correlatie (of covariantie) tussen de topicscores (itemscores) en het aantal topics (items).

Vervolgens is het mogelijk de gevonden totale scorebetrouwbaarheid op te splitsen in twee delen, de beoordelaarsbetrouwbaarheid en de topicbetrouwbaarheid, zodanig dat het product de totale scorebetrouwbaarheid moet leveren (Gulliksen, 1936). Het idee achter deze opsplitsing is dat we daardoor als het ware meer greep krijgen op de verschillende bronnen van onbetrouwbaarheid. We kunnen hierna zien of de beoordelaars of de topics de meeste onbetrouwbaarheid veroorzaken, lijkt het.

In de praktijk blijken zich vaak een aantal problemen voor te doen. Een eerste probleem is dat bij holistische beoordeling voortdurend gesproken wordt over beoordelaarsbetrouwbaarheid zonder dat precies duidelijk is, hoe die berekend moet worden of berekend is. Vaak gebeurt het bijvoorbeeld dat beoordelaars overleggen over moeilijk gevallen of standaard de beoordelingen met elkaar doornemen. Een andere gewoonte die wel gevolgd wordt, is dat essays die sterk verschillend worden beoordeeld, vervolgens worden beoordeeld door een derde beoordelaar. Door dit soort goedbedoelde acties wordt het begrip 'beoordelaarsbetrouwbaarheid' vooral een fraaie vlag, die geen feitelijke betekenis meer heeft.

Een volgend probleem is dat voor het berekenen van de beoordelaarsbetrouwbaarheid in een programma als SPSS twee verschillende manieren zijn, waarvan er slechts één juist is. De juiste manier is dat de beoordelingen van een beoordelaar een variabele (een kolom) vormen. De essays vormen dus de regels (de 'cases'). De berekening van coëfficiënt alfa via SPSS levert in dat geval de 'rater alpha' of de beoordelaarsbetrouwbaarheid van alle beoordelaars gezamenlijk. Tegelijkertijd kan dan eenvoudig de gemiddelde onderlinge ('off-diagonal') correlatie tussen de beoordelaars worden opgevraagd. In SPSS is het echter normaal de gewoonte alle resultaten van één persoon op een regel te vermelden. In dat geval gebruikt men dus voor vijf beoordelaars die zestig teksten beoordeelden, vijf regels en zestig variabelen. Wanneer men nu via SPSS coëfficiënt alfa berekent, berekent men de betrouwbaarheid van een toets bestaande uit zestig items (teksten) die voorgelegd is aan vijf 'studenten'. Deze 'toets' meet dan of een 'student' het goed of slecht doet, in dit geval wordt dus gemeten of een beoordelaar hoog of laag beoordeelt. Daar beoordelaars nogal verschillen in hoe kritisch ze zijn en het aantal items in dit geval uitermate groot is, is het resultaat dat de beoordeling uiterst betrouwbaar was. In werkelijkheid heeft men echter alleen aangetoond dat beoordelaars niet allemaal even kritisch zijn; iets wat al lang bekend was. Verderop in deze paragraaf wordt het onderzoek van Farmer (2008) besproken waar deze fout gemaakt werd.

Wanneer de beoordelingen wel onafhankelijk van elkaar plaatsvinden en de berekening van de beoordelaarsbetrouwbaarheid wel correct verloopt, is vervolgens de vraag, wat het nut daarvan precies is. Een holistisch beoordeelde topictekst heeft, zoals hiervoor besproken werd, drie soorten betrouwbaarheid: de beoordelaarsbetrouwbaarheid, de scorebetrouwbaarheid en de topicbetrouwbaarheid bij perfect betrouwbare beoordeling. Voor de berekening van de scorebetrouwbaarheid doet de beoordelaarsbetrouwbaarheid niet ter zake, maar dat is normaal niet iedereen onmiddellijk duidelijk. De kans is bijvoorbeeld aanwezig dat

de beoordelaarsbetrouwbaarheid verward wordt met de scorebetrouwbaarheid. Ook misverstanden rond de topicbetrouwbaarheid kunnen gemakkelijk ontstaan, doordat de topicbetrouwbaarheid bij een gegeven aantal beoordelaars gelijk is aan de scorebetrouwbaarheid. De 'topicbetrouwbaarheid bij perfect betrouwbare beoordeling' wijkt daar belangrijk van af, maar wordt in het gebruik nog wel eens afgekort tot de 'topicbetrouwbaarheid' zodat dit gemakkelijk een uitermate ambigu begrip wordt. Strikt genomen is het begrip ook niet noodzakelijk, omdat een toets bestaande uit holistisch beoordeelde essayopdrachten net als ieder andere toets, gekarakteriseerd kan worden met de totale scorebetrouwbaarheid. Informatie over de beoordelaarsbetrouwbaarheid vormt alleen extra informatie, die echter ook gemakkelijk complicerend kan werken.

In beginsel is het dus voldoende voor een holistisch beoordeelde toets bestaande uit meerdere essayopdrachten alleen de totale scorebetrouwbaarheid te vermelden en eventueel als extra informatie de 'beoordelaarsbetrouwbaarheid'. Het meest voor de hand liggend zou dan zijn de gemiddelde onderlinge correlatie tussen beoordelaars die hetzelfde topic beoordelen, te rapporteren. Deze correlatie ligt echter in de praktijk vaak nogal laag. Voor het onderzoek van Godshalk et al. (1966) lag die volgens Coffman (1966, p. 154) op 0.386. Een onderzoeker die vijf beoordelaars per topic gebruikt heeft, zal daarom in de praktijk liever de betrouwbaarheid van deze vijf beoordelaars rapporteren door via de formule voor testverlenging de gevonden gemiddelde correlatie op te waarderen tot een test van vijf beoordelaars. Voor het onderzoek van Godshalk et al. lag deze bij vijf beoordelaars voor één topic op 0.759 (Coffman, 1966, p. 154), een meer aansprekende waarde. Voor een collega-onderzoeker die vertrouwd is met de formule voor testverlenging, vormt deze opwaardering geen probleem, via dezelfde formule kan ook weer de oorspronkelijke correlatie terug worden gevonden. Aan de andere kant zal niet iedereen vertrouwd zijn met de formule voor testverlenging.

Wanneer een onderzoeker beschikt over meerdere topicteksten, kan men echter nog een stap verder gaan. De topicbetrouwbaarheid bij perfect betrouwbare beoordeling van een enkel topic was in het onderzoek van Godshalk et al. 0.681 (Coffman, 1966, p. 154). In totaal werden echter vijf topics gebruikt. Bij perfect betrouwbare beoordeling was de topicbetrouwbaarheid van vijf topics dan volgens de formule voor testverlenging 0.914. De totale scorebetrouwbaarheid op basis van de gemiddelde onderlinge correlatie tussen de topicscores en het aantal topics was bekend en was 0.841. De totale beoordelaarsbetrouwbaarheid van de test moest dan 0.921 zijn. De scorebetrouwbaarheid is immers gelijk aan het product

van de topicbetrouwbaarheid bij perfect betrouwbare beoordeling en de beoordelaarsbetrouwbaarheid. Het probleem met deze laatste opwaardering van de beoordelaarsbetrouwbaarheid is dat terugrekenen in de praktijk amper mogelijk is. Verder dient deze berekende beoordelaarsbetrouwbaarheid geen enkel praktisch doel. De gerapporteerde correlatie is ook niet daadwerkelijk waargenomen, hoewel Coffman ogenschijnlijk suggereerde van wel (1966, p. 154). Een andere vergissing in het artikel, is dat de vermelde Spearman-Brown formule voor testverlenging een fout bevat. In de teller van de breuk is de testverlengingsfactor 'n' weggevallen.

Samenvattend: er zijn veel verschillende soorten betrouwbaarheid in omloop zijn, waardoor de berekening van de juiste betrouwbaarheid, die voor onderzoekers bij holistische beoordeling toch al lastig kan zijn, nog complexer wordt. Verder realiseren onderzoekers zich de problemen op dit gebied niet altijd, waardoor men veronderstelt dat de betrouwbaarheid een objectief gegeven is, dat alleen nog maar even berekend hoeft te worden.

Er doet zich op dit gebied echter mogelijk nog een tweede probleem voor: door de opsplitsing van de scorebetrouwbaarheid in topicbetrouwbaarheid bij perfect betrouwbare beoordeling en beoordelaarsbetrouwbaarheid, wordt de indruk gewekt dat het topic verantwoordelijk is voor een bepaald deel van de variantie in de beoordeling. Verondersteld wordt dat het topic de schrijfvaardigheid van de student zou beïnvloeden, doordat bepaalde studenten over sommige onderwerpen meer weten dan over andere onderwerpen. Een variantie-analytische benadering versterkt die suggestie nog verder door de topics op te voeren als variantiebronnen. Het is echter de vraag of deze veronderstelling juist is. Het zou immers ook kunnen zijn dat beoordelaars zich op systematische wijze door de interactie tussen student en topic laten beïnvloeden. Met andere woorden: het zou kunnen zijn dat de topiconbetrouwbaarheid op zijn minst voor een deel ook beoordelaarsonbetrouwbaarheid is. In de volgende paragrafen wordt verder op dit punt ingegaan.

Hierna volgt een voorbeeld dat laat zien dat de berekening en de rapportage van betrouwbaarheid bij het (holistisch) beoordelen van teksten gemakkelijk grote problemen kan leveren. Farmer (2008) onderzocht het effect van een trainingsprogramma om de betrouwbaarheid van holistische beoordelaars te verhogen. Ze gebruikte een pretest-posttest design met zeventien beoordelaars die ieder dezelfde vier teksten beoordeelden afkomstig van vier verschillende topics. Hoewel ze meerdere bladzijden besteedt aan een overzicht van de problemen rond het meten van beoordelaarsbetrouwbaarheid (p. 20-27) formuleert ze de resulta-

ten voor de pretest als volgt (p. 55-56):

> *Cronbach's Alpha was the statistic used to asses inter-rater reliability in accordance with guidelines recommended by Atkinson and Murray (1987). A two-way mixed-model intraclass correlation was used because this study compared multiple raters that scored the same writing samples. For the four pre-test scores, the alpha coefficient was .913, well above the .80 standard for high intraclass correlations.*

Voor de posttest-resultaten vermeldt ze een soortgelijke passage met een gerapporteerde alfa van 0.919 (p. 76).

Op basis van deze uitkomsten zou men verwachten dat a. de training nogal overbodig was (een beoordelaarsbetrouwbaarheid van 0.913 laat weinig te wensen over) en b. dat de training geen duidelijk effect had (een verschil van 0.006 is niet dermate groot dat het een training lijkt te rechtvaardigen.) In haar conclusies volgde ze deze zienswijze niet. Ze merkte op: "Several researchers, however, have noted that the alpha coefficient is less than an ideal measurement when comparing multiple raters" (Farmer, 2008, p. 82). Als dit inderdaad zo zou zijn, rijst de vraag waarom ze eerst voor coëfficiënt alfa heeft gekozen.

Omdat er in dit geval slechts vier beoordeelde teksten beschikbaar waren per beoordelaar was het niet mogelijk rekening te houden met de vier verschillende topics: van ieder topic is slechts één tekst beschikbaar. De gebruikelijke manier om de beoordelaarsbetrouwbaarheid uit te rekenen is dan de correlaties tussen de beoordelaars te gebruiken. Coëfficiënt alfa zou dan berekend worden over zeventien beoordelaars (de zeventien 'items') waarbij de datafile slechts vier cases zou bevatten. Iedere beoordelaar zou op deze manier een eigen kolom/variabele hebben.

In werkelijkheid valt uit 'the four pre-test scores' af te leiden dat ze kennelijk niet zeventien variabelen in de berekening van coëfficiënt alfa heeft betrokken, maar slechts vier. Ze heeft dus de getransponeerde data-matrix gebruikt met zeventien cases (de beoordelaars) en vier variabelen (de vier pretest-scores). In dat geval heeft ze dus niet de overeenstemming tussen de beoordelaars berekend, zoals ze veronderstelt, maar de overeenstemming tussen de teksten/topics. Iedere beoordelaar heeft een eigen gemiddelde (de ene beoordelaar is wat kritischer dan gemiddeld, de andere wat minder kritisch dan gemiddeld) en de hoge alfa laat zien dat het beoordelen van vier teksten al voldoende is om beoordelaars op dit punt betrouwbaar in te delen.

Het vervolg van het verslag bevestigde dat de beoordelaars in de data-file inderdaad de cases vormden. Op pagina 53 is een t-test 'one sample' uitgevoerd blijkens de vormgeving van de tabel met SPSS over de holistische beoordeling van de zeventien beoordelaars van 'writing sample A'. In SPSS is dit alleen mogelijk als alle beoordelingen van tekst A één variabele vormen. Er lijkt dus weinig twijfel mogelijk te zijn dat in dit geval niet de beoordelaarsbetrouwbaarheid berekend is.

Wanneer de beoordelaars wel in de datafile waren opgenomen in de vorm van variabelen, had zich vervolgens een ander probleem voorgedaan. Voor de berekening van coëfficiënt alfa verwijst Farmer naar Atkinson en Murray (1987). Bij nazoeken blijkt in dit congrespaper de formule voor 'alfa' echter fout vermeld te zijn. In plaats van de gebruikelijke berekeningsformule te geven (Nunnally, 1967, p. 196, formule 6-26), geven Atkinson en Murray de formule voor alfa op basis van de gemiddelde onderlinge correlatie, dus in feite de Spearman-Brown formule voor testverlenging (Nunnally, 1967, p. 193, formule 6-18). In de noemer van de breuk vervangen ze (k-1) echter per abuis door (k). De vraag die daarmee ontstaat, is: welke formule is voor de berekening van alfa gebruikt, de juiste of de foute? Dit voorbeeld is alleen bedoeld te illustreren, hoe gemakkelijk er op het gebied van de berekening van de betrouwbaarheid veel mis kan gaan.

Het probleem dat in deze paragraaf gesignaleerd werd, is dat de berekening van de betrouwbaarheid bij holistische beoordeling gemakkelijk problemen kan geven, hoewel onderzoekers zich dat vaak niet realiseren en geneigd zijn betrouwbaarheid als een objectief gegeven te presenteren. Een groot aantal uiteenlopende betrouwbaarheidsindexen kan berekend worden, verder bestaan er verschillende soorten betrouwbaarheid. Ten slotte levert de berekening vaak onverwachte problemen op en rapporteren onderzoekers niet altijd duidelijk wat ze precies berekend hebben.

Voorgesteld wordt de totale scorebetrouwbaarheid van een test in navolging van Godshalk et al. (1966) en Coffman (1966) te baseren op de gemiddelde correlatie tussen de topicscores en het aantal topicscores. De basiscomponenten van de totale scorebetrouwbaarheid worden dan gevormd door de gemiddelde bb-correlaties (de correlaties tussen beoordeelaars die hetzelfde topic beoordeelden, de basis voor de beoordelaarsbetrouwbaarheid) en de gemiddelde sb-correlaties (de correlaties tussen beoordelaars die verschillende topics beoordeelden, de basis voor de scorebetrouwbaarheid). Op basis van deze twee waarden kan de topicbetrouwbaarheid bij perfect betrouwbare beoordeling geschat worden en is het mo-

gelijk de totale scorebetrouwbaarheid van een test met X topics en Y beoordelaars (of beoordelingen) per topic te voorspellen.

9.2 Benodigde tekstlengte voor meten basale schrijfvaardigheid

Men kan zich afvragen, hoe stabiel een maat als het aantal fouten PHW is. Het is mogelijk dat een student bij het schrijven van een tekst de ene keer veel fouten maakt per honderd woorden en een andere keer weinig. Met andere woorden: is het zo dat een student schrijft met een relatief vast aantal fouten PHW of wisselt dit sterk per tekstdeel? Om hier enig zicht op te krijgen zijn de 48 door beide onderzoekers op fouten beoordeelde teksten van het fouteneffect-onderzoek (fouten werden altijd onderstreept of aangestreept op de eigen kopie van de beoordelaar) opgesplitst in twee helften die evenveel of vrijwel evenveel (in geval van een oneven aantal) woorden telden. Voor het tellen van de woorden is gebruik gemaakt van de woordentelfunctie in OpenOffice Writer. Vervolgens is per onderzoeker het aantal fouten PHW in de eerste helft en in de tweede helft van iedere tekst bepaald. Wanneer studenten soms met veel fouten zouden schrijven en soms met weinig, dan was te verwachten dat er weinig overeenkomst tussen de twee helften in aantal fouten zou bestaan en dat de correlatie tussen beide helften laag zou zijn.

De correlatie tussen de aantallen fouten PHW in beide helften (48 teksten) bedroeg na standaardisatie per helft en per beoordelaar 0.817. Na correctie voor beoordelaarsonbetrouwbaarheid bedroeg de gecorrigeerde correlatie 0.972. Dit is de topicbetrouwbaarheid van een tekst ter lengte van een halve A4 bij perfect betrouwbare beoordeling. De totale betrouwbaarheid van beide helften samen bedroeg 0.899.

Deze hoge waarden wijzen erop dat het aantal fouten per honderd woorden een zeer betrouwbare maat vormt en dat een tekst van een halve A4 het al mogelijk maakt een student betrouwbaar in te delen qua basale schrijfvaardigheid. Een tekst van een A4 beoordeeld door twee beoordelaars levert een betrouwbaarheid die normaal alleen met objectieve tests voor selectiedoeleinden wordt bereikt.

De correlaties tussen de aantallen fouten PHW na standaardisatie voor beide helften en beide beoordelaars zijn weergegeven in Tabel 9.2. De tabel bevat twee xx-correlaties waarbij

de beoordelaars hetzelfde zijn, maar de teksthelften verschillend, twee bb-correlaties waarbij de teksthelften hetzelfde zijn, maar de beoordelaars verschillend en twee sb-correlaties waarbij zowel de teksthelften als de beoordelaars verschillend zijn. De bb-correlaties leveren een schatting van de beoordelaarsbetrouwbaarheid. De sb-correlaties leveren een schatting van de scorebetrouwbaarheid.

Tabel 9.2 Correlaties tussen aantallen fouten PHW per teksthelft (A/B) en beoordelaars (1/2); bb=beoordelaarsbetrouwbaarheid, sb=scorebetrouwbaarheid

	A1	B1	A2	B2
A1	---			
B1	0.674 xx	---		
A2	0.686 bb	0.599 sb	---	
B2	0.717 sb	0.668 bb	0.751 xx	---

In eerste instantie is het verleidelijk om de xx-correlaties voor een schatting van de topicbetrouwbaarheid te gebruiken. De beoordelaars zijn immers steeds hetzelfde, alleen het topic (hier: de teksthelft) verschilt. Beoordelaars hebben echter van zichzelf een bepaalde onbetrouwbaarheid. Een beoordelaar die dezelfde teksten een tweede maal beoordeelt, kan andere resultaten krijgen dan de eerste keer. Die intra-beoordelaarsonbetrouwbaarheid werkt ook door in de xx-correlaties waardoor deze niet een zuivere schatting van de topicbetrouwbaarheid leveren. Daar de intra-beoordelaarsonbetrouwbaarheid niet bekend is, valt de topicbetrouwbaarheid bij perfect betrouwbare beoordeling via deze correlaties niet te berekenen.

Bij de xx-correlaties worden twee teksthelften vergeleken en bij de sb-correlaties ook. Verder liggen de gemiddelden van beide soorten correlaties dicht bij elkaar. De gemiddelde xx-correlatie is: (0.674+0.751)/2=0.713. De gemiddelde sb-correlatie is: (0.599+0.717)/2= 0.658. Kennelijk maakt het dus in dit geval niet veel uit of de twee teksthelften beoordeeld worden door dezelfde beoordelaar of door een andere beoordelaar.

De xx-correlaties zijn niet goed te passen in het model van Coffman (1966) zoals dat in 9.1 werd besproken, omdat dit model uitgaat van twee empirisch bepaalde gemiddelde correlaties: de gemiddelde correlatie tussen de beoordelaars die hetzelfde topic beoordeelden en de gemiddelde correlatie tussen de beoordelaars die verschillende topics beoordeelden.

De gemiddelde sb-correlatie (0.658) gaat voor de tweede meting uit van een andere teksthelft en van een andere beoordelaar; dit is de scorebetrouwbaarheid van één teksthelft beoordeeld door één beoordelaar. In de score-onbetrouwbaarheid zit beoordelaarsonbetrouwbaarheid en topiconbetrouwbaarheid. Wat zou de correlatie zijn wanneer beide teksthelften perfect betrouwbaar beoordeeld waren? De beoordelaarsbetrouwbaarheid per teksthelft bedroeg 0.677 (het gemiddelde van de bb-correlaties). Via de correctie voor onbetrouwbaarheid valt dan uit te rekenen wat de overeenstemming tussen beide helften zou zijn bij perfect betrouwbare beoordeling (de topicbetrouwbaarheid van één teksthelft bij perfect betrouwbare beoordeling). Dat levert een voor onbetrouwbaarheid gecorrigeerde topicbetrouwbaarheid van 0.972. Hoewel deze waarde door kansfluctuaties relatief hoog kan zijn uitgevallen, lijkt deze waarde toch dermate hoog dat kennelijk beide teksthelften qua aantal fouten PHW overwegend soortgelijke resultaten geven bij betrouwbare beoordeling. Kennelijk maakt de ene teksthelft of de andere teksthelft weinig uit. Een halve A4 tekst lijkt daarmee in beginsel voldoende om het schrijfniveau van een student vast te leggen.

Deze uitkomst valt ook meer rechtstreeks af te leiden uit Tabel 9.2. De gemiddelde waargenomen bb en sb-correlaties (respectievelijk 0.677 en 0.658) zijn ongeveer even hoog. Of beide beoordelaars dezelfde teksthelft beoordelen of een verschillende teksthelft, maakt voor de correlatie kennelijk weinig verschil. Dat lijkt alleen mogelijk te zijn wanneer de aantallen fouten (samengenomen voor beide beoordelaars en na standaardisatie per beoordelaar en per helft) tussen de twee teksthelften in hoge mate overeenstemmen.

In Tabel 9.3 is via de methode van Coffman (1966) de scorebetrouwbaarheid en de beoordelaarsbetrouwbaarheid berekend voor verschillende aantallen topics (1-2) en verschillende aantallen beoordelaars (1-2). De gevonden totale scorebetrouwbaarheid (0.880) voor twee beoordelaars en twee teksthelften wijkt iets af van de op de normale wijze berekende betrouwbaarheid (0.899). Dit is mogelijk doordat in de berekening op de gebruikelijke wijze wordt uitgegaan van de correlatie tussen de somvariabelen van beide teksthelften die daarna via de formule voor testverlenging wordt opgewaardeerd tot de verwachte betrouwbaarheid voor beide teksthelften samen. De berekening volgens Coffman gaat echter uit van twee correlaties (aangegeven met '*') waarbij ook de verhouding tussen die twee wordt gebruikt. Dit maakt deze methode gevoeliger voor kansfluctuaties. Het doel van de methode was echter niet de scorebetrouwbaarheid beter te schatten, maar het verband met aantallen beoordelaars en topics te laten zien. In de tabel zijn op de een na laatste regel de topicbetrouwbaarheden voor één en twee teksthelften te vinden (respectievelijk 0.972 en 0.986) bij perfect betrouwbare beoordeling.

Tabel 9.3 Scorebetrouwbaarheid (bovenste waarde) en beoordelaarsbetrouwbaarheid (onderste waarde) bij verschillende aantallen beoordelaars en tekstdelen volgens de methode van Coffman (1966), * = geobserveerde waarde

aantal beoordelaars	aantal tekstdelen	
	1	2
1	0.658*	0.794
	0.677*	0.805
2	0.785	0.880
	0.807	0.892
oneindig	0.972	0.986
	1.000	1.000

Hier volgt kort de berekeningswijze. Eerst wordt voor kolom '1 topic' de scorebetrouwbaarheid berekend bij perfect betrouwbare beoordeling door de *-correlaties op elkaar te delen. Dit levert de waarde 0.972. Deze waarde moet ook gelden voor de overige verhoudingen tussen scorebetrouwbaarheid en beoordelaarsbetrouwbaarheid in deze kolom. Via de formule voor testverlenging kan de beoordelaarsbetrouwbaarheid van 2 beoordelaars bij 1 topic worden berekend uit die van 1 beoordelaar, 1 topic. Via de scorebetrouwbaarheid van 1 topic dat perfect betrouwbaar is beoordeeld, valt via de formule voor testverlenging ook de betrouwbaarheid van 2 topics bij perfect betrouwbare beoordeling af te leiden. Deze verhouding moet vervolgens weer gelden voor alle waarden in deze kolom. Via de scorebetrouwbaarheid van 1 topic valt via de formule voor testverlenging ook steeds de scorebetrouwbaarheid van 2 topics te berekenen.

Deze uitkomsten laten zien dat voor het betrouwbaar vaststellen van het aantal fouten per honderd woorden waarmee een student schrijft, een tekst van een halve A4 (ongeveer 250 woorden) in beginsel voldoende is. Bij twee beoordelaars is de verwachte betrouwbaarheid dan 0.817. Dit is de verwachte correlatie met een andere halve A4 tekst geschreven door dezelfde studenten en nagekeken door twee beoordelaars.

9.3 Is een enkele, korte tekst voldoende?

In de voorgaande paragraaf werd voor het aantal fouten PHW in een tekst van een halve A4 een topicbetrouwbaarheid van afgerond 0.97 gevonden, uitgaande van perfect betrouwbare beoordeling. Dit betekent dat op basis van een halve A4 tekst het aantal fouten PHW in een andere halve A4 tekst van dezelfde student vrijwel perfect voorspeld kan worden (bij inzet van voldoende beoordelaars). Voor het bepalen van het niveau van basale schrijfvaardigheid via het aantal fouten PHW lijkt daarmee een tekst ter lengte van een halve A4 in beginsel voldoende te zijn. Basale schrijfvaardigheid is dan niet moeilijk vaststelbaar, maar in beginsel vrij snel en eenvoudig.

Daar de in dit onderzoek gevonden waarde van 0.97 de verhouding vormt tussen twee correlaties die beide gebaseerd zijn op een beperkte steekproef van 48 teksten, is het mogelijk dat door toevalsfluctuaties bij een herhaald onderzoek een lagere waarde gevonden zal worden voor de topicbetrouwbaarheid van het aantal fouten PHW. Ook via een kansmodel gebaseerd op de uitkomsten van het foutenonderzoek (zie deelstudie 1, paragraaf 4.3) vonden we echter een zeer hoge waarde voor de topicbetrouwbaarheid van het aantal fouten PHW. Uitgaande van het gemiddelde aantal bevestigde fouten (16.1) PHW voor eerstejaars hbo-studenten en de SD (6.9) werd uitgaande van een normaalverdeling (de gevonden verdeling was inderdaad bij benadering normaal) via SPSS (versie 20) een gesimuleerde groep van tweeduizend studenten gegenereerd met ieder hun eigen schrijfvaardigheidsniveau (de kans op een bevestigde fout per woord). Een klein aantal studenten (zestien) kreeg hierbij een negatieve kans en werd bij de verdere berekening buiten beschouwing gelaten. Uitgaande van een binomiaalverdeling kon vervolgens voor iedere student het aantal fouten in twee teksten van 250 woorden gegenereerd worden. Dit leverde een topicbetrouwbaarheid (de correlatie tussen het aantal fouten in beide teksten) op van 0.888. Ook dit model leidde daarmee tot een zeer hoge topicbetrouwbaarheid voor het aantal fouten PHW.

Het model gaat per gesimuleerde student uit van twee parameters: het niveau van de basale schrijfvaardigheid uitgedrukt in het aantal fouten per woord dat de student gemiddeld maakt en het aantal woorden dat de tekst telt, twee voor de hand liggende assumpties. Verder bevat het model een toevalscomponent. Op basis van beide parameters valt het aantal fouten niet precies te voorspellen, maar slechts bij benadering. Dat uitgegaan wordt van een binomiaalverdeling (een vaste kans op een fout per woord) betekent niet dat verondersteld wordt dat de student daadwerkelijk afzonderlijke fouten maakt op basis van een toevalspro-

ces. Wel, wordt verondersteld dat het totaal aantal fouten in een tekst zich laat beschrijven via dit kansmodel.

Tegen dit model zou men kunnen inbrengen dat het uitgaat van een per student constant schrijfvaardigheidsniveau waardoor het vervolgens weinig verwonderlijk is, dat dit resulteert in een hoge topicbetrouwbaarheid. Deze assumptie is echter in overeenstemming met het gegeven dat basale schrijfvaardigheid met objectieve tests en via de TAVAN-score zeer betrouwbaar kan worden vastgesteld, terwijl de veronderstelling dat het basale schrijfvaardigheidsniveau van een student per topic belangrijk zou veranderen, zonder overtuigende data, vrij extreem lijkt.

Bij holistische beoordeling is een halve topictekst echter volstrekt niet voldoende. In het onderzoek van Godshalk et al. (1966) was de topic-betrouwbaarheid bij perfect betrouwbare beoordeling 0.681. Om een topicbetrouwbaarheid van 0.90 te bereiken, zouden dan - bij perfect betrouwbare beoordeling - vijf topicteksten nodig zijn volgens de formule voor testverlenging. Dit is voor de praktijk een belangrijk verschil met de eerder gevonden benodigde tekstlengte van een halve A4.

Men kan argumenteren dat het aantal fouten per honderd woorden een belangrijk andere maat is om de basale schrijfvaardigheid vast te stellen dan het holistische oordeel en dat het dus begrijpelijk is dat bij beoordeling op aantal fouten PHW volstaan kan worden met minder tekst. Uit het fouteneffect-onderzoek bleek echter dat beide maten zeer hoog correleren (ongecorrigeerd: -0.74, na correctie voor onbetrouwbaarheid: -0.89) en daarmee in hoge mate dezelfde uitkomsten leveren. Wanneer een half essay voldoende informatie bevat om de basale schrijfvaardigheid via het aantal fouten per honderd woorden te bepalen, zou men verwachten dat dit in beginsel via holistische beoordeling ook zou moeten kunnen. De vraag die dan rijst, is waarom dit bij holistische beoordeling niet lukt. Waarom valt de topicbetrouwbaarheid bij holistische beoordeling belangrijk lager uit dan bij beoordeling via het aantal fouten PHW?

Een eerste mogelijkheid om het verschil in topicbetrouwbaarheid te verklaren, leek te zijn dat de betrouwbaarheden mogelijk verschillend berekend waren. Zoals reeds in paragraaf 9.1 uitgebreider werd opgemerkt, kunnen er veel verschillende intraclasscorrelaties (betrouwbaarheden) berekend worden, die onderling uiteenlopende waarden kunnen geven, terwijl onderzoekers zich dit vaak niet bewust zijn en vaak ook niet precies vermelden wat en hoe ze hun betrouwbaarheid berekend hebben.

In het onderzoek van Godshalk et al. (1966, p. 12) werd een minder gangbaar variantie-analytisch design gebruikt voor het berekenen van de verwachte scorebetrouwbaarheid waardoor niet onmiddellijk duidelijk is welke betrouwbaarheid precies berekend werd. Via de in Godshalk et al. (1966) vermelde correlatiematrix (p. 53) bleek de gemiddelde onderlinge correlatie tussen de topicscores 0.515 te bedragen. Via de formule voor testverlenging zou men dan een totale scorebetrouwbaarheid van 0.841 verwachten voor een test bestaande uit deze vijf items, wat ook de waarde was die Godshalk et al. rapporteerden. Deze waarde lijkt daarmee vergelijkbaar met de door ons gerapporteerde betrouwbaarheid van 0.899 gebaseerd op twee 'items' (teksthelften). Godshalk et al. gebruikten vijf beoordelaars per topic en wij twee, maar in de berekening van de totale scorebetrouwbaarheid spelen die verder geen rol meer. Hun onbetrouwbaarheid zit al verwerkt in de topicscores. Het verschil in uitkomsten kan dus kennelijk niet verklaard worden door een verschil in de berekening van de scorebetrouwbaarheid.

Het probleem is niet dat de holistische beoordeling relatief onbetrouwbaar is, want dat valt op te heffen door veel beoordelaars in te zetten, maar dat uitgaande van holistische beoordeling zelfs bij perfect betrouwbare beoordeling meerdere topicteksten benodigd zijn. De correlatie tussen de topicscores is, zelfs bij perfect betrouwbare beoordeling, relatief laag. In het onderzoek van Godshalk et al. (1966) was de topic-betrouwbaarheid (de gemiddelde onderlinge correlatie tussen de topicscores na correctie voor beoordelaarsonbetrouwbaarheid) bij perfect betrouwbare beoordeling 0.68. Dit betekent dat van de variantie in de perfect betrouwbare topicscore slechts 68% gemeenschappelijk was met de 'ware' holistische score; 32% van de variantie was wel betrouwbaar (de beoordelaars waren het onderling eens), maar kwam niet terug in de holistische totaalscore. Ongeveer een derde van de *betrouwbare* variantie werd daarmee weggefilterd en had kennelijk betrekking op specifieke kenmerken van de teksten die wisselden per topic en per student.

Deze student-topic interacties ontstonden doordat de student 'vrij' moest schrijven over een verplicht onderwerp. Het is gemakkelijker te schrijven over een onderwerp waar men veel van weet, dan waar men weinig van weet. Een student zal daardoor op het ene topic beter presteren dan op het andere topic. Door veel topics te gebruiken hoopten Godshalk et al. deze topic-onbetrouwbaarheid te beteugelen. "Thus, an individual's rating would not depend on whether or not he could find something interesting and accurate to say on a single topic which he might never have considered before" (Godshalk et al., 1966, p. 4).

Men kan zich afvragen of dit niet een erg omslachtige manier van werken was. Als we niet willen dat de inhoud van het topic een rol gaat spelen in het eindresultaat, omdat we willen weten of iemand kan schrijven (formuleren), is het dan niet beter de taak aan te passen en alle studenten dezelfde inhoud te geven of een onderwerp op te geven, dat iedereen even goed kent? Bij het onderzoek naar het effect van het nieuwe TAVAN-programma (Deelstudie 3) kregen de studenten bij de pretest en de posttest een tekst met fouten te herschrijven. Bij het fouteneffect-onderzoek schreven de studenten een tekst over hun ervaringen met het TAVAN-programma (TAVAN2). Op beide manieren was het mogelijk de inhoud van de tekst per student zo gelijk mogelijk te houden.

De topiconbetrouwbaarheid werd toegeschreven aan de student. De uitkomsten van de variantie-analyse leken dit aan te tonen. De student-topic interactie was significant en vormde de topiconbetrouwbaarheid (Godshalk et al., 1966, p. 12). Holistische beoordeling was bekend onbetrouwbaar, maar een deel van de onbetrouwbaarheid werd niet veroorzaakt door de beoordelaar, maar doordat de student die bij het ene topic een totaal andere schrijfvaardigheid toonde dan bij het andere topic. Althans deze conclusie trok men. Godshalk et al. (1966, p. 4) merkten op:

> *At the time the study was designed, it was known that the unreliability of essay tests came from two major sources: the differences in quality of student writing from one topic to another, and the differences among readers in what they consider the characteristics of good writing.*

Een belangrijk punt dat Godshalk et al. en later ook anderen zich niet realiseerden, is dat naar alle waarschijnlijkheid ook de beoordelaars sterk werden beïnvloed door de inhoud van de essays. Iemand die weet waarover hij schrijft, maakt gemakkelijk een betere indruk op beoordelaars dan iemand die onzin debiteert, ook al is die onzin misschien perfect verwoord. Een beoordelaar is geneigd een artikel dat zijn opvattingen bevestigt, positiever te waarderen dan een artikel dat zijn opvattingen tegenspreekt. De student-topic interactie is daarmee vermoedelijk in belangrijke mate ook een beoordelaarseffect. Het idee bij holistische beoordeling is juist dat de beoordelaar zich door het totaal van de tekst laat beïnvloeden (Camara, 2003, p. 1). Het onvermijdelijke gevolg daarvan lijkt te zijn dat de inhoud van de tekst en eventuele andere kenmerken het holistische oordeel in belangrijke mate zullen beïnvloeden.

Overigens geldt dit argument in beginsel ook bij niet-holistische, analytische beoordeling. Alle in de tekst aanwezige componenten zullen in beginsel invloed (kunnen) uitoefenen op

de menselijke beoordelaar, zelfs wanneer die de expliciete opdracht krijgt alleen op bepaalde aspecten te letten. Zo valt bij een beoordeling op bijvoorbeeld alleen spelfouten in de tekst niet uit te sluiten dat de beoordelaar zich onbewust ook zal laten beïnvloeden door de inhoud van de tekst. Beoordelaars zullen immers in de praktijk niet alle fouten zien en soms fouten zien die er in feite niet zijn. Een inhoudelijke 'fout' kan er dan toe leiden dat de beoordelaars meer spelfouten gaan zien die aanwezig zijn en daarnaast ook nog spelfouten gaan signaleren die niet aanwezig zijn, terwijl een inhoud die goedkeuring en enthousiasme oproept, gemakkelijk kan leiden tot een minder kritische beoordeling.

Hoewel het principe dat ieder aspect dat in de tekst aanwezig is de beoordeling kan beïnvloeden, mogelijk vrij vanzelfsprekend lijkt en sterk doet denken aan het halo-effect bij persoonsbeoordeling (Thorndike, 1920), konden wij geen duidelijke bevestiging vinden voor het bestaan van een 'gegeneraliseerd' halo-effect (niet betrekking hebbend op personen). Wel laat het Stroop-effect zien, dat menselijke beoordelaars aangeleerde reacties op verbale stimuli niet eenvoudig kunnen onderdrukken en dat deze reacties sterk kunnen interfereren met de opgedragen beoordelingstaak.

Bij het Stroop-effect (Stroop, 1935) moet de beoordelaar de kleur waarin het woord afgebeeld is (de kleur van de inkt), benoemen, terwijl het woord zelf een andere kleur aangeeft. (Het woord 'ROOD' in groene inkt moet de response 'groen' opleveren.) In deze situatie heeft de beoordelaar de neiging het woord dat hij ziet, uit te spreken, terwijl de opdracht juist is de kleur van de inkt te zeggen. De twee factoren 'woord' en 'kleur van de inkt' werken elkaar tegen.

Omdat Godshalk et al. (1966) de student-topic interactie volledig toeschreven aan de student, een veronderstelling die overigens nog steeds gangbaar en gebruikelijk is, was de gevolgtrekking dat de schrijfvaardigheid van de student van topic tot topic sterk moest fluctueren. Anders leek niet verklaarbaar dat het holistische oordeel per student van topic tot topic sterk kon verschillen.

Stel dat een student een essay moet schrijven over inflatie. Een student die weet, wat dat is en in de materie thuis is en die verder goed kan schrijven, zal daar moeiteloos een verhandeling over schrijven. Een student die echter niet weet wat de term betekent, zich niet vertrouwd voelt op economisch gebied, maar wel goed kan schrijven, zal nu gemakkelijk inhoudelijke fouten maken, hoewel hij misschien prima formuleert. Beoordelaars hebben ver-

moedelijk normaal de neiging vooral te lezen op inhoud, terwijl het taalgebruik mogelijk past opvalt, zodra het afwijkt van wat de beoordelaar als normaal ziet. Op inhoudelijke fouten wordt volgens de eerder gegeven samenvatting van EEG-onderzoek op dit gebied (zie deelstudie 4, paragraaf 7.1) gereageerd met een N400-piek. De bedoeling van de schrijfopdracht is echter niet de economische kennis van de student te toetsen, maar alleen vast te stellen hoe goed de student kan schrijven. Het gebrek aan inhoudelijke kennis van de student kan daardoor bij holistische beoordeling onbedoeld doorwerken in de beoordeling van de basale schrijfvaardigheid.

De enige manier om er met zekerheid achter te komen waardoor de student-topic interactie veroorzaakt wordt (student, beoordelaar of een combinatie van beiden), is het toepassen van een andere meetmethode dan het holistische oordeel om de kwaliteit van de teksten onafhankelijk van de holistische beoordelaars vast te stellen. Godshalk et al. (1966) beschikten echter niet over een dergelijke methode.

Een bijkomend argument was dat deze situatie vergelijkbaar was met die van objectieve studietoetsen waar de student-item interactie ook de oorzaak vormt van de score-onbetrouwbaarheid. Een normale studietoets bestrijkt een groot gebied van mogelijke items, maar de items in de toets vormen slechts een kleine steekproef uit het grote aantal van alle mogelijke items. Een student zal bepaalde items niet weten en andere wel, waardoor hij bij een toets geluk kan hebben (er werd gevraagd wat hij wist) of pech (er werd toevallig vooral gevraagd over onderwerpen die hij niet goed bestudeerd had). Voor studietoetsen is inhoudsonbetrouwbaarheid (topiconbetrouwbaarheid) normaal en daarom werd verondersteld dat dit ook bij het meten van schrijfvaardigheid een normaal verschijnsel zou zijn.

Men kan zich echter afvragen of de vergelijking met een studietoets over de stof in een studieboek in het geval van basale schrijfvaardigheid wel opgaat. Om vast te stellen of iemand Engels spreekt, is het niet nodig een aantal gesprekken te voeren, een korte interactie volstaat. Voor andere complexe vaardigheden als bijvoorbeeld autorijden lijkt iets soortgelijks te gelden. Het lijkt dus moeilijk voorstelbaar dat een student bij het ene topic goed zou kunnen schrijven, maar dat bij een volgend topic opeens niet meer zou kunnen.

Het uitgangspunt bij holistische beoordeling is ook dat schrijven een vaardigheid is, waarvan de beheersing onmiddellijk via het waargenomen resultaat kan worden vastgesteld. Het

idee van topic-onbetrouwbaarheid lijkt daarmee in tegenspraak te zijn. Bij een kennistoets zijn alle items voortdurend belangrijk anders, maar bij een vaardighedentoets gaat het steeds om dezelfde vaardigheid in iets andere situaties, waardoor topic-onbetrouwbaarheid veel minder een rol speelt.

Op basis van de door ons gevonden resultaten rijst de vraag of topic-onbetrouwbaarheid bij het meten van schrijfvaardigheid wel een noodzakelijk verschijnsel is zoals altijd werd aangenomen? Wat is precies de empirische basis om te denken dat topic-onbetrouwbaarheid bestaat? En als topic-onbetrouwbaarheid bestaat, waar wordt die dan precies door veroorzaakt?

De correlatiematrix van de holistische beoordelingen van de topics uit het onderzoek van Godshalk et al. (1966) bevatte twee soorten correlaties, namelijk tussen beoordelaars die verschillende topics beoordeelden en tussen beoordelaars die hetzelfde topic beoordeelden. Het gemiddelde van de eerste groep correlaties (de scorebetrouwbaarheid, de correlatie tussen een topic beoordeeld door een beoordelaar met een ander topic beoordeeld door een andere beoordelaar) bedroeg 0.263. Het gemiddelde van de tweede groep correlaties (de beoordelaarsbetrouwbaarheid, de correlatie tussen twee beoordelaars die hetzelfde topic beoordeelden) bedroeg 0.386 (Coffman, 1966, p. 154, Table 3).

Dat de beoordelaarsbetrouwbaarheid hoger uitvalt dan de scorebetrouwbaarheid valt te verwachten. Beoordelaars zullen over dezelfde tekst meer overeenstemmen dan over twee verschillende teksten. Uit de verhouding van de twee betrouwbaarheden wordt de topicbetrouwbaarheid bij perfect betrouwbare beoordeling afgeleid: 0.263/0.368=0.68. Wanneer we topics perfect betrouwbaar zouden laten beoordelen (oneindig veel beoordelaars), zou de gemiddelde correlatie tussen de topicscores toch niet hoger uitkomen dan 0.68.

Wanneer we er echter van uitgaan dat het taalvermogen van studenten niet van topic tot topic zal fluctueren, betekent de veel hogere overeenstemming tussen de beoordelaars die hetzelfde topic beoordelen, dat ze er kennelijk niet in slagen volledig te focussen op de basale schrijfvaardigheid van de student, maar zich laten afleiden door de inhoud van het essay. De topicvariantie ontstaat kennelijk, doordat de holistische beoordelaar beïnvloed wordt door de inhoud van het essay. Het is normaal moeilijk inhoud en taalgebruik strikt te scheiden. Bovendien is het ook zeer de vraag of holistische beoordelaars een dergelijke splitsing nastreven. Een belangrijk idee bij holistische beoordeling is immers dat de tekst als geheel moet worden beoordeeld.

Valt er empirische evidentie voor dit standpunt te vinden? In paragraaf 8.5 werd vermeld dat de qua criteriumvaliditeit beste objectieve tests van Godshalk et al.(1966) om schrijfvaardigheid te meten na correctie voor onbetrouwbaarheid .87 correleerden met de holistische totaalscore. Het lijkt weinig plausibel om aan te nemen dat die via de objectieve tests vastgestelde schrijfvaardigheid vervolgens per topic zou kunnen gaan fluctueren. Het lijkt daarmee plausibeler dat de topic-onbetrouwbaarheid een bijproduct is van het holistische beoordelingsproces en mogelijk vooral ontstaat doordat de holistische beoordelaars zich laten beïnvloeden door de inhoud van de teksten.

Een tweede argument dat de topic-onbetrouwbaarheid vermoedelijk vooral veroorzaakt wordt door de methode van holistisch beoordelen, is het eerder door ons gerapporteerde resultaat voor de topicbetrouwbaarheid bij beoordeling op aantal fouten per honderd woorden van afgerond: 0.97 voor een tekst ter lengte van een halve A4. Hoewel niet valt uit te sluiten dat deze waarde door steekproeffluctuaties wat erg hoog is uitgevallen, wijst deze waarde er wel op dat basale schrijfvaardigheid gemeten via een korte tekst (ongeveer 250 woorden) in beginsel redelijk betrouwbaar kan worden vastgesteld. Studenten schrijven kennelijk met een vrij constante kwaliteit.

Een derde argument is dat ook het eerder besproken kansmodel dat slechts uitgaat van een bepaald schrijfvaardigheidsniveau per student en van het aantal woorden in de tekst, tot een zeer hoge topicbetrouwbaar leidt. Ook volgens dit model zou daarmee een korte tekst volstaan voor het bepalen van de basale schrijfvaardigheid.

De verklaring voor de lage topicbetrouwbaarheid van holistische beoordeling is dan als volgt. Ongeveer twee derde van de betrouwbare variantie in de topicscore heeft betrekking op het taalgebruik en eventueel de inhoud van het essay voor zover die correleert met het taalgebruik. De overige één derde van de betrouwbare variantie in de topicscore wordt veroorzaakt door toevallige inhoudsaspecten en eventueel andere toevallige aspecten van het essay. Wanneer hetzelfde essay door meerdere beoordelaars wordt beoordeeld, wordt de beoordeling wel betrouwbaarder, maar blijft deze systematische fout (resulterend in één derde van de betrouwbare variantie) aanwezig doordat gemiddeld genomen alle beoordelaars zich door de niet-talige aspecten van het essay op soortgelijke wijze laten beïnvloeden. Door meerdere topics samen te voegen (de scores te middelen of op te tellen) middelt deze specifieke inhoudscomponent echter uit.

Het lijkt niet plausibel om te veronderstellen dat de ongeveer één derde betrouwbare variantie die gebonden zit aan de student-topic interactie alle inhoudsvariantie vormt. Vermoedelijk zijn taalgebruik en keuze van een effectieve en goede inhoud sterk gecorreleerd. Het ene essay vormt een sterker verhaal of betoog dan het andere, maar de studenten met de betere taalbeheersing zullen vaak ook de betere inhoud construeren en selecteren. Met andere woorden: het lijkt waarschijnlijk dat de holistische beoordelaar zich sterk laat beïnvloeden door de inhoud van de essays, maar doordat het oordeel over die inhoud hoog correleert met de taalbeheersing van de student, kunnen objectieve tests de inhoud van de geschreven essays probleemloos negeren. De holistische beoordelaars laten zich echter door de inhoud van de essays wel beïnvloeden, waardoor ongeveer één derde van de betrouwbare variantie in hun oordeel niet langer betrekking heeft op het taalgebruik van de student, maar op de specifieke inhoud van het desbetreffende essay.

Of deze verklaring juist is, valt zonder verder onderzoek niet met zekerheid te stellen. Het is immers niet bekend hoe het holistische oordeel precies tot stand komt. Experimenteel onderzoek op dit punt ontbreekt grotendeels nog. Wel lijkt duidelijk en aangetoond dat het aantal fouten per honderd woorden sterk gecorreleerd is met het holistische oordeel en dit vermoedelijk ook voor een belangrijk deel bepaalt. Een vraag voor verder onderzoek is of valt aan te tonen dat naast het aantal fouten per honderd woorden ook (de kwaliteit van) de inhoud van de essays van invloed is op het holistische oordeel. Hoewel het moeilijk voorstelbaar lijkt dat dit niet het geval zou zijn, lijkt het wel van belang dit verband daadwerkelijk aan te tonen.

9.4 Samenvatting, conclusies en discussie

Lastige kwantificering betrouwbaarheid en soorten betrouwbaarheid

Betrouwbaarheid is een terugkerend probleem bij het meten van schrijfvaardigheid. De kwantificering van betrouwbaarheid blijkt echter minder simpel dan vaak wordt aangenomen. Er blijken veel verschillende soorten betrouwbaarheid te bestaan, terwijl er bij het berekenen gemakkelijk iets mis kan gaan. Vooral de situatie dat een aantal beoordelaars een aantal topics heeft beoordeeld, kan problemen geven. Wij volgden de oplossing die Godshalk et al. (1966) ook hebben gekozen, maar beschreven die in termen van correlatie.

De totale scorebetrouwbaarheid wordt gevonden via de waargenomen gemiddelde correlatie tussen de topicscores en wordt via de formule voor testverlenging gecorrigeerd voor het aantal topics. De beoordelaarsonbetrouwbaarheid zit bij deze benadering al verwerkt in de topicscores. Het voordeel van deze benadering is dat deze manier voor de berekening van de betrouwbaarheid volledig vergelijkbaar is met de gebruikelijke berekening (coëfficiënt alfa gestandaardiseerd) bij meerkeuze-toetsen.

Bij het werken met beoordelaars kan de beoordelaarsbetrouwbaarheid van een topic berekend worden. Dit is de gemiddelde onderlinge correlatie tussen de beoordelaars van een zelfde topic die vervolgens via de formule voor testverlenging gecorrigeerd wordt voor het aantal beoordelaars per topic. De scorebetrouwbaarheid van het topic is de gemiddelde onderlinge correlatie met de andere topicscores. Deze twee gegevens maken het mogelijk de topicbetrouwbaarheid bij perfect betrouwbare beoordeling uit te rekenen: de scorebetrouwbaarheid gedeeld door de beoordelaarsbetrouwbaarheid.

Het resultaat is dat het mogelijk is de scorebetrouwbaarheid van de topics op te splitsen in twee factoren: de beoordelaarsbetrouwbaarheid en de topicbetrouwbaarheid bij het desbetreffende aantal beoordelaars. Een deel van de onbetrouwbaarheid wordt veroorzaakt door de beoordelaars en een deel door de topics, is men doorgaans geneigd te veronderstellen. Deze veronderstelling blijkt echter discutabel gezien het grote verschil in topicbetrouwbaarheid tussen aantal fouten PHW en holistische beoordeling.

Hoeveelheid benodigde tekst voor meten schrijfvaardigheid

Voor de 48 teksten uit het fouteneffect-onderzoek was het mogelijk de teksten op te splitsen in twee (qua aantal woorden) even lange (of ongeveer even lange in het geval van een oneven aantal) helften. Door vervolgens voor iedere beoordelaar per teksthelft het aantal fouten PHW te tellen kon via de scorebetrouwbaarheid en de beoordelaarsbetrouwbaarheid de topicbetrouwbaarheid bij perfect betrouwbare beoordeling berekend worden. Deze bleek afgerond 0.97 te bedragen en daarmee zeer hoog te zijn. Deze hoge waarde duidt erop dat een tekst met een lengte van een halve A4 (250 woorden) in beginsel voldoende informatie bevat om via het aantal fouten PHW de basale schrijfvaardigheid van een student zeer betrouwbaar vast te stellen.

Daar de gevonden waarde erg hoog is en door steekproeffluctuaties mogelijk relatief hoog is uitgevallen, hebben we de topicbetrouwbaarheid ook bepaald door de gegevens die bij het foutenonderzoek waren gevonden over de schrijfvaardigheid van eerstejaars hbo-studenten in een simulatiemodel in te voeren. Dit kansmodel leverde een topicbetrouwbaarheid van afgerond 0.89 op. Ook via deze methode werd daarmee een zeer hoge waarde gevonden.

Deze twee uitkomsten lijken duidelijk in contrast te staan met wat bekend is over de topicbetrouwbaarheid bij holistische beoordeling. Zo vonden Godshalk et al. (1966) een topicbetrouwbaarheid bij perfect betrouwbare beoordeling van afgerond 0.68. Om een betrouwbaarheid van 0.90 te bereiken, zouden dan volgens de formule voor testverlenging vijf topicteksten per student benodigd zijn. Een belangrijk andere waarde dan een tekst van een halve A4.

Verklaring voor de discrepantie in topicbetrouwbaarheid

Onze verklaring voor dit grote verschil in topicbetrouwbaarheid tussen beide beoordelingsmethodes is dat de relatief lage topicbetrouwbaarheid bij holistische beoordeling voor een groot deel een beoordelaarseffect is, dat voorheen ten onrechte volledig werd toegeschreven aan de student. Om de lage topicbetrouwbaarheid bij holistische beoordeling te verklaren, werd verondersteld dat de schrijfvaardigheid van topic tot topic sterk zou fluctueren. Voor deze veronderstelling bestaan geen andere gronden dan de lage overeenstemming tussen topicscores bij holistische beoordeling, terwijl er wel goede argumenten zijn om aan te nemen dat de schrijfvaardigheid van een student een relatief stabiel kenmerk moet zijn.

Het lijkt daarom 'logischer' en eenvoudiger te veronderstellen dat de holistische beoordelaar zich door toevallige inhoudsaspecten van de tekst laten beïnvloeden ('meeslepen'). De beoordelaar moet de tekst beoordelen, maar heeft geen expliciete instructie. Bij het lezen van de tekst gaat het strikt genomen vooral om de effectiviteit van het taalgebruik. Tegelijkertijd gaat de tekst ook over een bepaald onderwerp en zijn beoordelaars normale lezers die geneigd zijn op inhoud en betekenis te lezen. Er werken daardoor twee verschillende factoren tegelijkertijd in op de beoordelaar. De beoordelaar moet idealiter vooral focussen op het taalgebruik, maar het is moeilijk niet op de inhoud van de tekst te reageren en door de inhoud 'meegesleept' te worden. De reactie op de inhoud kan vervolgens interfereren met de

reactie op het taalgebruik. Zodra de beoordelaar door de instructie meer gericht wordt op het taalgebruik door te beoordelen op fouten in de tekst verdwijnt dit inhoudseffect kennelijk grotendeels.

Bij beoordeling op aantal fouten PHW doet dit probleem zich kennelijk niet of amper voor, vermoedelijk door de expliciete instructie, waardoor in dat geval de topicbetrouwbaarheid belangrijk hoger uitvalt zodat in dat geval een halve A4 voldoende lijkt om de basale schrijfvaardigheid van een student betrouwbaar vast te stellen.

Conclusies

1a. Bij holistische beoordeling met meerdere beoordelaars dient de scorebetrouwbaarheid van de topics bepaald te worden via de gemiddelde onderlinge correlatie van de topic(totaal)scores. De onbetrouwbaarheid van de beoordelaars zit hierbij al verwerkt in de topicscores.
1b. De beoordelaarsbetrouwbaarheid dient bepaald te worden op basis van de gemiddelde onderlinge correlatie tussen de beoordelaars van dezelfde topics. Via de formule voor testverlenging kan vervolgens de betrouwbaarheid berekend worden voor het gebruikte aantal beoordelaars.
1c. De topicbetrouwbaarheid bij perfect betrouwbare beoordeling is vervolgens de verhouding tussen de scorebetrouwbaarheid en de beoordelaarsbetrouwbaarheid.

2. Voor het betrouwbaar vaststellen van het niveau van basale schrijfvaardigheid via het aantal fouten PHW lijkt een tekst ter lengte van een halve A4 (250 woorden) voldoende te zijn. Dit is belangrijk minder dan tot nu toe bij holistische beoordeling het geval was.

Hoewel wij een zeer hoge waarde vonden voor de topicbetrouwbaarheid, valt niet volledig uit te sluiten dat deze waarde, gebaseerd op de verhouding tussen twee via de steekproef te bepalen correlaties, belangrijk te hoog is uitgevallen. Een verdeling van één tekst in twee helften is verder minder overtuigend dan een waarde gebaseerd op twee echt verschillende teksten. Ook de hoge waarde die in de simulatie gevonden werd, kan dit probleem niet helemaal oplossen, omdat daar werd uitgegaan van de veronderstelling dat de basale schrijfvaardigheid per student een constante parameter was en niet van topic tot topic zou variëren. Voor deze uitkomst is derhalve bevestiging via verder onderzoek wenselijk.

3. De vermoedelijke reden dat de topicbetrouwbaarheid bij holistische beoordeling laag uitvalt, is niet dat de schrijfvaardigheid van de studenten van topic tot topic sterk fluctueert, maar dat de holistische beoordelaars zich laten beïnvloeden door de inhoud van de teksten waardoor interferentie optreedt met de beoordeling van het taalgebruik. De topicbetrouwbaarheid bij perfect betrouwbare beoordeling komt hierbij overeen met de proportie van de betrouwbare variantie in de holistische beoordeling die betrekking heeft op het taalgebruik van de student.

De enige evidentie tot nu voor de sterk variabele schrijfvaardigheid die men veronderstelt, bestaat uit het sterk fluctueren van de holistische beoordeling per student van topic tot topic. Doordat het holistische oordeel echter niet als erg betrouwbaar bekend staat, lijkt het plausibel dat dit een beoordelaarseffect is. Objectieve testscores, linears en de TAVAN-score vertonen deze sterke fluctuaties niet. Via het aantal fouten PHW is de veronderstelde sterk fluctuerende schrijfvaardigheid tussen topics echter eenvoudig te controleren. Ook dit vormt daarmee een punt voor verder onderzoek.

Een belangrijke conclusie van dit tweede deel van het deelonderzoek naar het meten van basale schrijfvaardigheid is, dat voor het beoordelen van de basale schrijfvaardigheid via de methode van het aantal fouten PHW een tekst ter lengte van een halve A4 (250 woorden) in de praktijk vaak voldoende zal zijn voor een redelijk betrouwbaar oordeel.

Dit lijkt in sterk contrast te staan met het aantal benodigde topics voor een even betrouwbaar holistisch oordeel. De kennelijke verklaring is dat holistische beoordelaars zich door de inhoud van de teksten laten afleiden van het te beoordelen taalgebruik.

Een consequentie zou kunnen zijn, dat het bij beoordelen van basale schrijfvaardigheid aanbeveling verdient de beoordelaar explicieter te richten op de beoordeling van het taalgebruik. Een andere mogelijkheid is de inhoud van de teksten gelijk te trekken door de studenten een slecht geschreven tekst met veel fouten te laten bewerken.

10

Samenvatting, conclusies en nabeschouwing

10.1 Korte samenvatting

Er zijn veel berichten over de tekortschietende schrijfvaardigheid van eerstejaars hbo-studenten. Het doel van het onderzoek was schrijfvaardigheid te kwantificeren en te remediëren.

Schrijfvaardigheid werd gemeten via het aantal bevestigde fouten (fouten gesignaleerd door ten minste twee onafhankelijke beoordelaars) per A4 (500 woorden). Eerstejaars hbo-studenten maken gemiddeld 81 bevestigde fouten per A4. Universitaire studenten maken er 'slechts' 42. Het voorafgaande schrijfonderwijs is kennelijk weinig effectief.

Aan methodes om tekortschietende schrijfvaardigheid te remediëren, ontbrak het niet. Zeventien papieren en negen digitale methodes werden gelokaliseerd en beoordeeld. Geen enkele methode bleek een duidelijke doelstelling te hebben. Geen enkele methode bleek empirisch onderzocht te zijn op effectiviteit. De verschillende methodes richten zich op een veelheid van taalproblemen, maar vaak niet op de fouten die studenten werkelijk maken.

Een nieuw ontwikkeld programma TAVAN (TAalVAardigheid Nieuw) is bij een groep eerstejaars hbo-studenten op effectiviteit onderzocht. De TAVAN-groep bleek in de eindtekst 20% minder fouten per honderd woorden te maken dan in de begintekst. De controlegroep die het traditionele onderwijsprogramma volgde, verbeterde niet. Het TAVAN-programma is daarmee zeer effectief.

TAVAN werkt met een online-programma dat feedback geeft zodra de student een zin herschreven heeft. TAVAN werkt niet met meerkeuzevragen, de student moet zelf formuleren. De score van studenten in het programma bleek een goede voorspeller van het aantal fouten dat de student per A4 maakt. Het programma traint niet alleen, het meet tegelijkertijd het niveau.

Maken fouten uit voor hoe een tekst overkomt? Teksten zonder fouten werden beoordeeld met gemiddeld 48, dezelfde teksten met fouten scoorden gemiddeld 30 (op een schaal van 0 tot 100). Dit laat zien dat fouten een zeer grote invloed hebben op het oordeel van lezers.

Het onderzoek leverde in totaal drie nieuwe methodes op voor het meten van schrijfvaardigheid. De eerste twee waren: het aantal fouten per A4 en de TAVAN-score (de

score behaald in het TAVAN-programma). Deze nieuwe methodes bleken hoog te correleren met het holistische oordeel van beoordelaars en dezelfde factor te meten: vaardigheid van studenten in het opsporen en corrigeren van fouten.

10.2 Samenvatting en conclusies

Inleiding

Het probleem dat in dit proefschrift centraal staat, is de tekortschietende schrijfvaardigheid van met name eerstejaars hbo-studenten. Over de tekortschietende schrijfvaardigheid van eerstejaarsstudenten zijn veel berichten te vinden, maar weinig harde gegevens. Verder blijkt de klacht dat de schrijfvaardigheid van jongeren tekortschiet, van alle tijden. Ook blijkt de klacht in andere landen voor te komen.

De reden dat er weinig harde (kwantitatieve) gegevens beschikbaar zijn over de schrijfvaardigheid van eerstejaars hbo-studenten, is dat schrijfvaardigheid moeilijk meetbaar is. De meest gangbare methode die docenten gebruiken om de schrijfvaardigheid van hun studenten vast te stellen is holistische beoordeling. Bij holistische beoordeling wordt een door een student geschreven tekst snel doorgenomen en becijferd. Deze methode is wel bruikbaar om te zorgen dat studenten schrijfopdrachten maken, maar heeft als meetmethode een aantal belangrijke bezwaren. Allereerst stemmen beoordelaars vaak amper overeen. Wat de ene beoordelaar een goede tekst vindt, vindt de volgende een slechte tekst. Een tweede probleem is dat de ene beoordelaar kritischer is dan de andere en daardoor gemiddeld lager of hoger becijfert. Een derde probleem is dat beoordelaars verschillen qua standaarddeviatie: de ene beoordelaar blijft dichter rond zijn eigen gemiddelde dan de andere. Verder is de methode arbeidsintensief en kunnen beoordelaars geleidelijk strenger of minder streng worden.

Een tweede methode om schrijfvaardigheid vast te stellen bestaat uit het gebruik van objectieve tests. Hoewel dit vermoedelijk moeilijk voorstelbaar is, blijken speciaal geconstrueerde tests bestaande uit meerkeuzevragen in staat het holistische oordeel uitstekend te voorspellen. Deze tests kunnen dus betrouwbaar en valide zijn. Doordat het moeilijk voorstelbaar is, dat men schrijven goed kan meten via meerkeuzevragen worden deze tests echter

weinig toegepast. Verder hebben deze tests als bezwaar dat ze niet toegepast kunnen worden op teksten van studenten.

Het eerste doel van het onderzoek was om na te gaan of er een methode te vinden was om schrijfvaardigheid te kwantificeren anders dan de twee hiervoor genoemde. Uitgangspunt was hierbij de constatering van een van de onderzoekers dat de eerstejaars hbo-studenten wel erg veel fouten maakten. Konden deze fouten objectief aangetoond en zo ja, hoeveel fouten maakten deze studenten dan?

Foutenonderzoek

In het foutenonderzoek kregen vier beoordelaars die goed konden schrijven 30 teksten te beoordelen op fouten die ze moesten onderstrepen en omschrijven. De teksten vormden een steekproef van teksten afkomstig van eerstejaars hbo-studenten (20) en van universitaire eerstejaars (10). De beoordelaars werkten onafhankelijk van elkaar en waren vrij in wat ze als 'fout' wilden signaleren.

Studenten die goed schrijven, produceren langere teksten dan studenten die slecht schrijven en maken daardoor in totaal meer fouten. Daarom is het noodzakelijk te werken met het aantal fouten PHW (per honderd woorden) of met het aantal fouten per A4 (500 woorden). Na deze correctie blijken studenten die langere teksten produceren belangrijk minder fouten per honderd woorden te maken dan studenten die korte teksten produceren.

De beoordelaars bleken zeer overeen te stemmen (gemiddelde onderlinge correlatie: 0.85) over het aantal fouten per honderd woorden in de teksten. Ook in het pilotonderzoek en het TAVAN-effectonderzoek werden zeer hoge waarden gevonden. Om teksten redelijk betrouwbaar te beoordelen op het aantal fouten per honderd woorden zal in veel gevallen een enkele beoordelaar al volstaan.

Wanneer men schrijfvaardigheid definieert als het aantal fouten per honderd woorden, is het probleem van de onbetrouwbare beoordeling daarmee opgelost. Om deze vorm van schrijfvaardigheid te onderscheiden van de holistisch beoordeelde schrijfvaardigheid, kozen we de term 'basale schrijfvaardigheid'. Het gaat er niet om dat de student een lang artikel kan schrijven, het gaat erom dat hij met niet te veel fouten een A4'tje (500 woorden of korter) kan schrijven.

Ondanks deze hoge overeenstemming verschilden de beoordelaars echter nog wel qua gemiddelde en spreiding. De ene beoordelaar was kritischer dan de andere, dat wil zeggen, signaleerde meer fouten. De aantallen fouten die beoordelaars signaleren in teksten zijn daardoor nog niet eenvoudig te interpreteren. Om dit probleem op te lossen is de methode van de bevestigde fouten ontwikkeld. Bevestigde fouten zijn fouten die door ten minste twee onafhankelijke beoordelaars zijn gesignaleerd. Aan het bestaan van een bevestigde fout kan daardoor moeilijk getwijfeld worden. Een beoordelaar kan nog zo veel fouten signaleren, als zijn fouten niet bevestigd worden door een andere beoordelaar, resulteren ze niet in 'bevestigde' fouten.

Beoordelaars bleken het over de aantallen bevestigde fouten zeer eens te zijn. De gemiddelde onderlinge correlatie tussen beoordelaars bedroeg 0.93. Hoewel aantallen bevestigde fouten eenvoudiger interpreteerbaar zijn dan gesignaleerde fouten, bleken ze verder niet tot wezenlijk andere uitkomsten te leiden. De correlatie tussen bevestigde fouten per honderd woorden en gesignaleerde fouten per honderd woorden bedroeg 0.93 en was daarmee zeer hoog.

Hoeveel fouten komen, objectief gedefinieerd, in teksten van eerstejaarsstudenten voor? Universitaire eerstejaarsstudenten bleken gemiddeld 42 bevestigde fouten te maken in een A4-tekst (500 woorden); eerstejaars in het hbo maakten gemiddeld 81 bevestigde fouten. Een vierde van de hbo-studenten maakte zelfs meer dan 100 bevestigde fouten per A4. Bij ongeveer 10% van de hbo-studenten werden waarden van rond de 150 bevestigde fouten of meer per A4 geconstateerd. Ook bij de herschrijfopdrachten die eerstejaars hbo-studenten later maakten in het kader van het onderzoek naar de effectiviteit van het nieuwe programma werden vergelijkbare aantallen fouten per honderd woorden gevonden.

Deze aantallen fouten zijn dermate groot dat ze moeilijk vallen voor te stellen. Kennelijk is het Nederlandse onderwijssysteem niet effectief om studenten in het hoger onderwijs redelijk foutloos te leren schrijven.

Het foutenonderzoek leverde ook een overzicht van de soorten fouten die eerstejaarsstudenten maken. De meest voorkomende fouten waren: 'Verkeerd woord', 'Niet-lopende zin', 'Interpunctie', Overbodig woord/overbodige zin', 'Alinea-indeling', 'Voorzetsel', 'Spelfout' en 'Ontbrekend woord'. Samen waren deze acht categorieën goed voor 75% van alle bevestigde fouten die eerstejaarsstudenten maken. D/t-fouten bleken wel

door iedere beoordelaar gesignaleerd te worden wanneer ze voorkwamen**,** maar relatief weinig (minder dan 2%) voor te komen.

Beoordeling bestaande methodes

Wat is de waarde van bestaande methodes om iets aan het probleem van de tekortschietende basale schrijfvaardigheid te doen? Het tweede deelonderzoek probeerde deze vraag te beantwoorden door onderwijsmethodes te beoordelen die verkrijgbaar zijn om studenten op dit punt bij te spijkeren. In totaal werden zeventien papieren methodes en negen digitale methodes beoordeeld.

Voor de didactische beoordeling van de bestaande onderwijsprogramma's is uitgegaan van het ABC-leermodel. ABC staat voor: Antecedents, Behavior, Consequences. Vertaald naar onderwijstermen: opdracht, antwoord, feedback. Dit model gaat ervan uit dat mensen leren door te doen. Kennis en vaardigheden moeten worden ingeoefend en feedback is daarbij van doorslaggevend belang. Veel kleine en duidelijke opdrachten werken beter dan enkele grote, vage opdrachten. Feedback moet snel, duidelijk en liefst positief zijn. Opdrachten moeten geleidelijk moeilijker worden (Cooper, Heron & Heward, 2007; Heward, 2005; Jenson, Sloane & Young, 1988; Malott, 2008; Vargas, 2009).

Bij iedere bestaande schrijfvaardigheidsmethode ontbrak een duidelijke doelstelling. Geen enkele methode bleek empirisch onderzocht te zijn op effectiviteit. De digitale methodes werden positiever beoordeeld op het punt van feedback. De hoeveelheid oefeningen en de geordendheid daarvan scoorde bij beide soorten methodes even hoog en liet te wensen over. Het beste digitale programma, Nedercom, scoorde qua feedback goed, maar werd wat betreft de hoeveelheid oefeningen beoordeeld als matig en met het oog op de ordening van de oefenstof als slecht.

Een probleem bij alle methodes was dat ze alle mogelijke taalproblemen behandelden, maar doorgaans niet de fouten die studenten werkelijk maken. 'Verkeerd woord' was de meest voorkomende foutsoort bleek in het foutenonderzoek, maar deze fout werd amper behandeld. Ook andere veel voorkomende fouten zoals 'Niet-lopende zin', 'Overbodig woord/overbodige zin', 'Alinea-indeling', 'Voorzetsel' en 'Ontbrekend woord' werden niet of nauwelijks geoefend in de onderzochte taalmethodes, hoewel deze fouten samen goed waren voor drie vierde van alle fouten.

Het oordeel over de geschiktheid van de bestaande methodes was daarmee negatief. De beste papieren en digitale methodes leken nog steeds belangrijke bezwaren te hebben. Op basis van dit negatieve oordeel werd besloten een nieuw programma te ontwikkelen: TAVAN (TAalVAardigheid Nieuw) en dit op effectiviteit te onderzoeken.

Effect van het nieuwe TAVAN-programma

Het nieuwe TAVAN-programma is op effectiviteit onderzocht bij een groep eerstejaarsstudenten in het hbo. De TAVAN-groep maakte in de eindtekst 3.8 fouten minder per honderd woorden dan in de begintekst. Dat betekent een reductie van 19 fouten per A4 (500 woorden) of meer dan 20%. De controlegroep die het traditionele taalvaardigheidsprogramma volgde, verbeterde qua aantal fouten niet. Het verschil met de controlegroep wat betreft de vermindering van het aantal fouten bedroeg meer dan 1 standaarddeviatie. Dit geldt als een groot effect.

De resultaten van het TAVAN-programma laten zien dat de hoge aantallen fouten die gevonden worden in schrijfproducten van studenten met een relatief korte training (twintig lesuur) aanzienlijk gereduceerd kunnen worden. Basale schrijfvaardigheid blijkt belangrijk en snel verbeterd te kunnen worden door te oefenen met het herschrijven van foute zinnen.

Het idee dat studenten veel fouten produceren doordat het hun ontbreekt aan een juiste schrijfattitude, bleek niet te kloppen. De schrijfattitude bleek niets te zeggen over hoe goed men schreef. TAVAN-studenten die slecht scoorden ten opzichte van het gemiddelde in het online-programma bleken een betere schrijfattitude te ontwikkelen, terwijl studenten die goed scoorden ten opzichte van het gemiddelde een slechtere schrijfattitude ontwikkelden. Deze verandering in schrijfattitude bleek echter niet samen te gaan met een vermindering van het aantal geproduceerde fouten (de gestandaardiseerde leerwinst). Schrijfattitude heeft kennelijk weinig te maken met schrijfvaardigheid.

De verwachting dat deelname aan het TAVAN-programma door de feedback mogelijk zou leiden tot een gemiddeld lagere inschatting van de eigen schrijfvaardigheid, werd niet bevestigd. Het nieuwe programma bleek gemiddeld genomen geen invloed te hebben op de eigen inschatting van de schrijfvaardigheid. Wel bleken studenten die bij het online-programma lager dan het gemiddelde scoorden, hun eigen schrijfvaardigheid lager te gaan

inschatten. Dit werd echter gecompenseerd door studenten die beter dan gemiddeld scoorden en die precies andersom reageerden. Het resultaat van het nieuwe programma op de eigen inschatting van de schrijfvaardigheid was daarmee dat men zich meer overeenkomstig de eigen prestaties ten opzicht van het groepsgemiddelde ging inschatten. In die zin ging men zichzelf realistischer inschatten.

De controlegroep ging zichzelf echter positiever inschatten, zonder dat men daadwerkelijk verbeterd was. De TAVAN-groep werd wel belangrijk beter, maar ging zichzelf gemiddeld niet positiever inschatten. De eigen inschatting van de schrijfvaardigheid is daarmee geen valide maat om basale schrijfvaardigheid vast te stellen.

Basale schrijfvaardigheid werd vastgesteld door studenten teksten met fouten te laten herschrijven en door studenten zinnen te laten herschrijven in het online-programma. Deze laatste manier bleek zeer betrouwbaar en zeer valide (qua correlatie met het aantal fouten per honderd woorden in begin- en eindtekst samen). Het online-programma blijkt daarmee een eenvoudige, betrouwbare en valide manier om basale schrijfvaardigheid vast te stellen, waarbij de student daadwerkelijk schrijft en tegelijkertijd ook nog beter leert schrijven.

Het niveau van basale schrijfvaardigheid blijkt verder reikende consequenties te hebben dan het aantal fouten per honderd woorden in een tekst. Studenten met een goede basale schrijfvaardigheid blijken langere teksten te schrijven, minder tijd nodig te hebben om te schrijven, een hogere vooropleiding te hebben, zichzelf positiever in te schatten qua schrijfvaardigheid en minder vaak te stoppen met de studie.

Constructie TAVAN-programma

Gedurende de tien TAVAN-lessen van twee uur is het eerste lesuur steeds geoefend met het herschrijven van foute zinnen via een speciaal ontwikkeld online-computerprogramma. Het tweede lesuur is gebruikt om in Word een korte tekst met fouten te herschrijven. Het TAVAN-programma bevatte opdrachten en feedback, maar geen 'theorie' die de student moet weten. Het programma was leerstof-vrij.

Waarom slaagde het TAVAN-programma er in het aantal fouten terug te dringen, terwijl bestaande onderwijsmethodes daar kennelijk vaak niet in slagen? Deze vraag valt niet met

zekerheid te beantwoorden, doordat TAVAN op veel punten afwijkt van gangbaar onderwijs. We noemen hierna de belangrijkste verschilpunten. Duidelijk lijkt dat het het online-programma in combinatie met de opzet volgens het ABC-model een doorslaggevende rol speelde.

1. TAVAN gaat uit van een expliciet doel: studenten moeten minder fouten per honderd woorden maken.
2. Om dit doel te bereiken wordt geoefend met het herschrijven van foute zinnen; niet met plannen of zelfstandig schrijven.
3. Het programma gaat niet uit van leerstof of theorie die door de student bestudeerd moet worden, maar is volledig gericht op oefenen door de student.
4. Er wordt voor de oefeningen niet uitgegaan van veronderstelde fouten, maar van de lijst met foutsoorten uit het foutenonderzoek en hun frequenties.
5. Er wordt niet uitgegaan van enkele grote opdrachten, maar van veel kleine. Per minuut maakt een student doorgaans drie à vier opdrachten.
6. Het online-programma zorgt voor onmiddellijke en duidelijke feedback.
7. Het online-programma werkt structurerend door automatisch de oefeningen te presenteren en de resultaten bij te houden.
8. De docent doceert niet, maar fungeert als coördinator en manager.

Fouteneffect-onderzoek

Hoe erg is een taalfout? Sommigen zullen iedere fout er één te veel vinden. Anderen zullen stellen dat taalfouten een normaal verschijnsel zijn, waar we ons niet al te druk over moeten maken. Het doel van het vierde deelonderzoek was na te gaan of fouten in een tekst effect hebben op de waardering van die tekst door de lezer.

In totaal werden 48 door studenten geschreven teksten door beide onderzoekers beoordeeld, eerst holistisch en daarna op grond van het aantal fouten per honderd woorden. De twee groepen hbo-studenten die de teksten hadden geschreven, werden zelf ook gevraagd teksten van medestudenten te beoordelen.

Tussen het aantal fouten per honderd woorden in een tekst dat was vastgesteld via de beide onderzoekers en het holistische oordeel over een tekst, bleek een zeer sterk verband te

bestaan bij zowel expert-beoordelaars als bij student-beoordelaars. In beide gevallen was de correlatie na correctie voor onbetrouwbaarheid -0.89. Het aantal fouten per honderd woorden en het holistische oordeel overlapten elkaar daarmee voor ongeveer 79% qua gemeenschappelijke variantie. Deze zeer hoge waarde laat zien dat beoordelaars zich bij hun holistische oordeel (bewust of onbewust) sterk laten beïnvloeden door het aantal fouten in een tekst.

De studenten bleken als holistische beoordelaars onderling belangrijk minder overeen te stemmen dan de beide onderzoekers (gemiddelde onderlinge correlatie 0.22 versus 0.65). Om dezelfde betrouwbaarheid te bereiken als beide onderzoekers samen waren twaalf studentbeoordelaars nodig.

Het holistische oordeel van de studenten correleerde echter na correctie voor onbetrouwbaarheid vrijwel perfect (0.99 voor 26 teksten) met het holistische oordeel van beide onderzoekers. Studenten hanteerden voor de beoordeling kennelijk dezelfde criteria en normen als beide onderzoekers. Hoewel het studentenoordeel minder betrouwbaar was gemiddeld, bleek het even (criterium)valide.

Er bleek een significant verband te bestaan tussen hoe goed studenten schreven volgens het holistische oordeel van beide onderzoekers en hoe goed ze holistisch beoordeelden ($r=0.31$, $p=0.041$, 2-zijdig, $N=44$). Studenten die goed schreven, waren beter in het holistische beoordelen van teksten dan studenten die slecht of matig schreven.

Om volledige zekerheid te krijgen dat het verband tussen het aantal fouten per honderd woorden en de beoordeling van de tekst inderdaad causaal was, is vervolgens een experiment uitgevoerd waarbij drie teksten van studenten in drie versies aan lezers zijn voorgelegd: de oorspronkelijke versie met veel fouten en twee gecorrigeerde versies. Iedere lezer kreeg hierbij slechts één tekst te lezen. De waardering op een schaal van 0 tot 100 voor de teksten met fouten was gemiddeld 30, voor de verbeterde teksten was de waardering gemiddeld 48, meer dan anderhalf maal zoveel. Dit verschil komt overeen met 1.4 standaarddeviatie en is daarmee zeer groot.

Zowel het correlationele onderzoek als het experimentele onderzoek naar het verband tussen aantal fouten per honderd woorden in de tekst en het (holistische) oordeel over die tekst laten daarmee zien dat fouten zeer negatief inwerken op het oordeel over de tekst. Het idee dat fouten er voor de lezer niet toe doen, blijkt onjuist.

Het meten van basale schrijfvaardigheid

Effectief schrijfonderwijs begint bij een goede meetmethode. Zonder goede meetmethode kan niet gecheckt worden of een programma effectief is. Verder moet een goede meetmethode ook omgezet kunnen worden naar een effectief trainingsprogramma. Anders kan men wel meten, maar niet trainen en heeft meten weinig nut. Vanuit dit uitgangspunt is in het vijfde deelonderzoek gekeken naar de mogelijkheden om basale schrijfvaardigheid te meten.

Tot nu toe waren er slechts twee methodes bekend voor het meten van basale schrijfvaardigheid: holistische beoordeling en objectieve tests. Holistische beoordeling is onbetrouwbaar en arbeidsintensief en door de trage en weinig specifieke feedback niet echt geschikt als trainingsmethode. Objectieve tests hebben als nadeel dat de student niet daadwerkelijk schrijft, maar alleen het beste alternatief kiest. Objectieve tests lijken daardoor niet geschikt als onderwijsmethode, maar alleen bruikbaar als selectiemethode.

Allereerst is getracht uit Godshalk, Swineford & Coffman (1966) af te leiden hoe goed objectieve test het holistische oordeel kunnen voorspellen. Afgeleid kon worden dat de door Godshalk et al. gebruikte objectieve tests na correctie voor onbetrouwbaarheid 0.87 correleerden met het holistische oordeel. Ook Breland en Gaynor (1979) vonden een soortgelijk waarde. Deze waarde komt overeen met 76% verklaarde variantie.

Hoe is het mogelijk dat objectieve tests het holistische oordeel zo goed kunnen voorspellen? De door ons voorgestelde verklaring is dat objectieve tests de FOC-factor (Fouten Opsporen en Corrigeren) rechtstreeks meten bij de student. De student drukt zijn FOC-niveau uit in de tekst via het aantal fouten per honderd woorden. De holistische beoordelaar wordt vervolgens beïnvloed door het aantal fouten per honderd woorden. Deze verklaring veronderstelt hoge positieve correlaties tussen alle drie variabelen.

Aan de hand van de gegevens uit het fouteneffect-onderzoek en TAVAN2 (de tweede keer dat het TAVAN-programma gegeven werd) kon dit verklaringsmodel bevestigd worden. Hierbij werd de TAVAN-score als objectieve test opgevat. Na correctie voor onbetrouwbaarheid bedroeg de gemiddelde onderlinge correlatie 0.86. Dit is een zeer hoge waarde, die aangeeft dat de drie variabelen in zeer hoge mate dezelfde factor meten.

Een derde nieuwe manier waarop basale schrijfvaardigheid op een criteriumvalide manier bleek te kunnen worden vastgesteld, was het meten van de kwaliteit van het holistische oordeel uitgebracht door de student over zes teksten geschreven door medestudenten. Naarmate een student het ideale holistische oordeel beter benaderde, bleek hij ook hoger te scoren op de overige drie schrijfvaardigheidsmaten. De gemiddelde onderlinge correlatie met de overige drie schrijfvaardigheidsmaten was na correctie voor onbetrouwbaarheid 0.88 en daarmee zeer hoog.

Doordat van de vier onderzochte maten (TAVAN-score*, aantal fouten PHW*, holistische oordeel, kwaliteit uitgebrachte holistische oordeel) twee maten (aangegeven met *) qua inhoud het vermogen meten om fouten op te sporen en te corrigeren (de FOC-factor), kan geconcludeerd worden, dat ook het holistische oordeel en de kwaliteit van het uitgebrachte holistische oordeel overwegend deze FOC-factor meten.

Van de drie nieuwe maten is de TAVAN-score bruikbaar gebleken als trainingsmethode (hoofdstuk 6). Met enige aanpassingen lijkt ook de kwaliteit van het holistische oordeel bruikbaar als trainingsmethode. Men kan dan denken aan twee vergelijkbare zinnen waaruit de student de beste moet kiezen. De praktische relevantie is dat er een tweede maat, naast TAVAN, beschikbaar komt voor trainingsdoeleinden. De wetenschappelijke relevantie is dat deze maat, mits aangetoond effectief als trainingsmethode, niet gebaseerd is op daadwerkelijk schrijven, maar op nauwkeurig lezen. Dit zou aantonen dat lezen en schrijven inderdaad één factor vormen, zoals de doorgaans hoge correlaties tussen lees- en schrijfvaardigheid ook lijken te suggereren.

Betrouwbaarheidsproblemen

Indien met meerdere holistische beoordelaars en meerdere topics wordt gewerkt, blijken er een groot aantal manieren te zijn waarop een betrouwbaarheidscoëfficiënt kan worden berekend, terwijl onderzoekers zich dit doorgaans niet realiseren. Verder blijken er verschillende soorten betrouwbaarheid onderscheiden te kunnen worden. Voorgesteld wordt de scorebetrouwbaarheid te berekenen op grond van de gemiddelde onderlinge correlatie tussen de topictotaalscores en het aantal topics. De topicbetrouwbaarheid bij perfect betrouwbare beoordeling kan geschat worden als de verhouding van de scorebetrouwbaarheid en de beoordelaarsbetrouwbaarheid.

Voor het betrouwbaar vaststellen van het niveau van basale schrijfvaardigheid via het aantal fouten PHW leek een tekst ter lengte van een halve A4 (250 woorden) voldoende te zijn. Dit lijkt belangrijk minder te zijn dan voor een even betrouwbare holistische beoordeling nodig is.

De verklaring dat bij holistische beoordeling belangrijk meer topics nodig zijn om een vergelijkbare betrouwbaarheid te bereiken, is vermoedelijk dat het oordeel van holistische beoordelaars over het taalgebruik interfereert met het oordeel over de inhoud van de tekst, op een wijze die te vergelijken valt met het Stroop-effect (zie 9.3). Tot nu toe werd altijd aangenomen dat de topiconbetrouwbaarheid het gevolg zou zijn van sterke wisselingen in de schrijfvaardigheid van de student. Daar zijn buiten de relatief lage holistische topicbetrouwbaarheid echter geen aanwijzingen voor, terwijl er wel duidelijke aanwijzingen zijn dat schrijfvaardigheid relatief stabiel is.

10.3 Nabeschouwing

In het kader van dit onderzoek zijn zeven onderzoeksvragen gesteld die beantwoord zijn in zes deelonderzoeken. Een groot aantal vervolgvragen is gesteld, talloze correlaties en gemiddelden zijn berekend en gepresenteerd. Het is tijd om de balans op te maken en na te gaan wat we uit dit complexe geheel kunnen concluderen.

Het doel van het totale onderzoek was het kunnen meten en maximaliseren van basale schrijfvaardigheid bij eerstejaars hbo-studenten. Naar verwachting zullen de conclusies ook van toepassing zijn op andere studenten en op leerlingen. Bij iedere conclusie wordt in de toelichting op de conclusie het hoofdstuk vermeld waarop de conclusie gebaseerd wordt. Eerst kijken we naar de opbrengst voor het meten van basale schrijfvaardigheid.

1. Een tekst van 250 woorden blijkt voldoende voor het betrouwbaar bepalen van basale schrijfvaardigheid door het aantal fouten per honderd woorden te bepalen.
Hoewel dit een van de laatste resultaten was uit het onderzoek (hoofdstuk 9), lijkt de importantie van deze conclusie groot. Het idee dat voor het vaststellen van basale schrijfvaardigheid meerdere teksten nodig zijn, blijkt niet juist. Het aantal fouten per

honderd woorden blijkt een zeer betrouwbare factor te zijn op basis waarvan basale schrijfvaardigheid eenvoudig kan worden vastgesteld. Ook het idee dat basale schrijfvaardigheid moeilijk vaststelbaar zou zijn, blijkt daarmee onjuist. Voor de betrouwbare bepaling van het aantal fouten per honderd woorden bleek een enkele beoordelaar doorgaans voldoende (foutenonderzoek, hoofdstuk 4).

2. In totaal blijken vijf verschillende maten beschikbaar te zijn voor het vaststellen van basale schrijfvaardigheid die criteriumvalide zijn (hoog correleren met het holistische oordeel) en constructvalide zijn (hoog correleren met alle andere basale schrijfvaardigheidsmaten). Alle vijf maten leveren soortgelijke uitkomsten.

De vijf maten zijn:

1. het holistische oordeel over een door de student geschreven tekst;
2. de score op een objectieve toets bedoeld de FOC-factor te meten;
3. het aantal fouten per honderd woorden in een door de student geschreven tekst;
4. de TAVAN-score behaald in het TAVAN-programma;
5. de kwaliteit van door de student uitgebrachte holistische oordelen.

De eerste twee methodes waren al langer bekend, de laatste drie methodes zijn in dit onderzoek gevonden en aangetoond (hoofdstuk 8). Het idee dat basale schrijfvaardigheid moeilijk meetbaar is, wordt door dit resultaat weerlegd. Er blijkt een behoorlijk aantal verschillende methodes te bestaan die allemaal aangetoond criterium- en constructvalide zijn om basale schrijfvaardigheid vast te stellen.

3. Alle vijf constructvalide maten voor het meten van basale schrijfvaardigheid meten vooral het vermogen om fouten op te sporen en te corrigeren (de FOC-factor).

Deze conclusie (hoofdstuk 8) geeft aan dat fouten (vermoedelijk onbewust) bij de beoordeling van teksten een grote rol spelen. Dit lijkt haaks te staan op de overtuiging van velen dat een tekst primair bedoeld is om betekenis te communiceren en dat fouten niet zo belangrijk zouden zijn.

4. Dat alle vijf constructvalide maten voor het meten van basale schrijfvaardigheid overwegend de FOC-factor meten, betekent niet dat er geen tweede, non-FOC factor zou kunnen bestaan.

In het in hoofdstuk 8 beschreven onderzoek werden geen aanwijzingen gevonden voor het bestaan van een tweede, non-FOC factor. Het is echter mogelijk dat er nooit een tweede

factor is gevonden dan wel aangetoond, doordat er nooit gericht en systematisch naar gezocht is. Voor een tweede, non-FOC factor lijkt vooral de inhoud die teksten communiceren een kandidaat.

5. Fouten blijken een aantoonbaar groot negatief effect te hebben op de beoordeling van een tekst door lezers.
Dit resultaat werd via een experiment in het fouteneffect-onderzoek (hoofdstuk 7) aangetoond en verklaart de derde conclusie, dat alle maten de FOC-factor meten. Dat de FOC-factor zo belangrijk is, komt doordat lezers en beoordelaars (mogelijk onbewust) sterk negatief door taalfouten worden beïnvloed.

Wanneer we de uitkomsten op meetgebied samenvatten, zien we dat basale schrijfvaardigheid en het aantal fouten per honderd woorden vrijwel synoniem zijn (conclusie 3). Verder zien we dat dat aantal fouten per honderd woorden op totaal verschillende manieren kan worden vastgesteld (conclusie 2) en dat een korte tekst van 250 woorden al voldoende is (conclusie 1) om basale schrijfvaardigheid vast te stellen. Waarom basale schrijfvaardigheidsmaten zo sterk gebaseerd zijn op fouten, komt doordat lezers en beoordelaars zich daar erg door laten beïnvloeden (conclusie 5). Hieruit kunnen we echter nog niet afleiden dat lezers zich niet ook door andere factoren, zoals de inhoud van een tekst kunnen laten beïnvloeden (conclusie 4). Het lijkt echter weinig twijfel te leiden dat fouten in de tekst het effect van een positief beoordeelde inhoud meer dan volledig teniet kunnen doen. Lezers en beoordelaars laten zich kennelijk sterker beïnvloeden door fouten in de tekst dan door de inhoud van de tekst.

Wat zijn de belangrijkste conclusies met betrekking tot het maximaliseren van basale schrijfvaardigheid die we op grond van het onderzoek kunnen trekken?

6. Van de vijf constructvalide maten voor het meten van basale schrijfvaardigheid is van één maat, de TAVAN-score, de 'trainingsvaliditeit' (de geschiktheid als trainingsmiddel) aangetoond.
In hoofdstuk 6 bleek dat het TAVAN-programma bij hbo-studenten leidt tot minder fouten bij het herschrijven van teksten. De TAVAN-items blijken daarmee geschikt als trainingsmiddel en kunnen daarnaast ook gebruikt worden om basale schrijfvaardigheid te meten.

7. Van de vijf constructvalide maten voor het meten van basale schrijfvaardigheid is een tweede maat, de kwaliteit van het door de student uitgebrachte holistische oordeel, mogelijk trainingsvalide.

In hoofdstuk 8 wordt aangetoond dat de maat 'kwaliteit uitgebrachte holistische oordeel' kan worden vertaald in items waarbij de student moet kiezen uit twee vergelijkbare zinnen, dat wil zeggen: hij moet beoordelen welke de beste is. De praktische relevantie hiervan is dat er naast de huidige TAVAN-items die gebaseerd zijn op foute zinnen, een tweede type item beschikbaar komt waarbij de student niet hoeft te schrijven, maar alleen hoeft te kiezen. Deze items zijn daardoor niet beperkt tot zinnen met fouten.

De wetenschappelijke relevantie is dat het mogelijk lijkt via dit soort items studenten gevoel bij te brengen voor goed taalgebruik (ze leren naar verwachting snel te discrimineren tussen goede en slechte zinnen). Men zou verwachten dat deze kennis of dit taalgevoel gebruikt wordt zodra de student zelf een tekst moet schrijven. Indien dit inderdaad zo zou blijken te zijn, is daarmee gedemonstreerd dat basale schrijfvaardigheid niet alleen 'indirect' (zonder te schrijven) meetbaar is, maar ook 'indirect' getraind kan worden.

8. TAVAN blijkt uitermate effectief in het verbeteren van de basale schrijfvaardigheid bij eerstejaars hbo-studenten.

Dit wordt aangetoond in hoofdstuk 6. De twintig lesuren TAVAN (waarvan iedere 2 uur slechts 1 uur met het online-programma werd geoefend) resulteerden in ruim 20% minder fouten per honderd woorden.

9. Gangbaar schrijfonderwijs lijkt in doorsnee niet-effectief te zijn voor het opbouwen van basale schrijfvaardigheid.

In het onderzoek naar de effectiviteit van TAVAN (hoofdstuk 6) bleek de controlegroep die het gangbare programma volgde, niet verbeterd te zijn. In het foutenonderzoek (hoofdstuk 4) bleek dat het gemiddelde aantal bevestigde fouten dat eerstejaars hbo-studenten in hun teksten maken 81 per A4 (500 woorden) bedroeg. In combinatie met het resultaat dat TAVAN erin slaagt in 20 lesuur een reductie van 20% te realiseren, lijkt dit te betekenen dat het Nederlandse onderwijssysteem op het gebied van schrijfonderwijs niet effectief is. Voor de beoordeelde zeventien papieren en negen digitale methodes (hoofdstuk 5) bestonden geen duidelijke doelstellingen, de methodes waren niet aangetoond als effectief, de

behandelde stof had overwegend niet betrekking op de fouten die het meeste werden gemaakt en ten slotte lieten alle methodes ook qua didactiek nog te wensen over. De ineffectiviteit van gangbaar schrijfonderwijs verklaart ook de persistentie en de verbreidheid van klachten over tekortschietende schrijfvaardigheid (hoofdstuk 1).

10. De TAVAN-methode ontleent haar effectiviteit vermoedelijk aan de opzet uitgaande van ABC-sequenties (A. opdracht; B. antwoord; C feedback) in combinatie met het online-programma voor de realisatie en lijkt na aanpassing ook op andere inhoudsgebieden toepasbaar.

In hoofdstuk 6 wordt de constructie van TAVAN beschreven. Het ABC-model (zie 3.3) was bij de constructie van TAVAN het uitgangspunt (Cooper et al., 2007, p. 42). Dit model stelt dat onderwijs gezien moet worden als een opeenvolging van bij voorkeur veel ABC-sequenties. Het online-programma maakte de realisatie van deze ABC-sequenties mogelijk door het presenteren van de opgaven en het geven van onmiddellijke en duidelijke feedback en registreerde de resultaten.

De ervaringen met effectieve onderwijsmethodes (zie 6.1) laten zien dat het voor een onderwijsmethode niet voldoende is om effectief te zijn; de methode moet voor docenten en studenten een meerwaarde inhouden. Op dit moment (oktober 2013) loopt het TAVAN-programma voor het derde jaar (TAVAN3) waarbij in totaal ruim 500 studenten (17 klassen) het programma 'volgen' onder leiding van een vijftal docenten. Dit wijst erop dat TAVAN voor studenten en docenten inderdaad een bepaalde meerwaarde levert.

Ten opzichte van TAVAN2 lijkt TAVAN3 een belangrijke verbetering door onder meer de belangrijk kortere responsetijden van het online-programma en het grotere aantal kortere lessen. Wie TAVAN2 vergelijkt met de eerste uitvoering van TAVAN waarvan de resultaten in dit proefschrift beschreven zijn, is vermoedelijk geneigd hetzelfde te denken. Werd bij TAVAN slechts de helft van de tijd uitgetrokken voor het online-programma, in TAVAN2 werd reeds alle beschikbare tijd gereserveerd voor de online-oefeningen. Dit laat zien dat TAVAN tot nu toe een snelle ontwikkeling doormaakt en dat de eerste versie van TAVAN kennelijk het begin van deze ontwikkeling vormt. Hieruit kunnen echter geen betrouwbare voorspellingen voor de verdere toekomst van TAVAN of TAVAN-achtige programma's worden afgeleid.

Beperkingen

De tien conclusies waarin we de belangrijkste uitkomsten van de zes deelstudies geprobeerd hebben samen te vatten, zijn gebaseerd op onderzoek en hebben daarmee beperkingen. Hierna proberen we per conclusie na te gaan hoe solide het onderliggende onderzoek is.

1. Een tekst van 250 woorden is voldoende voor het betrouwbaar bepalen van de basale schrijfvaardigheid door het aantal fouten per honderd woorden te bepalen.
Deze conclusie is gebaseerd op het opsplitsen van een door de studenten geschreven tekst in twee helften. Een overtuigender toetsing zou zijn de studenten bij twee verschillende gelegenheden een korte tekst te laten schrijven en op grond hiervan de betrouwbaarheid te bepalen. Deze conclusie is nog niet volledig overtuigend aangetoond.

2. Er blijken vijf verschillende maten te zijn voor het vaststellen van basale schrijfvaardigheid.
Deze conclusie is gebaseerd op de gegevens van TAVAN2 waarmee werd aangetoond dat de vier onderzochte variabelen onderling hoog correleren. Uit de literatuur blijkt verder een uitermate hoog verband tussen het holistische oordeel en objectieve testscores, zodat kennelijk alle vijf soorten variabelen dezelfde factor meten. Eleganter zou echter zijn om over een dataset te beschikken met metingen voor alle vijf soorten variabelen. Verder was de meting van de kwaliteit van het holistische oordeel weinig betrouwbaar en zou deze betrouwbaarheid verbeterd kunnen worden.

3. Alle vijf constructvalide maten voor het meten van basale schrijfvaardigheid meten vooral het vermogen om fouten op te sporen en te corrigeren (de FOC-factor).
De beperking bij conclusie 2 geldt ook hier. Conclusie 3 zal mogelijk op veel ongeloof stuiten en verdient om die reden replicatie. Doordat eerder in de literatuur een soortgelijk verband gerapporteerd werd, lijkt er op dit punt weinig twijfel mogelijk.

4. Dat alle vijf constructvalide maten voor het meten van basale schrijfvaardigheid overwegend de FOC-factor meten, betekent niet dat er geen tweede, non-FOC factor zou kunnen bestaan.
Deze stelling is vooral een inperking van conclusie 3: er bestaat misschien toch een non-

FOC factor. Op dit moment is er echter -- voor zover bekend -- geen onderzoek waarmee die non-FOC factor (het oordeel over de inhoud?) aangetoond wordt. Voorlopig lijkt dus te gelden dat het taalgebruik wel beoordeeld kan worden, maar inhoud niet, althans niet betrouwbaar, totdat het tegendeel wordt aangetoond.

5. Fouten blijken een aantoonbaar groot negatief effect te hebben op de beoordeling van een tekst door lezers.
Deze stelling werd aangetoond via een experimenteel onderzoek en is daarmee vrij hard. Wat echter niet duidelijk is, is de verklaring. Waarom hebben fouten zo'n sterk effect en vermoedelijk zelfs een groter effect dan de inhoud?

6. Van de vijf constructvalide maten voor het meten van basale schrijfvaardigheid is van één maat, de TAVAN-score, de 'trainingsvaliditeit' (de geschiktheid als trainingsmiddel) aangetoond.
Ook deze stelling werd aangetoond via een experiment en is daarmee relatief hard.

7. Van de vijf constructvalide maten voor het meten van basale schrijfvaardigheid is een tweede maat, de kwaliteit van het door de student uitgebrachte holistische oordeel, mogelijk trainingsvalide.
Deze conclusie is geformuleerd als mogelijkheid. Onderzoek zal moeten uitwijzen of het inderdaad mogelijk is de schrijfvaardigheid via trainen op deze maat te vergroten.

8. TAVAN blijkt uitermate effectief in het verbeteren van de basale schrijfvaardigheid bij eerstejaars hbo-studenten.
Deze stelling werd aangetoond via het TAVAN-experiment en is daarmee vrij hard. Wat nog niet duidelijk is, is wat de meest optimale opzet is voor TAVAN en welke praktijkproblemen grootschalige implementatie met zich mee brengt.

9. Gangbaar schrijfonderwijs lijkt in doorsnee niet-effectief te zijn voor het opbouwen van basale schrijfvaardigheid.
Deze conclusie is gebaseerd op de uitkomsten van het foutenonderzoek, op de beoordeling van de beschikbare methodes en op de constatering dat deugdelijk onderzoek waarin een duidelijke leerwinst wordt aangetoond voor het bestaande schrijfonderwijs vrijwel altijd ontbreekt. Verder blijkt uit de resultaten van TAVAN dat er een effectief alternatief is. Het

lijkt nu aan de voorstanders van het gangbare schrijfonderwijs te zijn om aan te tonen dat dit onderwijs wel effectief is.

10. *De TAVAN-methode lijkt na aanpassing ook op andere inhoudsgebieden toepasbaar.*
Het principe achter TAVAN is heel algemeen: gericht oefenen. Er lijkt geen reden te zijn om de TAVAN-methode niet ook op andere leerstofgebieden toe te passen, maar hoe dat in de praktijk uitwerkt, hangt ook af van allerhande andere factoren waardoor alleen daadwerkelijk uitproberen een definitief antwoord kan leveren over de gerealiseerde leerwinst.

Verder onderzoek

Enkele mogelijkheden voor vervolgonderzoek beschrijven we hierna.

-- De methode van het aantal bevestigde fouten per honderd woorden is nogal arbeidsintensief, terwijl het lastig is beoordelaars te vinden die zelf een goede basale schrijfvaardigheid bezitten. Het zou dan efficiënter zijn om voor een een aantal teksten van studenten het aantal bevestigde fouten per honderd woorden te bepalen en deze vervolgens via de TAVAN-score te koppelen aan het niveau van deze studenten. Op die manier kan men de basale schrijfvaardigheid bepalen via een TAVAN-test, terwijl de score vertaald kan worden naar het aantal bevestigde fouten PHW.

-- Het lijkt belangrijk dat er systematisch verzamelde kwantitatieve informatie beschikbaar komt over het niveau van basale schrijfvaardigheid bij Nederlandse studenten die periodiek opnieuw verzameld wordt.

-- In het fouteneffect-onderzoek werden beide gecorrigeerde versies van de teksten ongeveer gelijk beoordeeld, terwijl in de ene versie veel meer 'fouten' verbeterd waren dan in de andere versie. Dit duidt erop dat er fouten zijn die de tekst beschadigen en fouten die meer als een verbetering van de tekst moeten worden opgevat. Voor het beoordelen op fouten en het werken met fouten maakt dit onderscheid mogelijk veel uit. De vraag is of dit onderscheid op een of andere manier valt aan te tonen en hard valt te maken.

-- Het is mogelijk om via een klein computerprogramma de door studenten in teksten gebruikte woorden te scoren op frequentie en vervolgens per tekst een score te berekenen. De vraag is vervolgens of deze maat overeenkomt met de FOC-factor of iets anders meet.

-- Bestaat er een tweede (non-FOC) factor? Het lijkt goed mogelijk hier gericht naar te zoeken waarbij men van 'inhoud' zou verwachten dat die naast de FOC-factor doorwerkt in het holistische oordeel.

-- Het lijkt goed mogelijk studenten te laten trainen met 'kwaliteit uitgebrachte holistische oordeel'-items. Dit zijn items waarbij de student tussen twee zinnen of fragmenten de beste moet kiezen. De vraag is vooral of gericht oefenen over een langere periode leidt tot een waarneembaar betere basale schrijfvaardigheid.

-- Dit onderzoek begon met de waarneming dat eerstejaarsstudenten erg veel fouten in hun schriftelijk werk leken te maken. Het onderzoek was er vervolgens op gericht de aantallen fouten en het soort fouten in kaart te brengen en na te gaan welke mogelijkheden er waren om studenten een betere basale schrijfvaardigheid bij te brengen. In dat kader is uitgegaan van het ABC-leermodel dat gerealiseerd werd via het online-programma. Deze aanpak bleek effectief te zijn en het lijkt plausibel dat een soortgelijke aanpak eveneens in het basisonderwijs en voortgezet onderwijs zou kunnen werken. Ook lijkt het mogelijk deze methode toe te passen op andere leerstof. Onderzoek zal vervolgens moeten uitwijzen hoe effectief de geconstrueerde programma's zijn.

Bronnen

Abrahams, F. (2005, 3 november). Kafka. *NRC Handelsblad*.

Ahmed, W. (2010). *Expectancy-Value Antecedents and Cognitive Consequences of Students' Emotions in Mathematics*. Proefschrift. Rijksuniversiteit Groningen: GION.

Aiesec-congres (2009, maart). Internationaliseringscongres van Aiesec in samenwerking met VNO-NCW Noord, Groningen.

Al Fraidan, A. (2012). Evaluation of two ESP Textbooks. *English Language Teaching, 5* (6), 43-47. Geraadpleegd 25 januari 2014 via ccsenet.org/journal/index.php/elt/article/view/17463

Ansary, H. & Babaii, E. (2002). Universal Characteristics of EFL/ESL Textbooks: A Step Towards Systematic Textbook Evaluation. *The Internet TESL Journal, Vlll* (2). Geraadpleegd 15 januari 2014 via http://iteslj.org/

Anson, C. M. (2000). Response and the social construction of error. *Assessing Writing, 7,* 5-21.

Anson, C. M., Rashid Horn, S. & Schwegler, R. A. (2006, september). *In the Blink of an Eye: New research on error in Student Writing.* Paper gepresenteerd bij de Special Interest Group on Writing van EARLI (European Association for Research on Learning and Instruction), Antwerpen.

Atkinson, R. C. (2009, april). *The New SAT: A Test at War with Itself.* Paper gepresenteerd bij AERA (American Educational Research Association), San Diego. geraadpleegd 13 januari 2012 via rca.ucsd.edu/speeches/AERA_041509_Speech_Reflections_on_a_Century_of_College_Admissions_Tests.pdf

Atkinson, R. C. & Geiser, S. (2009). Reflections on a Century of College Admissions Tests. *Educational Researcher, 38* (9), 665-676.

Atkinson, D. & Murray, M. (1987, maart). *Improving Interrater Reliability.* Paper gepresenteerd bij de 38th Annual Meeting of the Conference on College Composition and Communication, Atlanta.

Bacon, D. R. & Scott Anderson, E. (2004). Assessing and Enhancing the Basic Writing Skills of Marketing Students. *Business Communication Quarterly, 67* (4), 443-454.

Bal, J., Berger, J., Jonge, J. de, Oudmaijer, S. & Tan, S. (2007). *Remediërende programma's rekenen en taal*. EIM-publicatie. Geraadpleegd 17 januari 2010 via minocw.nl/documenten/BrochureAP154.pdf

Baltzer, J. (1986). *Taalvaardigheid in het Hoger Onderwijs; Inleidend en samenvattend rapport*. Amsterdam: SCO, Universiteit van Amsterdam.

Baltzer, J., Glopper, K. de & Schooten, E. van (1988). *De taalvaardigheid van eerstejaars HBO-studenten*. Amsterdam: SCO, Universiteit van Amsterdam.

Basic Instructor Training. Geraadpleegd 4 april 2012 via www.tlcsem.com/blessonplan.htm

Beetsma, Y. (2010). *Effectieve kenmerken van een digitaal biologie practicum in het hoger onderwijs*. Proefschrift. Rijksuniversiteit Groningen: UOCG.

Beason, L. (2001). Ethos and Error: How Business People React to Errors. *College Composition and Communication, 53* (1), 33-64.

Beijer, J., Gangaram Panday, R. & Hajer, M. (2010). Taalbeleid in de steigers: naar een brede aanpak van taalonderwijs voor studie en beroep op de Hogeschool Utrecht. In E. Peters & T. Van Houtven (Red.), *Taalbeleid in het hoger onderwijs: de hype voorbij?* (pp. 165-176). Leuven: Acco.

Beijer, J. & Hajer, M. (2007). *Op weg naar taalbeleid in de Hogeschool Utrecht*. Notitie Lectoraat Lesgeven in de Multiculturele School, Hogeschool Utrecht.

Berntsen, A. & Gangaram Panday, R. (2007). Beter taal in het hoger onderwijs; een extra taak voor alle opleiders in het hbo. *Les, 25* (146), 8-10.

Bertina, M. (2006). *Gewoon goed Nederlands*. Amsterdam: Boom Onderwijs.

Bezooijen, R. van (2003). Stenen des aanstoots. Hoe ergerlijk kan taal zijn? *Onze Taal, 72*, (2), 36-38.

Binder, C. & Watkins, C. L. (1990). Precision Teaching and Direct Instruction: Measurably superior instructional technology in schools. *Performance Improvement Quarterly, 3* (4), 74-96.

Blom, A. (2006). Nt2-les zonder grammatica. *Levende Talen Tijdschrift, 7* (4), 20-27.

Bochardt, I. M. (1985). *Taalvaardigheid in het hoger onderwijs; Een enquête onder de eerstejaars studenten*. Amsterdam: SCO, Universiteit van Amsterdam.

Bochardt, I. M. (1986). Studenten hebben vooral moeite met schriftelijke taalvaardigheden: weinig verschil tussen hbo'ers en wo'ers. *Didaktief, 16* (4), 40-42.

Boes, A. W. (2002). *Toetsen, zin en onzin, een kritische beschouwing*. Geraadpleegd 2 januari 2012 via jenaplan.nl/cms/upload/docs/toetsen_zin_en_onzin.pdf

Bons, M. (2011, 16 november). Stop met al die zinloze bezigheidstherapietjes. *De Volkskrant*.

Bonset, H. (2011). Taalkundeonderwijs: Veel geloof, weinig empirie. *Levende Talen Magazine, 98* (2), 12-16.

Bonset, H. & Hoogeveen, M. (2007). *Schrijven in het basisonderwijs. Een inventarisatie van empirisch onderzoek in het perspectief van leerplanontwikkeling*. Enschede: SLO.

Bonset, H. & Braaksma, M. (2008). *Het schoolvak Nederlands opnieuw onderzocht. Een inventarisatie van onderzoek van 1997 tot en met 2007*. Enschede: SLO.

Borst, P. (2009, 5 december). Geestschrijven. *NRC Handelsblad*.

Bouma, J. (2007a, 13 januari). Ook de blondste leerlingen. *NRC Handelsblad*.

Bouma, J. (2007b, 20 januari). Maar ze kunnen wel goed praten. *NRC Handelsblad*.

Bout, M. & Bruijn, H. de (2007). *Basisvaardigheden Spelling voor de pabo*. Groningen: Wolters-Noordhoff.

Braas, C. & Krijgsman, J. (2005). *Taaltopics Formuleren* (2e druk). Groningen: Wolters-Noordhoff.

Braas, C. & Pas, L. van der (2006). *Taaltopics Spelling* (4e druk). Groningen: Wolters-Noordhoff.

Breland, H. M. (1983). *The Direct Assessment of Writing Skill: A Measurement Review.* College Board Report No. 83-6. New York: College Entrance Examination Board.

Breland, H. M., Bonner, M. W. & Kubota, M. Y. (1995). *Factors in Performance on Brief, Impromptu Essay Examinations*. College Board Report No. 95-4. New York: College Entrance Examination Board.

Breland, H. M. & Gaynor, J. L. (1979). A Comparison of Direct and Indirect Assessment of Writing Skill. *Journal of Educational Measurement, 16* (2), 119-128.

Breland, H. M. & Jones, R. J. (1982). *Perceptions of Writing Skill*. College Board Report No. 82-4. New York: College Entrance Examination Board.

Brink, T. van den (2007). *Rapportage taalvaardigheid*. Universiteit Utrecht, opleiding Geschiedenis, publicatie Onderwijscommissie. Geraadpleegd 20 augustus 2008 via www2.let.uu.nl/Solis/geschiedenis/mededelingen/Rapportage taalvaardigheid - Thomas van den Brink - 23 april 2007 - M (2).pdf

Broekkamp, H. & Hout-Wolters, B. van (2007). The Gap Between Educational Research and Practice. *Educational Research and* Evaluation, *13* (3), 209-220.

Bruffee, K. A. (1984). Collaborative Learning and the "Conversation of Mankind". *College English, 46* (7), 635-652.

Burt, M. (2011, november). *Scaling-up: The Right approach*. Debat WISE (World Innovation Summit for Education), Doha. Geraadpleegd 11 november 2011 viawise-qatar.org/content/25-scaling-right-approach

Butler, A. C., Karpicke, J. D. & Roediger, H. L. (2008). Correcting a Metacognitive Error: Feedback Increases Retention of Low-Confidence Correct Responses. J*ournal of Experimental Psychology: Learning, Memory, and Cognition, 34* (4), 918-928.

Camara, W. J. (2003). Scoring the Essay on the SAT Writing Section. *Research Summary,* College Board, 1-3.

Cambiumned. Geraadpleegd oktober 2008 via cambiumned.nl

Camstra, B., Van Dijk, T. & Van der Avoird, W. (1979). *Leren met de computer: eindrapport van het Plato-proefproject.* Amsterdam: COWO, Centrum Onderzoek voor Wetenschappelijk Onderwijs, Universiteit van Amsterdam.

Castagna, G. (2008, 4 februari). Correct spellen lukt niet meer. *Spits*.

CBS (2009). Geraadpleegd 10 april 2012 via www.cbs.nl/nl-NL/menu/themas/bedrijven/publicaties/artikelen/archief/2009/2009-arbeidsproductiviteit-exporterende-bedrijven-2005-art.htm

Chang, K. E., Sung, Y. T. & Chen, I. D. (2002). The Effect of Concept Mapping to Enhance Text Comprehension and Summarization. *The Journal of Experimental Education, 71* (1), 5-23.

Charney, D. (1984). The Validity of Using Holistic Scoring to Evaluate Writing: A Critical Overview. *Research in the Teaching of English, 18* (1), 65-81.

Coffman, W. E. (1966). On the Validity of Essay Tests of Achievement. *Journal of Educational Measurement, 3* (2), 151-156.

Connor, U. (1990). Linguistic/Rhetorical Measures for International Persuasive Student Writing. *Research in the teaching of English, 24* (1), 67-87.

Connors, R. J. & Lunsford A. A. (1988). Frequency of Formal Errors in Current College Writing, or Ma and Pa Kettle Do Research.*College Composition and Communication, 39* (4), 395-409.

Coombs, K. M. (1998, 24 maart). Honest follow-through needed on this project. *The Washington Times*. Geraadpleegd 10 januari 2009 via mathematicallycorrect.com/honestft.htm

Cooper, P. L. (1984). *The Assessment of Writing Ablitiy: A Review of Research*. GRE Board Research Report GREB No. 82-15R. Princeton: Educational Testing Service.

Cooper, J. O., Heron, T. E. & Heward, W. L. (2007). *Applied Behavior Analysis* (2nd edition). Upper Saddle River, NJ: Pearson/Merrill/Prentice Hall.

Coughlan, S. (2011, 14 juli). Spelling mistakes 'cost millions' in lost online sales. *BBC News.* Geraadpleegd 15 september 2011 via bbc.co.uk/news/education-14130854

Coughlan, S. (2012, 13 februari). MIT launches free online 'fully automated' course. *BBC News.* Geraadpleegd via bbc.co.uk/news/education-17012968

Creemers, B. P. M. (1991). *Effectieve instructie. Een empirische bijdrage aan de verbetering van het onderwijs in de klas.* Den Haag: SVO.

Creemers, B. P. M. (1994). *The effective classroom.* Londen: Cassell.

Creemers, B. P. M. & Kyriakydes, L. (2008). *The dynamics of educational effectiveness.* Abingdon: Routledge.

Cyr, M. D. (2011). Do Them No Favors, Tell Them No Lies. *The Chronicle of Higher Education*. Geraadpleegd 22 augustus 2011 via chronicle.com/article/D0-Them-No-Favors-Tell-Them/128583/

Daniëls, W. (2006). *Wolters' Nederlands in je pocket* (herziene druk). Groningen: Wolters-Noordhoff.

DeSantis, N. (2012, January 23). Stanford Professor Gives Up Teaching Position, Hopes to Reach 500,000 Students at Online Start-Up. *The Chronicle of Higher Education*. Geraadpleegd 15 februari 2012 via chronicle.com/blogs/wiredcampus/stanford-professor-gives-up-teaching-position-hopes-to-reach-500000-students-at-online-start-up/35135

De Stentor (2011, 22 december). Helft scripties journalistiek Windesheim onder de maat. Geraadpleegd 28 december 2011 viadestentor.nl/nieuws/algemeen/binnenland/10125651/Helft-scripties-journalistiek-Windesheim-onder-de-maat.ece

Deygers, B. & Kanobana, S. (2010). Taaltoetsen: waarom, wat en hoe? In E. Peters & T. Van Houtven (Red.), *Taalbeleid in het hoger onderwijs: de hype voorbij?* (pp. 23-35). Leuven: Acco.

Dijck, M. van, Elshout, J. van den & Hoogeveen, K. (1985). *Van voortgezet onderwijs naar HBO: Problemen, Oorzaken, Oplossingen*. Den Haag: SVO.

Dijkma, I. K. (2010, maart). *Creativity.* Presentatie bij IMEET (International Master of Entrepeneurship Education & Training), Hanzehogeschool Groningen.

Dijkstra, E.W. (1969). *Notes on Structured Programming. TH Report 70-WSK-03. Second Edition, April 1970. EWD 249*. Eindhoven: Technische Hogeschool Eindhoven, Onderafdeling der Wiskunde. Geraadpleegd via http://en.wikipedia.org/wiki/Edsger_W._Dijkstra op 24-03-2014.

Dijkstra, B. A. & Delden, J. van (1996). *Repetitieboekje Nederlands* (5e druk). Groningen: Wolters-Noordhoff.

Dikli, S. (2006). An Overview of Automated Scoring of Essays. *Journal of Technology, Learning, and Assessment, 5* (1) 1-36. Geraadpleegd 20 februari 2014 via ejournals.bc.edu/ojs/index.php/jtla/article/view/1640

Driessen, C., Westhoff, G., Haenen, J. & Brekelmans, M. (2008). A qualitative analysis of language learning tasks: the design of a tool. *Journal of Curriculum Studies, 40* (6), 803-820.

Driscoll, M. P. (2000). *Psychology of Learning for Instruction*. Boston: Allyn & Bacon.*dtkompas*. Geraadpleegd september 2008 via dtkompas.nl

Eerden, A. van & Es, M. van (2010, mei). *Measurement of Basic Writing Skill of First-Year Students in Higher Education*. Paper gepresenteerd bij de 10th ABC Conference (Association for Business Communication), Antwerpen.

Eerden, A. van, Es, M. van & Werf, M. P. C. van der (2011a, januari). *Two Reliable Methods to Measure Basic Text Quality bij Counting Errors in Texts*. Paper gepresenteerd bij ICSEI (International Congress for School Effectiveness and Improvement), Limassol, Cyprus.

Eerden, A. van & Es, M. van (2011b, januari). *A Reliable Method to Measure Basic Writing Skill by Counting Confirmed Errors which Shows that Dutch First-Year Students in Higher Education on Average Make 40 to 80 Errors in One Side A4.* Paper gepresenteerd bij de RIS3 Conference (Rhetoric in Society lll), Antwerpen.

Eerden, A. van, Es, M. van & Werf, M. P. C. van der (2011c, juni). *Een betrouwbare methode voor het meten van basale schrijfvaardigheid.* Paper gepresenteerd bij de ORD (Onderwijs Research Dagen), Maastricht.

Elbers, H. (2011). Universitair diploma goedkoper. In *Webmagazine* van het CBS. Geraadpleegd 11 februari 2012 via cbs.nl/nl-NL/menu/themas/onderwijs/publicaties/artikelen/archief/2011/2011-3528-wm.htm

Elliot, N. (2005). *On a Scale: a Social History of Writing Assessment in Amercia*. New York: Peter Lang Publishing.

Ellis, R. (2003). *Task-based Language Learning and Teaching*. Oxford: Oxford University Press.

Engelstalige Wikipedia (2014). Equivalence principle. Geraadpleegd 15 februari 2014 via en.wikipedia.org/wiki/Equivalence_principle

Engelstalige Wikipedia (2014). Cohen's kappa. Geraadpleegd 25 februari 2014 via en.wikipedia.org/wiki/Cohen's_kappa

Es, M. van (1980). *Zeer korte methodologie voor onderwijsevaluatie*. Amsterdam: COWO, Universiteit van Amsterdam.

Es, M. van, Velthuijsen, A., Neervoort, T. (1980). *De konstruktie van een werkboek om de bestudering van een studieboek te vereenvoudigen en de meningen van de studenten over het werken hiermee.* Amsterdam: COWO, Universiteit van Amsterdam.

Es, M. van (1985). Wat is een goede studietekst? In G. van der Veen (Red.), *Onderwijs in druk. Leerteksten kiezen, schrijven, vormen en drukken* (pp. 21-38). Deventer: Kluwer/Van Loghum Slaterus.

Es, M. van, Dijkhuizen A.D. (1987). *Lesmaker voor MSX en MSX2 met diskdrive.* Oosterend: Stark-Texel.

Es, M. van, Dijkhuizen A.D. (1988). *Handleiding Lesmaker voor DOS.* Oosterend: Stark-Texel.

Examenblad.nl (2011). Geraadpleegd op 7 juni 2011 via examenblad.nl

Expertgroep Doorlopende Leerlijnen Taal en Rekenen (2008). *Over de drempels met taal.* Enschede: SLO.

Farmer, L. (2008). *A Study of an Attempt to Improve the Reliability of Teachers' Holistic Scores of Elementary Writing through In-House Professional Development.* Proefschrift. University of Central Forida.

Feenstra, H. (2011, juni). *Objectieve schrijfvaardigheidstoetsen: een onderzoek naar validiteit.* Poster ORD (Onderwijs Research Dagen), Maastricht.

Foer, J. (2011, oktober). *Het geheugen na de zondvloed*. Van der Leeuw-lezing, Groningen.

Fox, E. J. (2004). The Personalized System of Instruction: A flexible and Effective Approach to Mastery Learning. In D. J. Moran & R. W. Malott (Eds.), *Evidence-Based Educational Methods* (pp. 201-221). San Diego: Elsevier Academic Press.

Garinger, D. (2002). Textbook Selection for the ESL Classroom. *Eric Clearinghouse on Languages and Linguistics*. Geraadpleegd 5 januari 2014 via www.cal.org/resources/Digest/0210garinger.html

Gamaroff, R. (2000). Rater Reliability in Language Assessment: The Bug of all Bears. *System, 28* (1), 31-53.

Gangaram Panday, R., Droop, M. & Rutten, E. (2008). Taalvaardigheid in beeld. Een verkennend onderzoek binnen de pilots Taalcompetenties voor studie en beroep. *Onder de loep 13*.

Gein, J. van de (2010). Komd een kind van de basisschool. Onderwijscommissie onderschat spelvaardigheden basisscholieren. *Onze Taal, 79* (9), 228-231.

Gelder, L. van, Oudkerk Pool, T., Peters, J. & Sixma, J. (Red.). (1973). D*idactische analyse: werk- en studieboek 1* (2e druk). Groningen: Wolters-Noordhoff.

Genootschap Onze Taal (2013). Geraadpleegd 22 juli 2013 via onzetaal.nl.

Gerrits, R. (2008, 25 januari). Rekenen en taal moeten beter - maar hoe? *De Volkskrant*.

Gertsbakh, I. (2003). *Measurement Theory for Engineers*. Berlijn/New York: Springer Verlag.

Gilbert, M. B. (2004). Grammar and Writing Skills: Applying Behavior Analysis. In D. J. Moran & R. W. Malott (Eds.), *Evidence-Based Educational Methods* (pp. 361-374). San Diego: Elsevier Academic Press.

Godshalk, F. I., Swineford, F. & Coffman, W. E. (1966). *The Measurement of Writing Ability.* New York: College Entrance Examination Board.

Google Books Ngram Viewer. Geraadpleegd 18 mei 2012 via books.google.com/ngrams/graph?content=learning+machine&year_start= 1800 &year_end=2000&corpus=5&smoothing=0

Google Search: College Board. Geraadpleegd 13 september 2013 via www.google.nl/search?q=college+board&ie=utf-8&oe=utf-8&rls=org.mozilla:nl:official&client=firefox-a&gws_rd=cr&ei=e61JUpaxGILNtAb35IH4Bg

Graham, S. (2006). Writing. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of Educational Psychology* (pp. 457-478). Mahwah: Lawrence Erlbaum Associates.

Graham, S. & Perin, D. (2007a). *Writing Next: Effective strategies to improve writing of adlolescents in middle and high Schools.* New York: Alliance for Excellent Education.

Graham, S. & Perin, D. (2007b). A Meta-Analysis of Writing Instruction for Adolescent Students. *Journal of Educational Psychology, 99* (3), 445-476.

Greenberg, K. L. (1992). Validity and Reliability Issues in the Direct Assessment of Writing. *Writing Program Administration, 16* (1-2), 7-22.

Groene Boekje: Woordenlijst Nederlandse Taal (2005). Nederlandse Taalunie. Den Haag: Sdu.

Groot, A. D. de (1983a). Is de kwaliteit van onderwijs te beoordelen? In B. P. M. Creemers, W. T. J. G. Hoeben & K. Koops (Red.), *De kwaliteit van het onderwijs* (pp. 54-72). Haren: RION/Groningen: Wolters-Noordhoff.

Groot, A. D. de (1983b). Gebruik en nut van studietoetsen. In A. G. Vroon & S. E. M. Everwijn (Red.), *Handboek voor de onderwijspraktijk* deel 4 (pp. 1-8). Deventer: Van Loghum Slaterus.

Groot, A. D. de (1993). *Denken over onderwijs*. *Analyses en kritieken van A. D. de Groot.* Den Haag: SVO.

Gulliksen, H. (1936). The Content Reliability of a Test. *Psychometrika, 1* (3) 189-194.

Hajer, M. (2005). Taalgericht vakonderwijs, Tijd voor een nieuw vijfjarenplan. *Levende Talen Tijdschrift, 6* (1), 3-11.

Hanushek, E. A. & Rivkin, S. G. (2010, januari). Generalizations about Using Value-Added Measures of Teacher Quality. Paper gepresenteerd bij The Annual Meeting of the American Economic Association, Atlanta, GA. Geraadpleegd 10 januari 2012 via usapr.org/paperpdfs/54.pdf

Harm, Y. (2008). *Het effect van taalfouten op tekstwaardering*. Scriptie. Universiteit Utrecht, Taal- en Cultuurstudies. Geraadpleegd op 15 december 2009 via igitur-archive.library.uu.nl/student-theses/2008-0902-203534/UUindex.html

Hayes, J. R. & Flower, L. S. (1980). Identifying the Organization of Writing processes. In L. Gregg & E. R. Steinberg (Eds.), *Cognitive Processes in Writing* (pp. 3-30). Hillsdale, NJ: Lawrence Erlbaum.

Heijmer, T. & Vonk, R. (2002). Effecten van een regionaal accent op de beoordeling van de spreker. *Nederlands Tijdschrift voor de Psychologie*, *57*, 108-113.

Heward, W. L. (2005). Reasons Applied Behavior Analysis Is Good for Education and Why Those Reasons Have Been Insufficient. In W. L. Heward, T. E. Heron, N. A. Neef, S. M. Peterson, D. M. Sainato, G. Cartledge, . . . J. C. Dardig (Eds.), *Focus on Behavior Analysis in Education: Achievements, Challenges, and Opportunities* (pp. 316-348). Upper Saddle River, NJ: Pearson/Merrill/Prentice Hall.

Hofstee, W. K. B. (2008). De mythe van de sterke benen. In M. P. C. van der Werf (Red.), *Mythes in het onderwijs* (pp. 38-47). Amsterdam: SWP.

Hogen, R. van (1997). *Praktische cursus Formuleren* (2e druk). Groningen: Wolters-Noordhoff.

Hogen, R. van & Rietstap, E. (2007). *Basisvaardigheden Taal*. Groningen: Wolters-Noordhoff.

Hogeschooltaal (2008, april). Deventer: Kluwer.

Hogeweg, R. (2003). *Dat d/t gedoe*. Groningen: Wolters-Noordhoff.

Holland, J. G. (1960). Teaching Machines: An Application of Principles from the Laboratory. *Journal of the Experimental Analysis of Behavior, 3*, 275-287.

Holland, J. G. & Skinner, B. F. (1961). The Analysis of Behavior: A Program for Self-Instruction. New York: McGraw-Hill.

Holland, J. G., Solomon, C., Doran, J. & Frezza, D. A. (1976). *The Analysis of Behavior in Planning Instruction.* Reading: Addison-Wesley.

Horst, J. ter & Molenaar, A. (2006). *Zakelijk schrijven*. Bussum: Coutinho.

Hudson, R. (2001). Grammar teaching and writing skills: the research evidence. *Syntax in the Schools*, *17*, 1-6.

Hull, C. L. (1932). The Goal Gradient Hypothesis and Maze Learning. *Psychological Review, 39* (1), 25-43.

Hyslop, N. B. (1990). Evaluating Student Writing: Methods and Measurement. *ERIC Clearinghouse on Reading and Communication.* Geraadpleegd 6 januari 2012 via ericae.net/db/edo/ED315785.htm

Inspectie van het Onderwijs (2009). *Het taalonderwijs op taalzwakke en taalsterke scholen. Een onderzoek naar de kenmerken van het taalonderwijs op basisscholen met lage en hoge taalresultaten*. Utrecht: Inspectie van het Onderwijs.

Inspectie van het Onderwijs (2010). *Het onderwijs in het schrijven van teksten. De kwaliteit van het schrijfonderwijs in het basisonderwijs*. Utrecht: Inspectie van het Onderwijs.

Jager, O. & Neijndorff, R. (2009). Taaltrivia. *Digitale nieuwsbrief, 5*.

Jager, S. (2009). *Towards ICT-Integrated Language Learning. Developing an Implementation Framework in terms of Pedagogy, Technology and Environment*. Proefschrift. Rijksuniversiteit Groningen.

Jansen, F. (2007). Spelfouten op internetfora corrigeren? Taaladviesblog *Onze Taal*. Geraadpleegd 17 januari 2010 via onzetaal.nl/homofkuit/h0710.php

Jansen, M. M. & Wesdorp, H. (1973). De waarde van eindexamenopstelcijfers. *Levende Talen* (297*)*, 191-204.

Janssen, T., Dam, G. ten & Hout-Wolters, B. van (2002). *Vaardigheden voor zelfstandig leren*. Assen: Koninklijke Van Gorcum BV.

Janssen, D., Jansen, F. & Kinkhorst, G. (2007). *Zakelijke Communicatie deel 1* (5e druk). Groningen/Houten: Noordhoff.

Janssen, D., Jansen, F. & Kinkhorst, G. (2007). *Zakelijke Communicatie deel 2* (5e druk). Groningen/Houten: Noordhoff.

Jaynes, E. T. (1989). Clearing up Mysteries - The Original Goal. In J. Skilling (Eds.), *Maximum Entropy and Bayesian Methods* (pp. 1-27). Dordrecht: Kluwer Academic.

Jenson, W. R., Sloane, H. N. & Young, K. R. (1988). *Applied Behavior Analysis In Education: A Structured Teaching Approach*. Englewood Cliffs, NJ: Prentice Hall.

Joling, E. (2001). *Onderwijzen als Doelgerichte Activiteit*. Geraadpleegd 27 oktober 2011 via staff.science.uva.nl/~joling/vakdidactiek/mda.html

Johnson, K. & Street, E. M. (2004). The Morningside Model of Generative Instruction: An Integration of Research-Based Practices. In D. J. Moran & R. W. Malott (Eds.), *Evidence-Based Educational Methods* (pp. 247-265). San Diego: Elsevier Academic Press.

Juf Melis. Geraadpleegd oktober 2008 via jufmelis.nl

Kam, F. de (2009, 13 juni). Heeft Nederland de aardgasbaten goed besteed? *NRC Handelsblad*.

Karpicke, J. D. & Roediger, H. L. (2008). The Critical Importance of Retrieval for Learning. *Science, 319* (5865), 966-968.

Karpicke, J. D. & Blunt, J. R. (2011). Retrieval Practice Produces More Learning than Elaborative Studying with Concept Mapping. *Science, 331* (6018), 772-775.

Kas, W. (1996). *Spelbewust* (2e druk). Zutphen: Thieme.

Karstanje, P. N. (1983). Over doeldenken en kwaliteitsbepaling. In B. P. M. Creemers, W. T. J. G. Hoeben & K. Koops (Red.), *De kwaliteit van het onderwijs* (pp. 73-77). Haren: RION/Groningen: Wolters-Noordhoff.

Keller, F. S. (1968). "Goodbye, Teacher...". *Journal of Applied Behavior Analysis, 1,* 78-89.

Kirschner, P. A., Sweller, J. & Clark, R. E. (2006). Why Minimal Guidance During Instruction Does Not Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiental and Inquiry-Based Teaching. *Educational Psychologist, 41* (2), 75-86.

Klein, M. & Visscher, M. (2006). *Praktische cursus spelling* (5e druk). Groningen: Wolters-Noordhoff.

Kloet, L., Renkema, J. & Wijk, C. van (2003). Waarom foutloos schrijven? Het effect van taalfouten op tekstwaardering, imago en overtuigingskracht. In L. van Waes (Red.), *Studies in taalbeheersing 1* (pp. 270-279). Assen: Koninklijke Van Gorcum.

Knispel, K. (2008). *Zakelijke Communicatie - Schriftelijk* (2e druk). Amsterdam: Pearson Education.

Krüger, M. (2008, november). *Evidence bases werken in het onderwijs. Het belang van een onderzoekende houding*. Lezing Lectoraat Integraal Jeugdbeleid, Hanzehogeschool Groningen.

Kuhn, K. F. (1996). *Basic Physics. A Self-Teaching Guide* (2nd edition). New York: John Wiley & Sons, Inc.

Kuiken, F. (2010). Taalbeleid in het hoger onderwijs: verslag van werk in uitvoering. In E. Peters & T. Van Houtven (Red.), *Taalbeleid in het hoger onderwijs: de hype voorbij?* (pp. 195-200). Leuven: Acco.

Kuiper, M. (1996). *Meer uitleg, meer kennis? Het effect van minimale versus uitgebreide uitleg bij feedback na een fout antwoord op het leerresultaat in Computer Ondersteund Onderwijs. Doctoraalscriptie Toegepaste Taalkunde.* Groningen: Instituut voor Toegepaste Taalkunde, Rijksuniversiteit Groningen.

Kulik, J. A., Kulik, C.-L. C. & Cohen, P. A. (1979). A Meta-Analysis of Outcome Studies of Keller's Personalized System of Instruction. *American Psychologist*, *34* (4), 307-318.

Laar, F. van de (2009, 19 april). Ontdooi het Nederlands. 'Het meisje die' kan best. *NRC Handelsblad*.

Lambay, F. (2011, november). *Scaling-up: The Right approach*. Debat WISE (World Innovation Summit for Education), Doha. Geraadpleegd 11 november 2011 via wise-qatar.org/content/25-scaling-right-approach

Lamie, E. (2000 April-May). *SED 590*. Geraadpleegd 20 februari 2014 via www.cs.csustan.edu/~lamie/sed590/main.htm

Lindvall, C. M. & Bolvin, J. O. (1967). Programmed Instruction in the Schools: An Application of Programing Principles in 'Individually Prescribed Instruction'. In *Programmed Instruction: Sixty-Sixth Yearbook of the National Society for the Study of Education, part II* (pp. 217-254). Chicago: The University of Chicago Press.

Loerts, H. (2012). *Uncommon Gender. Eyes and brains, native and second language learners, & grammatical gender.* Proefschrift. Rijksuniversiteit Groningen: Faculteit der Letteren.

Lowyck, J. (1994). Teaching Effectiveness: An overview of studies. *Tijdschrift voor Onderwijsresearch, 19*, 17-25.

Lücker-de Boer, F. (2010, maart). *Creativity.* Presentatie bij IMEET (International Master of Entrepeneurship Education & Training), Hanzehogeschool Groningen.

Lunsford, A. A. & Lunsford, K. J. (2008). "Mistakes Are a Fact of Life": A National Comparative Study. *College Composition and Communication, 59* (4), 781-806.

Macdonald, A. (2013). *General Relativity in a Nutshell*. Geraadpleegd 23 februari 2014 via http://faculty.luther.edu/~macdonal

Macrorie, K. (1971). *Telling Writing*. Springfield, MO: Hayden Book Company.

Malott, R. W. (2008). *Principles of Behavior* (6th edition). Upper Saddle River, NJ: Pearson/Prentice Hall.

Mante, J. (2006). *Een Goede Spelling*. Utrecht: ThiemeMeulenhoff.

Matthews, W. J. (2003). Constructivism in the Classroom: Epistemology, History, and Empirical Evidence. *Teacher Education Quarterly, 30* (3), 51-64.

Mayer, R. E. (2008). *Learning and Instruction* (2nd edition). Upper Saddle River, NJ: Pearson/Prentice Hall.

Mechner, F. (1965). Science Education and Behavioral Technology. In R. Glaser (Ed.), Teaching Machines and Programed Learning, II (pp. 441-508). Washington, D.C.: National Education Association.

Meuffels, B. (2002). De beoordeling van schrijfvaardigheid in de tweede fase. In L. Derriks (Red.), *Retoriek en praktijk van het schoolvak Nederlands 2002* (pp. 129-136). Gent: Academia Press.

Miekley, J. (2005). ESL Textbook Evaluation Checklist. *The Reading Matrix, 5* (2). Geraadpleegd 4 januari 2014 via www.readingmatrix.com/reading_projects/miekley/project.pdf

Miller, B. & McCardle, P. (2011). Reflections on the need for continued research on writing. *Reading and Writing, 24,* 121-132.

Minne, B., Steeg, M. van der & Webbink, D. (2007). *De maatschappelijke opbrengsten van onderwijs.* Den Haag: CPB.

Mirande, M. (2006). *De onstuitbare opkomst van de leermachine*. Assen: Van Gorcum.

Montens, F. & Sciarone, A. G. (1992). *De Delftse methode: Nederlands voor buitenlanders* (8e druk). Amsterdam: Boom.

Moons, A., Bovenhoff, M. & Latjes, G. (2008). *Basisboek Spelling*. Groningen: Wolters-Noordhoff.

Muiswerk (2008, september). Uithoorn: Muiswerk Educatief.

Mukundan, J., Hajimohammadi, R. & Nimehchisalem, V. (2011). Developing An English Language Textbook Evaluation Checklist. *Contemporary Issues In Education Research, 4* (6), 21-28.

Muralidharan, K. & Sundararaman, V. (2011). Teacher Performance Pay: Experimental Evidence from India. *Journal of Political Economy, 119* (1), 39-77.

Nadler, R. (1998, 1 juni). Failing Grade. *National Review, 50* (10), 38-39. Geraadpleegd 10 januari 2009 via old.nationalreview.com/01jun98/nadler060198.html

Nedercom (2008, augustus). Roden: Nedercom Eduware.

Nieuwenhuijsen, P. (2010). Fouten, vergissingen en Nederlands-B. In E. Peters & T. Van Houtven (Red.), *Taalbeleid in het hoger onderwijs: de hype voorbij?* (pp. 210-213). Leuven: Acco.

Nunnally, J. C. (1967). *Psychometric Theory.* New York: McGraw-Hill.

Oderkerk, B. (2007, 27 juni). Journalisten kunnen ook niet rekenen. *Spits*.

Onderzoek Onderwijs Nederlands 1969-2002. *Taalunieversum*. Geraadpleegd 12 maart 2010 via taalunieversum.org/onderwijs/onderzoek/1969-1997/

Onrust, M., Verhagen, A. & Doeve, R. (1999). *Formuleren*. Houten: Bohn Stafleu Van Loghum.

Otter, M. E. & Schoonen, R. (1996). *Aap, noot, niets .... of het spook van de ontlezing in het basisonderwijs*. Amsterdam: SCO, Universiteit van Amsterdam.

Ottjes, A. (2009). *Communiceren met een reisbrochure. Het effect van taalkundige fouten.* Onderzoek Onderzoekspracticum. Rijksuniversiteit Groningen, Communicatie- en Informatiewetenschappen.

Pak, D. (2007). *Vlekkeloos Nederlands: spelling en stijl compleet* (2e druk). Den Haag: Dick Pak.

Palmer, O. (1961). Sense or Nonsense? The Objective Testing of English Compositon. *The English Journal, 50* (5), 314-320.

Parsons, J. A. & Polson, D. (2000). Engelmann's Direct Instruction and Project Followthrough. *Psychology 387: Learning*. Geraadpleegd op 4 november 2011 via psych.athabascau.ca/html/387/OpenModules/Engelmann/

Paulson, E. J., Alexander, J. & Armstrong, S. (2007). Peer Review Re-Viewed: Investigating the Juxtaposition of Composition Students' Eye Movements and Peer-Review Processes. *Research in the Teaching of English, 41* (3), 304-335.

Pear, J. J. & Martin, T. L. (2004). Making the Most of PSI with Computer Technology. In D. J. Moran & R. W. Malott (Eds.), *Evidence-Based Educational Methods* (pp. 223-243). San Diego: Elsevier Academic Press.

Peters, E. (2010). Inleiding. In E. Peters & T. Van Houtven (Red.), *Taalbeleid in het hoger onderwijs: de hype voorbij?* (pp. 19-21). Leuven: Acco.

Peters, E., Houtven, T. van & El Morabit, Z. (2010). Is meten echt meer weten? Taalvaardigheid van instromende studenten in het hoger onderwijs in kaart gebracht. In E. Peters & T. Van Houtven (Red.), *Taalbeleid in het hoger onderwijs: de hype voorbij?* (pp. 51-66). Leuven: Acco.

Popham, W. J. (1983). Hedendaagse opvattingen over onderwijsevaluatie. In A. G. Vroon & S. E. M. Everwijn (Red.), *Handboek voor de onderwijspraktijk* deel 4 (pp. 1-24). Deventer: Van Loghum Slaterus.

Prenger, J, & Glopper, c. m. de (2011). *Schrijven om te leren bij Nederlands, Geschiedenis en Science*. Presentatie op de vijfentwintigste conferentie Het Schoolvak Nederlands, Den Haag.

Project X 2002. Geraadpleegd oktober 2008 via projectx2002.org

Purves, A. C. (1992). Reflections on Research and Assessment in Written Composition. *Research in the Teaching of English, 26* (1), 108-122.

Raedts, M. (2011, november). *De leereffecten van 'leren-door-observeren' in vergelijking met 'leren-door-doen' en 'leren via modeloplossingen'*. Presentatie op de vijfentwintigste conferentie Het Schoolvak Nederlands, Den Haag.

Raseks, A. E., Esmae'li, S., Ghavamnia, M. & Rajabi, S. (2010). Don't Judge a Book by its Cover: Text Book Evaluation in the EFL Settings. *The Journal of International Social Research, 3* (14), 448-461.

Reijn, G. (2008, 30 januari). Eerstejaars VU krijgen taaltoets. *De Volkskrant*.

Reijn, G. (2011, 20 december). Meer geld betekent niet altijd meer eten. *De Volkskrant*.

Renkema, J. (2005). *Schrijfwijzer* (4e druk). Den Haag: Sdu.

Renner, K. E. (1964). Delay of Reinforcement: A Historical Review. *Psychological Bulletin, 61* (5), 341-361.

Research Brief. The Center for Comprehensive School Reform and Improvement (2007, september). Writing Next. What does the research indicate concerning specific teaching techniques that will help adolescent students develop necessary writing skills? Geraadpleegd 30 november 2011 via centerforcsri.org/files/Center_RB_Sept07.pdf

Richards, J. C. (2001). *Curriculum Development in Language Teaching*. New York: Cambridge University Press.

Richards, J. C. (2010). Theories of Teaching in Language Teaching. In J. C. Richards & W. A. Renandya (Eds.), *Methodology in Language Teaching. An Anthology of Current Practice* (14th edition) (pp. 19-25). New York: Cambridge University Press.

Riesen, A. H. (1940). Delayed Reward in Discrimination Learning by Chimpanzees. *Comparative Psychology Monographs, 15* (5), 1-54.

Rijlaarsdam, G. (2011, november). *Academisch schrijven in voortgezet en hoger onderwijs: effectieve didactiek.* Inleiding subplenaire lezing op de vijfentwintigste conferentie Het Schoolvak Nederlands, Den Haag.

Rijlaarsdam, G., Braaksma, M., Couzijn, M., Janssen, T., Raedts, M., Steendam E. van, . . . Bergh, H. van den (2008). Observation of peers in learning to write, Practice and Research. *Journal of Writing Research, 1* (1), 53-83.

Rijlaarsdam, G. & Wesdorp, H. (1984). *Het beoordelen van taalvaardigheid in het onderwijs.* Amsterdam*:* SCO, Universiteit van Amsterdam.

Robinson, W. S. (1998). Towards a Theory of Error. *Teaching English in the Two-Year College, 26* (1), 50-60.

Rohrer, D. & Pashler, H. (2010). Recent Research on Human Learning Challenges Conventional Instructional Strategies. *Educational Researcher, 39* (5), 406-412.

Rose, M. (1985). The Language of Exclusion: Writing Instruction at the University. *College English, 47* (4), 341-359.

Rosenshine, B. (1997, maart). *The Case for Explicit, Teacher-led, Cognitive Strategy Instruction.* Paper gepresenteerd bij AERA (American Educational Research Association), Chicago. Geraadpleegd 13 maart 2008 via formapex.com/barak-rosenshine/616-the-case-for-explicit-teacher-led-cognitive-strategy-instruction

Rossen-Knill, D. & Lynch, K. (2000). A Method For Describing Basic Writers And Their Writing: Lessons From A Pilot Study. *Journal of Basic Writing*, *19* (2), 93-123.

Sanders, E. (2007, 16 februari). Nooit een zin met 'ik' beginnen? *Woordhoek*. Geraadpleegd 17 januari 2010 via weblogs.nrc.nl/woordhoek/2007/02/16/nooit-een-zin-met-ik-beginnen/

Sanders, E. (2008, 25 maart). Groter als. *NRC Handelsblad*.

Scheerens, J. (1997). *De bevordering van schooleffectiviteit in het basisonderwijs. Mogelijkheden tot "flankerend beleid" bij klassenverkleining*. Enschede: Vakgroep Onderwijsorganisatie en -management.

Scheerens, J & Bosker, R. J. (1997). *The foundation of educational effectiveness*. Oxford: Pergamon.

Scheerens, J., Luyten, H. & Ravens, J. van (2011). *Visie op onderwijskwaliteit met illustratieve gegevens over de kwaliteit van het Nederlandse primair en secundair onderwijs*. Onderzoek gesubsidieerd door NWO/PROO: Universiteit Twente.

Schilder, J. (2008). *Van verslag tot rapport*. Amsterdam: Boom Onderwijs.

Schooten, E. van (1988). *De constructie van een meerkeuzetoets voor het meten van schrijfvaardigheid*. Amsterdam: SCO, Universiteit van Amsterdam.

Schutte, I. & Veenker, H. (2009). *Oog voor etnische en culturele diversiteit*. Hanzehogeschool Groningen: Afdeling Studentzaken.

Shaw, E. J. & Kobrin, J. L. (2012). *The Sat Essay and College Performance: Understanding What Essay Scores Add to HSGPA and SAT*. College Board Research Report 2012-9. Geraadpleegd 5 september 2013 via research.collegeboard.org/publications/sat-essay-and-college-performance-understanding-what-essay-scores-add-hsgpa-and-sat

Sheldon, L. E. (1988). Evaluating ELT textbooks and materials. *ELT Journal*, *42* (4), 237-246.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin, 86* (2), 420-428.

Simons, P. R. J. (2000). Towards a Constructivistic Theory of Self-Directed Learning. In G. A. Straka (Ed.), *Conceptions of Self-Directed Learning: Theoretical and Conceptual Considerations* (pp. 155-169). Münster: Waxmann.

Simons, P. R. J. (2006). Waarom nieuw leren? *Blind, 7*.

Skinner, B. F. (1954). The Science of Learning and the Art of Teaching. *Harvard Educational Review, 24* (2), 86-97.

Skinner, B. F. (1958). Teaching Machines. *Science, 128*, 969-977.

Skinner, B. F. (1968). *The technology of learning*. New York: Appleton-Century-Crofts.

Slecht Nederlands. (15 april 2009). Weblog *nrc.nl.* Geraadpleegd op 14 maart 2010 via weblogs.nrc.nl/commentaar/2009/04/15/slecht-nederlands/

Steeg, M. van der, Vermeer, N. & Lanser, D. (2011). Nederlandse onderwijsprestaties in perspectief. *CPB Policy Brief*. Den Haag: Centraal Planbureau.

Steehouder, M., Jansen, C., Maat, K., Staak, J. van der, Vet, D. de, Witteveen, M. & Woudstra, E. (2006). *Leren Communiceren. Handboek voor mondelinge en schriftelijke communicatie* (5e herziene druk). Groningen/Houten: Noordhoff.

Steenbergen, H. (2009). *Vrije en reguliere scholen vergeleken. Een onderzoek naar de effectiviteit van Vrije scholen en reguliere scholen voor voortgezet onderwijs.* Proefschrift. Rijksuniversiteit Groningen: GION.

Steinmann, M. (1967). A Conceptual Review. The Measurement of Writing Ability by F. I. Godshalk; Frances Swineford; W. E. Coffman. *Research in the Teaching of English, 1* (1), 79-84.

Straalen, E. van (2009, november). *De taaltoets voor eerstejaars van de Vrije Universiteit Amsterdam*. Presentatie op de conferentie VO-HO, Wageningen.

Stroop, J. R. (1935). Studies of Interference in Serial Verbal Reactions. *Journal of Experimental Psychology, 18* (6), 643-662.

Stroop, J. (2007, 20 januari). Nieuw Nederlands Spelpeil. *NRC Handelsblad*.

Studiemeter (2008, september). Amersfoort: Deviant.

Taalniveau studenten schiet tekort. (2009). *Taaljournaal*, *14*. Geraadpleegd op 7 april 2010 via taalcentrum-vu.nl/fileadmin/user_upload/Documenten/Taaljournaal/2009_TJ_december.pdf

TaalONLINE (2008, september). Den Haag: Jager & Neyndorff.

Taba, H. (1962). *Curriculum Development: Theory and Practice*. New York: Harcourt, Brace & World.

Thorndike, E. L. (1920). A Constant Error in Psychological Ratings. *Journal of Applied Psychology, 4* (1), 25-29.

Tiggeler, E. (2007). *Vraagbaak Nederlands* (herziene druk). Den Haag: Sdu.

Tijd voor Onderwijs (2008). Den Haag: Sdu.

TiO (2008, april). Rosmalen: Bureau voor Educatieve Ontwerpen.

TLC Seminars (2009). Geraadpleegd 15 maart 2012 via www.tlcsem.com/binstructor.htm

Tversky, A & Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science, 211* (4481), 453-458.

Tyler, R. W. (1949). *Basic Principles of Curriculum and Instruction*. Chicago: The University of Chicago Press.

Tyree, A. L. (1997). *The Keller Plan at Law School*. Geraadpleegd 17 december 2008 via austlii.edu.au/~alan/j_leged.html

Van Dale onlinewoordenboek Nederlands. Geraadpleegd 21 oktober 2011 via surfdiensten3.vandale.nl.proxy-ub.rug.nl/

Vanmaele, L. (2002). *Leren Schrijven van Informatieve Teksten: Een ontwerponderzoek bij beginners secundair onderwijs*. Leuven: Studia Paedagogica 32.

Vargas, J. S. (2009). *Behavior Analysis for Effective Teaching*. New York: Routledge.

Veenman, S. (2001). *Directe Instructie.* Paper Katholieke Universiteit Nijmegen. Geraadpleegd 15 maart 2008 via daltondeventer.nl/ogw/DIRECTE-INSTRUCTIE-Veenman.doc

Ven, P. H. van de (1986). Honderd jaar kommer en kwel. *Moer, Tijdschrift voor het onderwijs in Nederlands*, *3*, 2-11.

Vernooy, K. (2011, oktober). *Taal en lezen is cruciaal*. Presentatie op de conferentie Op weg naar de excellente school, Pedagogische Academie, Hanzehogeschool Groningen.

Wagen-Huijskes, D. van der (2011). Taal is het belangrijkste wapen van een jurist. *Platform Communicatie. Nieuwsmagazine voor docenten en opleidings-management*, 4.

Wall, S. V. & Hull, G. A. (1989). The semantics of error: What do teachers know? In C. M. Anson (Ed.), *Writing and response: theory, practice, and research* (pp. 261-292). Urbana, IL: National Council of Teachers of English.

Wang, J. (2006). *Evaluating an EFL Textbook--A New English Course*. Geraadpleegd 14 februari 2014 via www.ling.lancs.ac.uk/groups/crile/docs/crile31wang.pdf

Werf, M. P. C. van der & Weide, M. G. (1991). Effectief onderwijs voor allochtone studenten.*Tijdschrift voor Onderwijsresearch*, *16* (4), 231-243.

Werf, M. P. C. van der (2005). *Leren in het studiehuis. Consumeren, construeren of engageren?* Oratie. Rijksuniversiteit Groningen: GION.

Werf, M. P. C. van der (2008). De mythe van de sterke benen. In M. P. C. van der Werf (Red.), *Mythes in het onderwijs* (pp. 27-38). Amsterdam: SWP.

Westen, W. van der (2003). Ondersteunend onderwijs Nederlands: het perspectief op een goede taalvaardigheid. In A. Mottard (Red.), *Retoriek en praktijk van het schoolvak Nederlands 2002* (pp. 207-219). Gent: Academia Press.

Westen, W. van der (2005). *Welgespeld*. Bussum: Coutinho.

Westen, W. van der (2006). 'Maatregelen die fruit brengen!' Een integrale aanpak van taalontwikkeling in een hogere beroepsopleiding. In D. Ebbers (Red.), *Retoriek en praktijk van het moedertaalonderwijs 2006* (pp. 115-123). Gent: Academia Press.

Westen, W. van der (2011a). Helder taalbeleid gaat verder dan eenmalig toetsen en beoordelen. *Platform Communicatie. Nieuwsmagazine voor docenten en opleidings-management*, 4.

Westen, W. van der (2011b, november). *Ontwikkeling van een instrument voor zelfbeoordeling schrijfvaardigheid*. Presentatie op de vijfentwintigste conferentie Het Schoolvak Nederlands, Den Haag.

Wikipedia (2013). Taalfout. Geraadpleegd 22 juli 2013 via nl.wikipedia.org/wiki/Taalfout

Williams, J. M. (1981). The Phenomenology of Error. *College Composition and Communication, 32* (2), 152-168.

Williams, D. (1983). Developing criteria for textbook evaluation. *ELT Journal, 37* (3), 251-255.

Williamson, D. M., Bennett, R. E., Lazer, S., Bernstein, J., Foltz, P. W., Landauer, T. K., . . . Sweeney, K. (2010). *Automated Scoring for the Assessment of Common Core Standards*. Educational Testing Service/The College Board.

Witte boekje: *Spellinggids van het Nederlands* (2006). Genootschap Onze Taal. Utrecht: Het Spectrum.

Wubs, E. & Nauta, M. (2010). *Syllabus Commercieel correct schrijven.* Hanzehogeschool Groningen: Instituut voor Marketing Management.

Zijlstra, H. (2012, 27 april). *Antwoorden op vragen van het lid Jadnanansing over de taalvaardigheid van hbo-studenten*. Ministerie van Onderwijs, Cultuur en Wetenschap.

Zimmerman, B. & Kitsantas, A. (2002). Acquiring writing revision and self-regulatory skill through observation and emulation. *Journal of Educational Psychology, 94* (4), 660-668.

Zuidweg, M. (2006, 25 november). Dat eeuwige wachten. *NRC Handelsblad.*

Zwiers, R. (2010). Taalbeleid voor de pabo. In E. Peters & T. Van Houtven (Red.), *Taalbeleid in het hoger onderwijs: de hype voorbij?* (pp. 201-212). Leuven: Acco.

Bijlagen

Bijlage 1 **Korte toelichting op enkele psychometrische begrippen**

De psychometrische begrippen die in dit proefschrift soms ter sprake komen, zijn niet algemeen bekend. Daarom worden in deze bijlage enkele begrippen kort toegelicht.

Percentage verklaarde variantie

Wat in de psychometrie afwijkt van veel andere wetenschappen, is dat men niet focust op het significant zijn van de correlatie (dat wil zeggen: in de populatie aantoonbaar verschilt van 0), maar vooral in de hoogte van de correlatie geïnteresseerd is. Twee correlaties van 0.30 en 0.90 kunnen beide significant zijn, maar verschillen belangrijk in waarde. Wanneer de correlatie tussen A en B 0.30 zou zijn, is dat dermate weinig dat er in de praktijk vaak gedaan zal kunnen worden alsof er geen correlatie is. Wanneer de correlatie 0.90 is, vallen beide maten vrijwel samen en kan men vaak volstaan worden met het meten van slechts één van de twee.

De grootte van de correlatie kan geïnterpreteerd worden door de correlatie te kwadrateren en te vermenigvuldigen met 100. Dit geeft het percentage verklaarde variantie. Wanneer twee variabelen A en B 0.80 correleren, is het kwadraat daarvan 0.64 en dat levert 64%. De beide variabelen hebben dan 64% variantie gemeenschappelijk en 36% niet. Dit wil zeggen dat wanneer we A gebruiken om via lineaire regressie B te voorspellen, de voorspelde B een variantie heeft die 64% is van de oorspronkelijke variantie van B. Het stukje van B dat niet voorspeld is door A (het residu), heeft nog maar een variantie van 36% van de oorspronkelijke variantie van B. In feite hebben we daarmee de totale variantie van B opgesplitst in twee componenten: een component die volledig door A verklaard wordt en een restcomponent die volledig niet door A verklaard wordt. De eerste component correleert 1 met A, de tweede component correleert 0 met A.

Wanneer twee variabelen 0.30 correleren, betekent dit dat ze slechts 9% variantie gemeenschappelijk hebben. Op basis van de ene variabele kan men de andere variabele slechts voor een klein deel voorspellen. De onderlinge overlap is slechts 9%. Bij een correlatie van 0.90 hebben beide variabelen 81% variantie gemeenschappelijk. Ze overlappen bijna volledig. Op basis van de ene variabele kan de andere variabele bijna volledig voorspeld worden.

Correlatie als t-test

In de psychometrie is het gebruikelijk ook correlaties te berekenen wanneer één van beide variabelen dichotoom is (slechts twee waarden heeft) of wanneer beide variabelen dichotoom zijn. Een test kan bijvoorbeeld gemaakt zijn door jongens en door meisjes. Wanneer sekse gecodeerd is als 1=man en 2=vrouw kan vervolgens de correlatie met de testscore berekend worden. De precieze codering van de dichotome variabele (0/1, 1/2 of 3/5) maakt hierbij voor de hoogte van de gevonden correlatie niet uit. Iedere codering levert (afgezien van het teken) dezelfde correlatie. De significantie van de gevonden correlatie (de p-waarde) is dezelfde waarde die gevonden wordt via een t-test voor onafhankelijke steekproeven.

Het voordeel van een correlatie boven een t-test is dat de sterkte van het gevonden verband onmiddellijk zichtbaar is via de waarde van de correlatie. Verder kan men het verband eenvoudiger vergelijken met de andere verbanden waarvoor correlaties berekend waren.

Gemiddelde correlatie

Bij een test die bestaat uit een groot aantal items, is het mogelijk om via de computer alle correlaties tussen de items te berekenen en vervolgens deze correlaties samen te vatten in een enkel getal: de gemiddelde onderlinge correlatie. (De correlaties op de diagonaal van de correlatiematrix, met de waarde 1, worden hierbij buiten beschouwing gelaten.) Stel, men heeft tien items. Dat levert (10x9)/2=45 verschillende correlaties. Dat is te veel om eenvoudig te overzien. Daarom worden deze correlaties samengevat via het gemiddelde. De gemiddelde onderlinge correlatie van de tien items geeft aan of ze veel of weinig gemeenschappelijk hebben. Wanneer alle items onderling hoog correleren wordt wel gesteld dat de items of de variabelen dezelfde (onderliggende) factor meten.

Standaardiseren

Wanneer twee beoordelaars 0.90 correleren, betekent dat nog niet dat ze het echt eens zijn over de beoordeelde teksten. Beoordelaars verschillen namelijk op gemiddelde en op standaarddeviatie (SD). Het is dus mogelijk dat veel studenten bij de ene beoordelaar een voldoende hebben en bij de andere een onvoldoende. Een hoge correlatie betekent wel dat de beoordelaars het in hoge mate eens zijn over de rangordening van de teksten.

Om het probleem van de uiteenlopende oordelen op te lossen, is er een eenvoudige statistische procedure, die echter bij docenten vaak onbekend is. Voor beide beoordelaars wordt het gemiddelde en de standaarddeviatie van de beoordelingen bepaald. Vervolgens wordt van de beoordelingen het gemiddelde van de desbetreffende docent afgetrokken en daarna wordt de rest (de afwijkingsscore) gedeeld door de standaarddeviatie van de docent. Voorbeeld: het gemiddelde van een beoordelaar is 5, de SD 2. Een beoordeling van 7 wordt dan gestandaardiseerd als: (7-5)/2=1. Dit wordt gedaan voor alle beoordelingen van beoordelaar A. Daarna wordt hetzelfde gedaan voor beoordelaar B. De beoordelingen van beide beoordelaars liggen nu vlak bij elkaar. Deze procedure heet standaardiseren op gemiddelde 0 en SD 1.

Het is echter ook mogelijk te standaardiseren op een ander gemiddelde en een andere SD. Zo levert standaardiseren op gemiddelde 6 en SD 1 ongeveer uitkomsten op als bij een schoolcijfer.

Betrouwbaarheid en maximale correlatie

Na correlatie is betrouwbaarheid vermoedelijk het meest centrale begrip in de psychometrie. Wanneer bij mensen tweemaal dezelfde eigenschap wordt gemeten, zullen de uitkomsten vaak sterk verschillen. De gemeten variabele is onbetrouwbaar. De maat voor de betrouwbaarheid is de correlatie tussen twee afnames van soortgelijke testen op dezelfde groep personen.

Door die twee benodigde afnames is betrouwbaarheid in de praktijk lastig vast te stellen: men zou de onderzochte personen tweemaal moeten vragen om medewerking. Psychometrici werken mede daarom bij voorkeur met multi-item tests. Men kan vragen of iemand vaak zenuwachtig is en dat antwoord noteren. Wanneer men echter even later ook nog vraagt of men wel eens nerveus is en ook dat antwoord noteert, leveren die twee antwoorden samen een betrouwbaardere score op dan de afzonderlijke vragen. De betrouwbaarheid van een test die uit soortgelijke items bestaat, neemt toe met het aantal items, mits die items allemaal dezelfde factor meten. De verklaring hiervoor is dat de ruis uitmiddelt, terwijl het signaal sterker wordt.

De consequentie van onbetrouwbaarheid is dat een variabele nooit maximaal, dat wil zeggen 1, kan correleren met een andere variabele. Wanneer variabele A gemeten is met een betrouwbaarheid van 0.50 en variabele B perfect betrouwbaar is, is de maximaal mogelijke correlatie tussen A en B in beginsel gelijk aan de wortel uit de betrouwbaarheid. De correlatie tussen A en B kan in dit geval in beginsel maximaal 0.71 worden.

Spearman-Brown formule voor testverlenging

De betrouwbaarheid van een test van 20 items die onderling gemiddeld 0.25 correleren, kan voorspeld worden via de Spearman-Brown formule voor testverlenging. Deze formule luidt als volgt (Nunnally, 1967, p. 193, formule 6-18):

$$R = k.\ r / (1 + (k-1)\ .\ r)$$

waarin:

R= betrouwbaarheid na testverlenging

k= testverlengingsfactor

r = betrouwbaarheid voor testverlenging

Voorbeeld: de gemiddelde correlatie tussen de items bedraagt 0.25. De betrouwbaarheid van een één item test is dan 0.25. Voor een test bestaande uit twintig van deze items, wordt de betrouwbaarheid:

$$R= 20 \times 0.25 / (1 + (19 \times 0.25) = 5 / (1 + 4.75) = 5 / 5.75 = 0.87.$$

Het is ook mogelijk via de formule na te gaan, wat er gebeurt als een test korter wordt gemaakt, bijvoorbeeld half zo lang wordt genomen. Verder kan via de formule ook de betrouwbaarheid van een aantal beoordelaars voorspeld worden. Wanneer beoordelaars gemiddeld 0.30 correleren, zal de beoordelaarsbetrouwbaarheid van 5 beoordelaars $5x0.30/(1 + 4x0.30) = 0.68$ worden.

Coëfficiënt alfa, alfa standardized, raters alfa

Voor de betrouwbaarheid wordt doorgaans coëfficiënt alfa berekend. Coëfficiënt alfa (ook aangeduid als: Cronbach's alfa) gaat uit van de gemiddelde onderlinge correlatie (of bij niet gestandaardiseerde items van de covariantie) tussen de items of beoordelaars. Op deze waarde wordt vervolgens de formule voor testverlenging toepast. Coëfficiënt alfa is daar-

mee gebaseerd op twee parameters: de gemiddelde onderlinge correlatie en het aantal items van de test. Wanneer deze twee bekend zijn, valt coëfficiënt alfa daaruit te berekenen.

Voor de berekening van de betrouwbaarheid wordt normaal niet uitgegaan van de correlaties tussen de items, maar van de gemiddelde onderlinge covariantie. Items die een iets grotere variantie hebben, kunnen daardoor het totaal meer beïnvloeden dan items die vrij constant zijn. Bij het combineren van beoordelaars is dit niet altijd wenselijk. Een enkele beoordelaar die heel erg fluctueert zou daardoor relatief veel gewicht krijgen. De bedoeling is doorgaans dat alle beoordelaars ongeveer evenveel invloed hebben. Om dat te bereiken zou men de beoordelingen eerst kunnen standaardiseren. Alfa standardized geeft de betrouwbaarheid van deze gestandaardiseerde beoordelingen zonder dat men de desbetreffende variabelen al heeft gestandaardiseerd.

Normaal wordt coëfficiënt alfa berekend over de items van een toets. Het is echter ook mogelijk coëfficiënt alfa te berekenen over beoordelaars die hetzelfde 'topic' beoordeeld hebben. Hierbij moeten in SPSS de beoordelaars de variabelen (de kolommen) vormen en vormen de teksten van de studenten de regels (de 'cases'). Dit levert de 'raters alfa' of de beoordelaarsbetrouwbaarheid.

Correctie voor onbetrouwbaarheid

In de psychometrie draait het om de vraag hoe hoog twee variabelen met elkaar correleren. Stel dat men twee objectieve tests A en B gebruikt om de schrijfvaardigheid te meten en dat ze onderling 0.70 correleren. Betekent dit dat de twee tests dezelfde vaardigheid meten of meten ze allebei een net iets andere vaardigheid? Uitgaande van de correlatie overlappen ze elkaar voor 49%, er blijft dan voor iedere test nog 51% niet-verklaarde variantie over. Beide tests lijken schrijfvaardigheid te meten, maar daarnaast lijkt iedere test ook nog een eigen factor te meten. Dit zou betekenen dat we met deze twee tests drie verschillende soorten schrijfvaardigheid aangetoond hebben: de gemeenschappelijke component van A en B, de unieke component van A en de unieke component van B.

Er is echter geen rekening gehouden met de onbetrouwbaarheid van de metingen. Coëfficiënt alfa blijkt voor beide tests 0.80 te bedragen. Door die onbetrouwbaarheid zit er in beide maten een behoorlijke toevalscomponent (ruis). Dat beperkt de maximale correlatie. Via de correctie voor onbetrouwbaarheid kan men voor die onbetrouwbaarheid corrigeren.

De formule om de gevonden correlatie te corrigeren voor de betrouwbaarheid, luidt (Nunnally, 1967, p. 204, formule 6-36):

$$R = r \; / \; \sqrt{(Bx \; . \; By)}$$

waarin:

R = gecorrigeerde correlatie tussen X en Y

r = geobserveerde correlatie tussen X en Y

Bx = betrouwbaarheid X

By = betrouwbaarheid Y

$\sqrt{}$ = vierkantswortel

Voorbeeld: de correlatie was 0.70. De betrouwbaarheden zijn ieder 0.80. Dit levert:

$R = 0.70 \; / \; \sqrt{(0.80 \times 0.80)}$

$R = 0.70 \; / \; 0.80$

$R = 0.88$

Na correctie voor onbetrouwbaarheid is de correlatie 0.88. Beide tests zijn, indien ze volledig betrouwbaar zouden zijn, vrijwel perfect gecorreleerd. Het deel dat er overblijft is zo minimaal dat het te verwaarlozen valt. Dit betekent dat de 'lage' correlatie tussen de twee schrijfvaardigheidsmaten verklaard kan worden uit hun onbetrouwbaarheid.

Doordat de correlaties en betrouwbaarheden soms geschat worden via vrij kleine steekproeven en de gecorrigeerde correlatie de verhouding is van twee schattingen, kan de gecorrigeerde correlatie soms boven 1.00 uitkomen. Dat is een schattingsfout veroorzaakt door een kleine steekproef. In het algemeen kan het betrouwbaarheidsinterval bij kleine steekproeven voor gecorrigeerde correlaties groot zijn.

Bijlage 2 **Instructie beoordelaars foutenonderzoek**

Beoordeel de teksten in de volgorde waarin ze in de map zitten.

Beoordeel niet meer dan twee teksten achter elkaar om het missen van fouten door vermoeidheid te voorkomen.

Als dezelfde fout vaker voorkomt, deze steeds opnieuw onderstrepen en nummeren.

Voor iedere beoordelaar wordt achteraf het percentage gevonden fouten berekend van fouten die ook door andere beoordelaars gevonden zijn.

0. Vermeld op het bijgaande schrijfpapier het nummer van de tekst die u beoordeelt.

1. Geef in de tekst alle fouten aan door die te onderstrepen. Vermeld hierbij ook een nummer (bij iedere volgende fout een volgend nummer gebruiken, ook al gaat het om dezelfde soort fout).

Voorbeeld beoordeelde tekst

Multinationals ***06***

Multinationals vindt[1] je over de hele wereld. Shell is een bekend voorbeeld van een onderneming dat[2] een multinational is. Wie denkt dat zulke ondernemingen geen moeite moeten doen om winst te veroveren,[3] heeft niet goed nagedachd.[4]

---

2. Zet vervolgens op het bijgaande papier het nummer van de fout en geef een korte omschrijving van de fout.

Voorbeeld ingevuld schrijfpapier

---

tekst **06**

1. d/t-fout
2. die/dat
3. verkeerd woord
4. d/t
5. ...

Bijlage 3* **Bevestigde fouten in de originele tekst met commentaar beoordelaars (in de dertig teksten)**

Te raadplegen via: **http://goo.gl/6o4Rw**

* Bijlage 3, 5 en 6 zijn vanwege hun omvang niet opgenomen in dit proefschrift, maar staan online.

Bijlage 4 **Overzicht van de 76 achteraf als niet-kloppend beoordeelde bevestigde fouten (in de 30 teksten)**

1. light producten
2. Light-producten
3. light producten
4. light producten
5. light producten
6. light producten
7. light producten
8. waardering van de smaak
9. light frisdranken
10. light producten
11. light producten
12. light product
13. light frisdranken
14. light product
15. light producten
16. light frisdranken
17. light producten
18. light producten
19. light producten
20. light producten
21. light producten
22. term ... zaait ... verwarring
23. light producten
24. waardering van
25. light frisdranken
26. light producten
27. light producten
28. term verwarring zaait
29. hun
30. van jongs af aan
31. allang
32. hebben, dat
33. onderscheiden.
34. Ten slotte
35. prijs/kwal. verh.
36. prijs/kwal. verh.

37. om te beginnen
38. als 2e
39. _ het midden
40. zullen
41. Ten slotte
42. light producten
43. light producten
44. light producten
45. van
46. Light producten
47. dit product
48. light producten
49. light producten
50. light producten
51. light drankjes
52. light drankjes
53. Light producten
54. Light producten
55. light producten
56. Light producten
57. niet-light producten
58. light frisdranken
59. light producten
60. light product
61. light frisdranken
62. light frisdranken
63. light producten
64. light producten
65. light producten
66. light product
67. geeft al snel een beeld
68. het niet light product
69. light producten
70. light frisdranken
71. light producten
72. light producten
73. light producten
74. light product
75. ... druivensuiker, die
76. light producten

Bijlage 5 **Papieren methodes**

Te raadplegen via: **http://goo.gl/6o4Rw**

Bijlage 6 **Digitale methodes**

Te raadplegen via: **http://goo.gl/6o4Rw**

Bijlage 7 **Opbouw TAVAN-lessen**

Les 1

1e uur	100 zinnen
	1 fout per zin
	fouten onderstreept
	korte zinnen
2e uur	stukje tekst laten herschrijven
	omvang 125 woorden; aantal fouten 15
	via klassikale bespreking van de tekst de eigen herschreven tekst verder laten verbeteren
	deze laatste verbeterde versie laten mailen

Les 2

1e uur	100 zinnen
	1 fout per zin (niet meer onderstreept)
	vrij eenvoudige zinnen
	hetzelfde genre fouten soms bij elkaar gezet
2e uur	stukje tekst laten herschrijven
	omvang 125 woorden; aantal fouten 20
	via klassikale bespreking verder laten verbeteren
	deze laatste verbeterde versie laten mailen

Les 3

1e uur	100 zinnen
	1 fout per zin
	naast eenvoudige zinnen ook wat complexere zinnen
	hetzelfde genre fouten zelden bij elkaar gezet
2e uur	stukje tekst laten herschrijven
	omvang 125 woorden; aantal fouten 25
	via klassikale bespreking verder laten verbeteren
	deze laatste verbeterde versie laten mailen

Les 4

1e uur	100 zinnen
	1 fout per zin
	doorgaans wat complexere zinnen
	doorgaans iets lastiger fouten

2e uur stukje tekst laten herschrijven
omvang 125 woorden; aantal fouten 30
via klassikale bespreking verder laten verbeteren
deze laatste verbeterde versie laten mailen

Les 5

1e uur 100 zinnen
1 fout per zin
complexe taak door samenstelling zin en soort fout
2e uur stukje tekst laten herschrijven
omvang 125 woorden; aantal fouten 35
via klassikale bespreking verder laten verbeteren
deze laatste verbeterde versie laten mailen

Les 6

1e uur 100 zinnen
1 fout per zin
complexe taak door samenstelling zin en soort fout
2e uur tekst laten herschrijven
omvang 150 woorden; aantal fouten 50
via klassikale bespreking verder laten verbeteren
deze laatste verbeterde versie laten mailen

Les 7

1e uur 100 zinnen
1 fout per zin
complexe taak door samenstelling zin en soort fout
2e uur tekst laten herschrijven
omvang 200 woorden; aantal fouten 65
geen klassikale bespreking meer van verbeterpunten
herschreven versie laten mailen

Les 8

1e uur 100 zinnen
1 fout per zin
complexe taak door samenstelling zin en soort fout
2e uur tekst laten herschrijven
omvang 250 woorden; aantal fouten 74
herschreven versie laten mailen

Les 9

1e uur	100 zinnen
	2 fouten per zin
	complexe taak door verbetering zin op meer dan een manier
	in enigszins complexe zinnen
2e uur	tekst laten herschrijven
	omvang 250 woorden; aantal fouten 82
	herschreven versie laten mailen

Les 10

1e uur	100 zinnen
	2 fouten per zin
	complexe taak door verbetering zin op meer dan een manier in complexe zinnen
2e uur	tekst laten herschrijven
	omvang 300 woorden; aantal fouten 100
	herschreven versie laten mailen

Bijlage 8 **TAVAN: verbeteren van zinnen (10 zinnen uit les 1)**

Geef eerst antwoord en check pas daarna via het docentantwoord.

1 De term light wekt veel verwarring in onze samenleving.

Jouw antwoord: ….....

Check je antwoord via 38: ….....

2 Wat bedoeld de verkoper hiermee?

Jouw antwoord: ….....

Check je antwoord via 56: ….....

3 Consumenten zijn niet op de hoogte van de inhout van de verpakking.

Jouw antwoord: ….....

Check je antwoord via 14: ….....

4 Jarenlang gebruikte hij te veel slaapmiddelen, maar wil hij nu daarmee stoppen.

Jouw antwoord: ….....

Check je antwoord via 72: ….....

5 Dit product bevat te veel caloriën.

Jouw antwoord: ….....

Check je antwoord via 99: ….....

6 Supermarkten krijgen veel van deze producten in de winkels om te verkopen, maar als er een slecht beeld van bestaat, verkopen ze niets.

Jouw antwoord: ….....

Check je antwoord via 65: ….....

7 Geldhandelaren denken wel gemakkelijk voor dit probleem.

Jouw antwoord: ….....

Check via 43: ….....

8 Albert Heijn filialen hebben veel van deze producten, maar als consumenten er een slecht beeld bij hebben, verkopen ze niets.

Jouw antwoord: ….....

Check 21: ….....

9 De belangstelling voor de product zal teruglopen.

Antwoord: ….....

Check 04: ….....

10 En daardoor krijgen de ondernemers er ook genoeg van.

Antwoord: ….....

Check 93: ….....

Docentantwoorden TAVAN (door elkaar)

38 term 'light'
56 bedoelt
14 inhoud
72 hij wil
99 calorieën.
65 [weg]
43 over
21 van
4 het product
93 Daardoor

Bijlage 9 **TAVAN: herschijfopdracht**

H&M verkoopd er veel mode als groente Manshooge borden van de Historisch Museum ontrekken het nieuwe winkel van H&M Hennes & Mauritz aan de dam op Amsterdam nu nog aan het zigt.

In de monumentale pand heropend zweedse-modegigant dit najaar het grootste filiaal in Nederland. Retailkenner Pim van den Berg zeggen; een locatie waar je als ondernemer alleen in kunt dromen.

Het **mode bedrijf** H&M, die er is bezig in 37 landen, aasde al langer achter het gebouw, wat eerder toebehoorde aan ABN Amro. Kiezen voor top-locaties met veel vloer oppervlakte beproefde strategie is vanH&M.

Op die plaatsen je trekt nu eenmaal veel publiek : aldus van den Berg. Het bedrijf H&M heeft dat ook wel nodig. Hun winstmarges zijn heel erg klein, daarom dit bedrijf moet het hebben in volumes en omzetsnelheid. Winkel aan de Dam word er de vierde ter wereld waar H&M behalfe kleding ook interieurtextiel gaat verkopen: theedoeken beddengoed en kussentjes. Grote winkelketens als Zara en Sissy Boy ging de concern voor.

waar andere merken er twee keer per jaar nieuwe collectie presenteren, doet H&M dat aan de lopende markt.

De H&M winkels verkoopt mode zo als groente, zegt Van den Berg.

“Net als in versspeciaalzaak is er bijna elke dag een ander aan bod.”

Bijlage 10 **Schrijfopdracht tv (taalvaardigheid)**

Voor eerstejaars IBL-studenten is deze opdracht een voorwaarde voor het behalen van een voldoende resultaat op taalvaardigheid in de propedeuse.

De volgende tekst is niet altijd even duidelijk en correct geformuleerd. De opdracht is om deze tekst te verbeteren. Je hebt twee uur de tijd, dan moet je tekst af zijn en per mail verstuurd worden.

De tekst is via internet beschikbaar op:

http://dl.dropbox.com/u/6861883/tekstA128.doc

- Typ de bovenstaande link in de adresbalk van je browser.
- Open de file in Word.
- Sla de tekst daarna eerst op in je eigen directory als: opdracht tv
- Hierna kun je de tekst verbeteren.

* Vermeld in het Word-doc eerst je naam en studentnummer.
* Schrijf correct Nederlands.
* Zorg ervoor dat je tekst prettig leest.
* Zorg ervoor dat je tekst begrijpelijk is.
* Zorg ervoor dat de lengte ongeveer gelijk blijft (300 woorden).
* Sla regelmatig op.

Stuur het word-document als bijlage aan:

a.van.eerden@pl.hanze.nl

- Check of je naam en studentnummer vermeld zijn.
- Check of je mailtje als bijlage bevat: opdracht tv

Schrijfopdracht tv (taalvaardigheid)

Voor eerstejaars IBL-studenten is deze opdracht een voorwaarde voor het behalen van een voldoende resultaat op taalvaardigheid in de propedeuse.

De volgende tekst is niet altijd even duidelijk en correct geformuleerd. De opdracht is om deze tekst te verbeteren. Je hebt twee uur de tijd, dan moet je tekst af zijn en per mail verstuurd worden.

De tekst is via internet beschikbaar op:

http://dl.dropbox.com/u/6861883/tekstB37.doc

- Typ de bovenstaande link in de adresbalk van je browser.
- Open de file in Word.
- Sla de tekst daarna eerst op in je eigen directory als: opdracht tv
- Hierna kun je de tekst verbeteren.

* Vermeld in het Word-doc eerst je naam en studentnummer.
* Schrijf correct Nederlands.
* Zorg ervoor dat je tekst prettig leest.
* Zorg ervoor dat je tekst begrijpelijk is.
* Zorg ervoor dat de lengte ongeveer gelijk blijft (300 woorden).
* Sla regelmatig op.

Stuur het word-document als bijlage aan:

a.van.eerden@pl.hanze.nl

- Check of je naam en studentnummer vermeld zijn.
- Check of je mailtje als bijlage bevat: opdracht tv

Bijlage 11 **Tekst Vragenlijst taalvaardigheid**

Vragenlijst Taalvaardigheid

Vul eerst je naam en studentnummer in.

Naam:

Studentnummer:

Omcirkel het gewenste antwoord.

1	Het maken van een verslag lukt me altijd wel.	nee / ? / ja
2	Ik ken veel moeilijke woorden.	nee / ? / ja
3	Schrijven kost nu eenmaal veel tijd.	nee / ? / ja
4	Een woord dat ik niet ken, zoek ik op.	nee / ? / ja
5	Het schrijven van een verslag moet snel gaan.	nee / ? / ja
6	Correct schrijven is voor mij belangrijk.	nee / ? / ja
7	Wat ik schrijf, moet goed zijn.	nee / ? / ja
8	Kranten vormen voor mij de belangrijkste nieuwsbron.	nee / ? / ja
9	Ik vind schrijven wel leuk.	nee / ? / ja
10	Schrijven kost altijd te veel tijd.	nee / ? / ja
11	Mijn mails zijn meestal nogal kort.	nee / ? / ja
12	Ik heb een grote woordenschat.	nee / ? / ja
13	Ik ben bereid veel tijd in het schrijven van een verslag te steken.	nee / ? / ja
14	Ik weet dat ik goed kan schrijven.	nee / ? / ja
15	Ik ben vrij goed in spelling.	nee / ? / ja
16	De computer is voor mij de belangrijkste nieuwsbron.	nee / ? / ja
17	Ik ben niet zo goed in schrijven.	nee / ? / ja
18	Ik weet dat ik nogal wat spelfouten maak.	nee / ? / ja
19	Gratis kranten lees ik altijd als ik ze tegenkom.	nee / ? / ja
20	Het schrijven van een mailtje gaat me gemakkelijk af.	nee / ? / ja
21	Als wat ik schrijf voldoende is, ben ik tevreden.	nee / ? / ja
22	Spelling vind ik eigenlijk niet zo belangrijk.	nee / ? / ja
23	Ik heb wel eens een dagboek bijgehouden.	nee / ? / ja
24	Schrijven gaat bij mij zo snel mogelijk.	nee / ? / ja
25	Ik schrijf vaak fouten die ik zelf niet zie.	nee / ? / ja
26	De televisie is voor mij de belangrijkste nieuwsbron.	nee / ? / ja
27	Ik kom vaak woorden tegen die ik niet ken.	nee / ? / ja
28	Als ik niet oppas, schrijf ik vaak meer dan mag.	nee / ? / ja

29 Ik weet dat ik soms moeite heb met schrijven. nee / ? / ja

30 De spellingchecker haalt de spelfouten er wel uit. nee / ? / ja

31 Hoeveel lees je? weinig / normaal / veel

32 Hoe vaak mail je?

nooit / iedere week / iedere dag / meerdere keren per dag

33 Hoe vaak heb je moeite met schrijfopdrachten?

nooit / soms / regelmatig / vaak / altijd

34 Hoe schrijf je in vergelijking met anderen? slechter / even goed / beter

35 Hoe vaak maak je uittreksels van te bestuderen stof?

nooit / soms / regelmatig/ vaak / altijd

36 Geef een schoolcijfer (1-10) voor je eigen schrijfvaardigheid: ….....

37 De klas krijgt een dictee. Welk cijfer haal je, denk je? …..….....

38 Hoe vaak kijk je per week naar actualiteitenprogramma's? ….....

39 Hoe vaak kijk je per week naar het journaal? …..….....

40 Hoeveel uur tv kijk je per dag? ...….....

41 Hoe vaak per week bezoek je een nieuwssite? …..….....

42 Hoeveel minuten per dag breng je op nieuwssites door? ….....…..

43 Hoeveel minuten lees je per dag een betaalde krant? ….....….

44 Hoeveel minuten lees je per dag een gratis krant? ….....

45 Hoe vaak lees je per week een betaalde krant? ….....

46 Hoe vaak per week lees je een gratis krant? ….....

Bijlage 12 **Schrijfopdracht 1: Evaluatie van TAVAN**

1. Inleidende alinea(s).	Wat is je onderwerp? Wat is het doel van TAVAN?
2. Beschrijving van TAVAN.	Hoe ziet het programma eruit? Hoe werkt het?
3. Beoordeling van TAVAN.	Wat zijn positieve aspecten met het oog op het doel? Wat zijn negatieve aspecten met het oog op het doel?
4. Conclusie.	In hoeverre wordt het doel van TAVAN bereikt? Wat is je eindoordeel over TAVAN?

=======

- Formuleer helder en foutloos. Je tekst is bedoeld voor publicatie in een studentenblad voor aankomende eerstejaars.
- Zorg voor een prettige lay-out (titel, tussenkoppen, witregels tussen de tekstblokken).

- Omvang van je tekst**: maximaal 1 A4 (minimaal 400 woorden)**

- Zet je naam en studentnummer op de tekst die je naar je docent stuurt.
  Bij de beoordeling van je tekst door studenten, zullen die niet kunnen zien door wie de tekst geschreven is.

*** mail je tekst als bijlage naar: a.van.eerden@pl.hanze.nl**

Bijlage 13 **Beoordelingsformulier teksten studenten**

Voornaam: ..... Studentnummer: .....

Achternaam: ..... Klas: .....

1) Lees de 6 teksten en beantwoord per tekst de volgende punten.

Codeletters tekst: .....

Sterke punten: .....

.....

Zwakke punten: .....

.....

Codeletters tekst: .....

Sterke punten: .....

.....

Zwakke punten: .....

.....

Codeletters tekst: .....

Sterke punten: .....

.....

Zwakke punten: .....

Ga op de achterzijde verder met de beoordeling.

Codeletters tekst: .....

Sterke punten: .....

.....

Zwakke punten: .....

.....

Codeletters tekst: .....

Sterke punten: .....

.....

Zwakke punten: .....

.....

Codeletters tekst: .....

Sterke punten: .....

.....

Zwakke punten: .....

---

2) Orden nu de teksten van slecht (1) naar goed (6), met behulp van de codeletters van de tekst.

Slechtste tekst *beste tekst*

1	2	3	4	5	6
Codeletters	Codel.	Codel.	Codel.	Codel.	Codel.
.....	.....	.....	.....	.....	.....

Bijlage 14 **Negen tekstversies op basis van drie teksten**

BE- 11/04/2013

Evaluatie Tavan

Onderwerp: Tavan

Wat is het doel van tavan?

Tavan is bedoeld om het taal/spelling niveau van jezelf omhoog te krikken.

Hoe ziet het progamma eruit/hoe werkt het ?

Bij tavan ga je via de computer verschillende opgaven van 50 voorbeeld zinnen maken, waar fouten in staan om deze vervolgens te verbeteren je moet 3 lesjes per keer maken A,B,C. er word aangegeven hoeveel fouten erin een zin staan,meestal was het zo A 2 foutjes per zin ,B 1 fout, C 2 fout . Na deze opgave gemaakt te hebben komt er een percentage uit van 0% tot 100% dat is dan je score. Voor een voldoende (5.5) moet je minimaal 75% goed hebben van de 50 items per keer.

Wat zijn de positieve aspecten met het oog op het doel ?

Na veel oefenen leer je spel/taal fouten makkelijker te herkennen . zelf ga sta je hier ook dan vaker bij stil. Het is goed om te zien dat je je zelf iedere keer weer een stapje verbeterd. Hierdoor kunnen je bespaar je tijd met het nachecken van je fouten.

Wat zijn de negatieve aspecten met het oog op het doel?

Het programma Tavan dat via de pc gemaakt moest worden liep sommige keren vast dat was wel irritant of dat de sever niet werkte dit waren voor mij de enige negatieve aspecten van Tavan,

In hoeverre wordt het doel bereikt van Tavan?

Het doel om beter te formulieren/spellen/taal Word zeker bereikt, je hoort veel mensen over dat ze zicht steeds verbeteren in deze onderdelen. Dus in dat opzicht word het doel van Tavan behaald.

Wat is je eindoordeel over tavan?

Tavan is een goed leermiddel om spelling/taal presentaties te verbeteren. Het enige wat ik een nadeel vond is dat je lang geconcentreerd achter een pc scherm zat te kijken, dit maakte je aan het einde van de oefeningen wel moe.

Evaluatie Tavan

Evaluatie Tavan

Wat is het doel van Tavan?

Tavan is bedoeld om het niveau van taal en spelling te verhogen.

Hoe ziet het progamma eruit?

Bij Tavan ga je via de computer verschillende opgaven van 50 voorbeeldzinnen maken, waar fouten in staan om deze vervolgens te verbeteren. Je moet 3 lesjes per keer maken A, B en C. Er wordt aangegeven hoeveel fouten er in een zin staan. Meestal was het zo: A 2 foutjes per zin, B 1 fout, C 2 fout . Na deze opgaven gemaakt te hebben, komt er een percentage uit van 0% tot 100%. Dat is dan je score. Voor een voldoende (5.5) moet je minimaal 75% goed hebben van de 50 items per keer.

Wat zijn de positieve aspecten met het oog op het doel ?

Na veel oefenen leer je spel- en taalfouten gemakkelijker te herkennen. Zelf sta je hier dan ook vaker bij stil. Het is goed om te zien dat je je zelf iedere keer weer een stapje verbetert. Hierdoor bespaar je tijd met het checken van je fouten.

Wat zijn de negatieve aspecten met het oog op het doel?

Het programma Tavan dat op de computer gemaakt moest worden, liep soms vast. Dat was wel irritant of de server werkte niet. Dit waren voor mij de enige negatieve aspecten van Tavan.

In hoeverre wordt het doel bereikt van Tavan?

Het doel om beter te leren formulieren en te spellen wordt zeker bereikt. Je hoort van veel mensen dat ze zich steeds verbeteren in deze onderdelen, dus in dat opzicht wordt het doel van Tavan behaald.

Wat is je eindoordeel over Tavan?

Tavan is een goed leermiddel om je spelling en taal te verbeteren. Het enige wat ik een nadeel vond, is dat je lang geconcentreerd naar een computerscherm zat te kijken. Dit maakte je aan het einde van de oefeningen wel moe.

BE- 13/04/2013

Leren schrijven via de computer

Het studieonderdeel Tavan heeft tot doel de schrijfvaardigheid van studenten te verbeteren.

De opzet van Tavan bestaat eruit dat de student via de computer zinnen waar fouten in zitten, verbetert. Per keer moet je 3 lesjes maken van 50 zinnen: A, B en C. Er wordt aangegeven hoeveel fouten in een zin staan, meestal had A twee fouten per zin, B één fout en C twee fouten.

Na het lesje gemaakt te hebben, levert een percentage van 0 tot 100 procent dan de score. Voor een voldoende (5.5) moet je minimaal 75 procent goed hebben van de 50 items van het lesje.

Positief aan Tavan is dat je door het vele oefenen fouten gemakkelijker gaat herkennen. Zelf ga je hier dan ook beter op letten. Het is goed om te zien dat je iedere keer weer een stapje verbetert. Hierdoor kun je tijd besparen met het nalopen op fouten.

Een minpunt van Tavan is dat het programma sommige keren vastliep, omdat de server niet werkte. Dat was wel irritant. Dit waren voor mij de enige negatieve aspecten van Tavan.

Mijn eindoordeel over Tavan is dat het doel van een betere schrijfvaardigheid zeker wordt bereikt. Je hoort veel studenten erover dat ze steeds beter worden op dit punt. Tavan is een goed middel om de schrijfvaardigheid te verbeteren.

Het enige wat ik een nadeel vond, is dat je lang en geconcentreerd achter een computerscherm zat te kijken. Dat maakte je aan het einde van de drie lesjes wel moe.

Het onderwerp van TAVAN is: taalvaardigheid in het Nederlands.
Het doel van deze les is het verminderen van het aantal fouten in geschreven teksten.

Hoe ziet het programma eruit? Hoe werkt het?

TAVAN werkt als een computerprogramma. Iedere studenten werkt indvidueel vanaf zijn eigenen computer. TAVAN wordt per lesuur in 3 lessen weer gegeven A, B en C. Iedere les bevatt 50 zinnen met telkens één of twéé fouten, dus in totaal 150 zinnen ter corrigeren. Je heb daarvoor 2 uur tijd, maar iedereen werkt in zijn eigen tempo. Dus als je het niet op tijd lukt, moet je de rest thius maken voordat de nieuwe les begint. Van A naar C wordt het steeds moeilijker de fouten te herkennen. Als je bedoelt een fout te zien, verbeter je deze en als je alles juist gemaakt hebt wordt “ANTWOORD JUIST” getoont en je ga naar de volgende zin. Als je fouten gemist hebt toont het programma de volledige juiste zin, zodat je de volgende keer de fout herkent.
Omdat de computerlessen vermoeiend zijn, is het verstandig tussen de computerlessen (A, B of C) steeds even (maximaal 10 minuten) te pauzeren. Naar iedere les toont het programma de resultaat van de les, bijvoorbeeld je heb 25 zinnen goed gecontorleerd van de 50 zinnen, dan heb je 50% juist gemaakt.
De resultaten van het doorwerken van de computerlessen worden wekelijks gepubliceed op Blackboard. De computerlessen bestaan uit uit hetzelfde soort items als de toetsen. Wie goed wil scoren op de toets, moet de computerlessen serieus doorwerken. Maar de score in de computerlessen is geen toetscijfer! Het cijfer wordt op basis van de toetsresultaten later door de docent vastgesteld. De toets van iedere periode bepaalt voor 85% het cijfer van die periode. Maar alleen het cijfer voor periode 3 wordt berekend als het gewogen gemiddelde van de cijfers van periode 1 (40%) en periode 2 (60%). Bij de toets is het voor buitenlandse studenten mogelijk een online-woordenboek te gebruiken.

Positieve aspecten met het oog op het doel?

- je leer het verminderen van het aantal fouten in geschreven teksten
- het doorwerken van het programma leidt tot gemiddeld 20% minder fouten
- je leert sneller fouten te herkennen

◊ dus je leert voor jouw toekomst duidelijker en sneller teksten zonder fouten te schrijven

Negatieve aspecten met het oog op het doel?

- de computerlessen zijn vermoeiend, du je kan je naar een tijd niet met zo goed concentreren
- er is geen verdere uitleg waarom het bijvoorbeeld een ‘t’ of een ‘d’ aan het eind van een werkwoord moet zijn
- voor buitenlandse studenten is het veel moeilijker de programma te doen ofwel de toets dan voor nederlanders, maar zij worden gelijk beoordeelt

TAVAN is wel een vermoeiend programma, maar helpt je op lange termijn veel in het schrijven van teksten.

AG-15/04/2013

Evaluatie van TAVAN

Deze tekst gaat over TAVAN (Taalvaardigheid in het Nederlands). TAVAN is een online-schrijfvaardigheidsprogramma dat als doel heeft het verminderen van het aantal fouten in geschreven teksten.

Hoe ziet het programma eruit?

TAVAN werkt als een computerprogramma. Iedere student werkt individueel vanaf zijn eigen computer. Iedere TAVAN-les bestaat uit 3 lessen: A, B en C. Iedere les bevat 50 zinnen met telkens een of twee fouten, dus in totaal moeten 150 zinnen gecorrigeerd worden. Je heb daarvoor 2 uur tijd, maar iedereen werkt in zijn eigen tempo. Dus als het je niet op tijd lukt, moet je de rest thuis maken, voordat de nieuwe les begint. Van A naar C wordt het steeds moeilijker de fouten te herkennen. Als je denkt een fout te zien, verbeter je deze en als je alles juist gemaakt hebt, wordt 'ANTWOORD JUIST' getoond en ga je naar de volgende zin. Als je fouten gemist hebt, toont het programma de volledige juiste zin, zodat je de volgende keer de fout herkent.

De computerlessen zijn vermoeiend en daarom is het verstandig tussen de computerlessen (A, B of C) steeds even (maximaal 10 minuten) te pauzeren. Na iedere les toont het programma het resultaat van de les, bijvoorbeeld je hebt 25 zinnen goed gecontroleerd van de 50 zinnen, dan heb je 50% juist gemaakt.

De resultaten van het doorwerken van de computerlessen worden wekelijks gepubliceerd op Blackboard. De computerlessen bestaan uit uit hetzelfde soort items als de toetsen. Wie goed wil scoren op de toets, moet de computerlessen serieus doorwerken. Maar de score in de computerlessen is geen toetscijfer! Het cijfer wordt op basis van de toetsresultaten later door de docent vastgesteld. De toets van iedere periode bepaalt voor 85% het cijfer van die periode. Alleen het cijfer voor periode 3 wordt berekend als het gewogen gemiddelde van de cijfers van periode 1 (40%) en periode 2 (60%). Bij de toets is het voor buitenlandse studenten mogelijk een online-woordenboek te gebruiken.

Positieve aspecten met het oog op het doel

- Je leert het aantal fouten in geschreven teksten te verminderen.
- Het doorwerken van het programma leidt tot gemiddeld 20% minder fouten.
- Je leert sneller fouten te herkennen.

Je leert dus om snel duidelijke teksten te schrijven.

Negatieve aspecten met het oog op het doel

- De computerlessen zijn vermoeiend, dus je kan je na een tijd niet meer zo goed concentreren.
- Er is geen verdere uitleg waarom het bijvoorbeeld een ‘t’ of een ‘d’ aan het eind van een werkwoord moet zijn.
- Voor buitenlandse studenten is het veel moeilijker het programma te doen ofwel de toets dan voor Nederlanders, maar zij worden gelijk beoordeeld.

TAVAN is wel een vermoeiend programma, maar helpt je op lange termijn veel in het schrijven van teksten.

TAVAN: leren schrijven met minder fouten

TAVAN staat voor: Taalvaardigheid Nederlands. Het doel van dit werkcollege is studenten te leren met minder fouten te schrijven.

Wat is de opzet van dit werkcollege? Hoe gaat het?

TAVAN werkt met een computerprogramma. Iedere student werkt individueel vanaf zijn eigen computer. TAVAN bestaat per werkcollege uit 3 lessen: A, B en C. Iedere les bevat 50 zinnen met telkens één of twee fouten, in totaal dus 150 zinnen te corrigeren. Je hebt daarvoor 2 uur de tijd, maar iedereen werkt in zijn eigen tempo. Als je niet op tijd klaar bent, moet je de rest thuis maken voor het volgende werkcollege begint. Van A naar C wordt het steeds moeilijker de fouten te herkennen. Als je denkt een fout te zien, verbeter je deze. Als je alles goed hebt, wordt “ANTWOORD JUIST” getoond en ga je naar de volgende zin. Als je fouten gemist hebt, toont het programma de goed gecorrigeerde zin, zodat je de volgende keer de fout herkent.

Omdat de computerlessen vermoeiend zijn, is het verstandig tussen de lessen (A, B of C) steeds 10 minuten te pauzeren. Na iedere les toont het programma het resultaat van de les. Als je bijvoorbeeld 25 van de 50 zinnen goed gecorrigeerd hebt, scoor je 50 procent.

De resultaten van het doorwerken van de computerlessen worden wekelijks gepubliceerd op Blackboard. De computerlessen bestaan uit hetzelfde soort items als de toetsen. Wie goed wil scoren op de toets, moet de computerlessen serieus doorwerken. Maar de score in de computerlessen is geen toetscijfer! Het cijfer wordt op basis van de toetsresultaten later door de docent vastgesteld. De toets van iedere periode bepaalt voor 85% het cijfer van die periode. Alleen het cijfer voor periode 3 wordt berekend als het gewogen gemiddelde van de cijfers van periode 1 (40%) en periode 2 (60%), omdat het werkcollege in die periode niet gegeven wordt. Bij de toets is het voor buitenlandse studenten toegestaan een online-woordenboek te gebruiken.

Wat zijn de pluspunten van TAVAN? Je leert met minder fouten te schrijven. Het doorwerken van het programma leidt tot gemiddeld 20% minder fouten. Je gaat fouten ook sneller herkennen. Je leert dus duidelijker en sneller te schrijven zonder fouten.

Minpunten heeft TAVAN ook. De computerlessen zijn vermoeiend. Na een tijd kun je je niet meer zo goed concentreren. Verder wordt er geen uitleg gegeven, waarom het bijvoorbeeld een ‘t’ of een ‘d’ aan het einde van een werkwoord moet zijn. Voor buitenlandse studenten is het veel moeilijker het programma en de toets te maken dan voor Nederlanders, maar ze worden gelijk beoordeeld.

Mijn eindoordeel is dat TAVAN een vermoeiend programma is, maar op langere termijn helpt het je goed om beter te worden in het schrijven van teksten.

Evaluatie van Tavan

Het Nederlandse computerprogramma Tavan, het doel van Tavan is het ontwikkelen van de Nederlandse taal. Dit wordt gedaan op een computer met als doel het inzien van spelfouten.

Het programma ziet er als volgt uit: Als het programma wordt gestart krijg je vijftig zinnen met daarin; spelfouten, fouten in gezegden en zinsdeel fouten. Deze moeten dan worden verbeter (door de persoon die met het programma bezig is), als deze zinnen goed worden opgelost ga je door naar de volgende zin. Ook kan het gebeuren dat de zin niet goed word opgelost en krijg je het beste antwoord van de docent te zien. Bij zo'n antwoord van de docent is het de bedoeling dat je goed naar het antwoord van de docent kijkt en inziet wat je fout hebt gedaan.

De positieve aspecten van Tavan met het oog op het doel zijn voornamelijk dat je leert van je fouten. Je leert van je fouten om zoveel mogelijk te oefenen, hoe vaker je het oefent hoe meer je inziet hoe de zinnen moeten lopen en welke gezegden er hadden moeten staan.

De negatieve aspecten van Tavan met het oog op het doel zijn voornamelijk dat je niet precies inziet wat je fout doet. Met als voorbeeld: Als er een spelfout wordt gemaakt krijg je het docenten antwoord te zien en zie je wat je fout hebt gedaan, maar er word niet exact bij verteld waarom jou antwoord fout is.

In hoeverre het doel van Tavan wordt bereikt is lastig te zeggen, voor sommige mensen werkt het heel goed en voor andere mensen minder goed. Mensen die vanaf het begin met Nederlands moeten beginnen kunnen beter een ander programma nemen of leren uit een boek, om zo eerst te weten waarom de Nederlandse grammatica zo werkt.
Voor de mensen die de Nederlandse taal al gemiddeld beheersen is dit een heel fijn programma om mee te werken.

Mijn eindoordeel over Tavan
Ik ben zelf niet zo goed in Nederlands en vind het dus lastig, omdat er geen uitleg wordt gegeven hoe het wel moet, je krijgt alleen de fout te zien. Door de zinnen vaak te oefenen leer je de woorden automatisch uit je hoofd, maar je weet dan uiteindelijk nog steeds niet waarom dit zo word gedaan. Toch zijn er ook positieve punten over Tavan te zeggen met als voorbeeld; Als er Nederlandse spreekwoorden in de zin staan en je deze nog niet zo goed kent. Leer je deze voornamelijk met Tavan door veel te oefen, ook leer je door veel te oefenen, meer woorden kennis en dat is voor mij wel erg fijn.

Evaluatie van Tavan

Deze tekst gaat over het online-schrijfvaardigheidsprogramma Tavan. Het doel van Tavan is het ontwikkelen van de Nederlandse schrijf- en spelvaardigheid.

Het programma

Het programma ziet er als volgt uit: als het programma wordt gestart, krijg je vijftig zinnen met daarin spelfouten, fouten in gezegden en zinsdeelfouten. Deze moeten dan worden verbeterd (door de persoon die met het programma bezig is). Als deze zinnen goed worden opgelost, ga je door naar de volgende zin. Ook kan het gebeuren dat de zin niet goed wordt opgelost en dan krijg je het beste antwoord van de docent te zien. Bij zo'n antwoord van de docent is het de bedoeling dat je goed naar het antwoord van de docent kijkt en inziet wat je fout hebt gedaan.

Positieve en negatieve aspecten

De positieve aspecten van Tavan met het oog op het doel zijn voornamelijk dat je leert van je fouten. Je leert van je fouten door zoveel mogelijk te oefenen. Hoe vaker je het oefent, hoe meer je inziet hoe de zinnen moeten lopen en welke gezegden er hadden moeten staan.

De negatieve aspecten van Tavan met het oog op het doel zijn voornamelijk dat je niet precies inziet wat je fout doet, bijvoorbeeld: als er een spelfout wordt gemaakt krijg je het docentenantwoord te zien en zie je wat je fout hebt gedaan, maar er wordt niet exact bij verteld waarom jouw antwoord fout is.

In hoeverre het doel van Tavan wordt bereikt, is lastig te zeggen. Voor sommige mensen werkt het heel goed en voor andere mensen minder goed. Mensen die vanaf het begin met Nederlands moeten beginnen, kunnen beter een ander programma nemen of leren uit een boek, om zo eerst de Nederlandse grammatica te leren kennen. Voor de mensen die de Nederlandse taal al gemiddeld beheersen, is dit een heel fijn programma om mee te werken.

Mijn eindoordeel over Tavan

Ik ben zelf niet zo goed in Nederlands en vind Tavan dus lastig, omdat er geen uitleg wordt gegeven hoe het wel moet. Je krijgt alleen de fout te zien. Door de zinnen vaak te oefenen, leer je de woorden automatisch uit je hoofd, maar je weet dan uiteindelijk nog steeds niet waarom dit zo wordt gedaan. Toch zijn er ook positieve punten over Tavan te zeggen met als voorbeeld: als er Nederlandse spreekwoorden in de zin staan en je deze nog niet zo goed kent, leer je ze door veel te oefenen met Tavan. Ook krijg je door veel te oefenen meer woordenkennis en dat is voor mij wel erg fijn.

CG-19/04/2013

Beter Nederlands met Tavan?

Het doel van het computerprogramma Tavan is je te leren beter Nederlands te schrijven. Je werkt op de computer om je bewust te worden van taalfouten.

Het programma werkt als volgt. Na het starten krijg je vijftig zinnen met spelfouten, fouten in uitdrukkingen en zinsdeelfouten. Deze fouten moeten worden verbeterd door de student. Als de zin goed is verbeterd, ga je door naar de volgende zin. Ook kan het gebeuren dat de zin niet goed is verbeterd. Dan krijg je het antwoord van de docent te zien. Bij het antwoord van de docent is het de bedoeling dat je goed naar het antwoord kijkt en begrijpt wat je fout hebt gedaan.

Het positieve aspect van Tavan is vooral dat je leert van je fouten. Je leert door zoveel mogelijk te oefenen. Hoe vaker je oefent, hoe beter je begrijpt hoe de zinnen moeten lopen en welke uitdrukkingen er hadden moeten staan.

Het belangrijkste minpunt van Tavan is vooral dat je niet precies begrijpt, wat je fout doet. Als er bijvoorbeeld een spelfout wordt gemaakt, krijg je het docentantwoord te zien. Je ziet wat je fout hebt gedaan, maar er wordt niet precies bij verteld waarom het fout is.

In hoeverre het doel van Tavan wordt bereikt, is lastig te zeggen. Voor sommige mensen werkt het heel goed, maar voor andere mensen minder goed. Mensen die moeten starten met Nederlands, kunnen beter een ander programma nemen of een boek, om er eerst achter te komen hoe de Nederlandse grammatica werkt. Voor studenten die de Nederlandse taal al redelijk beheersen, is dit een heel fijn programma om mee te werken.

Wat is mijn eindoordeel over Tavan? Ik ben zelf niet zo goed in Nederlands en vind Tavan lastig, omdat er geen uitleg wordt gegeven over hoe het moet. Je krijgt alleen de fout te zien. Door de zinnen vaak te oefenen, leer je de woorden automatisch uit je hoofd, maar je weet dan nog steeds niet, waarom dit zo wordt gedaan. Toch heeft Tavan ook positieve punten. Als er bijvoorbeeld Nederlandse uitdrukkingen in de zin staan die je nog niet goed kent, leer je deze door met Tavan veel te oefenen. Ook leer je door te oefenen meer woorden, dat vind ik fijn.

Bijlage 15 **Beoordelingsformulier lezers**

Geef je oordeel over de tekst door op onderstaande schalen een kruisje te zetten.

Saai Leuk
|----------|

Onduidelijk Duidelijk
|----------|

Niet informatief Wel informatief
|----------|

Slordig Verzorgd
|----------|

Slecht geschreven Goed geschreven
|----------|

Vervelend Interessant
|----------|

Subjectief Objectief
|----------|

Zwak Sterk
|----------|

Ondeskundig Deskundig
|----------|

Ongeschikt voor publicatie Geschikt voor publicatie
|----------|

Bijlage 16 **Instructie lezers**

Beste student,

Bijgaande tekst is bedoeld voor een studentenblad. Lees s.v.p. de tekst door en geef daarna je mening op het beoordelingsformulier.

Bedankt voor het meedoen.

Anouk van Eerden

Bijlage 17 **Een tweede beoordeling van twee methodes**

Een tweede beoordeling van twee methodes

Mik van Es, februari 2014

Inleiding

De beoordeling van onderwijsmethodes (zie hoofdstuk 5) op grond van didactische criteria is een lastige zaak (Van Es, 1985). Allereerst is er het probleem van de juistheid en de volledigheid van de criteria waarvan wordt uitgaan. Een tweede probleem is dat een onderwijsmethode door onvoorziene factoren anders kan uitwerken dan het plan was. De waarde van het uitgebrachte oordeel valt daardoor pas achteraf te bepalen door de methode in de praktijk uit te proberen bij studenten en de leerwinst te bepalen.

Een derde probleem bij de beoordeling van onderwijsmethodes is de betrouwbaarheid van de beoordeling. Leidt een tweede beoordeling tot dezelfde conclusie als de eerste? Wanneer verschillende beoordelaars tot belangrijk verschillende conclusies komen, lijkt een beoordeling op grond van dat gegeven al weinig waarde te hebben. Een eerste check op de beoordeling van een onderwijsmethode is daarom een tweede beoordeling. Vanuit dit uitgangspunt verzocht mijn mede-auteur om twee methodes, een papieren en een digitale, te willen beoordelen. Kwam ik tot soortgelijke conclusies als zij bij de eerste beoordeling (zie hoofdstuk 5)?

Idealiter zouden alle beoordeelde methodes aan een tweede beoordeling worden onderworpen. De kosten aan tijd van een dergelijke tweede beoordeling leken in dit geval echter niet op te wegen tegen de mogelijke opbrengst. De vraag waar het immers uiteindelijk om ging, is of een van de onderzochte methodes effectief en geschikt is om de schrijfvaardigheid van eerstejaars hbo-studenten te verbeteren. In beginsel is het dan voldoende om naar de twee best beoordeelde methodes te kijken. Op dit punt bleek echter dat niet alle methodes meer beschikbaar waren en moest daarom genoegen worden genomen met de beste, beoordeelde methodes die nog wel beschikbaar waren.

Bij de beoordeling van onderwijsmethodes op didactische criteria speelt de keuze van de onderwijskundige theorie een grote rol. In dit geval was echter al gekozen voor het ABC-model dat was uitgewerkt in de vorm van een schema (zie Tabel 5.1 en 5.2). Het idee achter

dit schema, in overeenstemming met het ABC-model, was dat een studieboek niet alleen uitleg moest geven, maar vooral ook vragen en opdrachten en verder ook de antwoorden, zodat een student zichzelf kon checken. Voor een digitaal programma geldt uiteraard hetzelfde.

Het herziene schema (Tabel 5.2) bevat onder feedbackmiddel een toegevoegde categorie uitleg. Bij de eerdere beoordelingen is deze categorie niet gebruikt, dat gebruik heb ik overgenomen. Ik heb geprobeerd mijn oordeel te kwantificeren (op een schaal van 1 tot en met 5, waarbij 1 minimaal en 5 maximaal is), maar doordat iedere beoordelaar in de praktijk zijn eigen gemiddelde en standaarddeviatie heeft en het slechts om twee beoordelingen gaat, mag daar niet al te veel waarde aan worden gehecht.

Als tweede beoordelaar stond ik in dit geval niet geheel blind en onafhankelijk tegenover de te beoordelen methodes. Ik had eerder de uitkomsten van mijn mede-auteur gelezen en was op de hoogte van haar eindconclusies. Om de beoordeling zo zuiver mogelijk te houden, heb ik eerst beoordeeld en pas daarna de resultaten van de eerste beoordeling opnieuw nagelezen.

Een probleem bij de beoordeling vormt de doelstelling. In beginsel gaat het om een methode die geschikt moet zijn om eerstejaars hbo-studenten bij te spijkeren op het gebied van schrijfvaardigheid. De methodes hebben echter vaak een beperktere doelstelling, bijvoorbeeld beter spellen. Voor die beperkte doelstelling kan een methode optimaal zijn, terwijl de methode voor de wijdere doelstelling tekortschiet. Dit punt komt vooral terug bij de volledigheid en de relevantie van de oefeningen. Ik heb geprobeerd uiteindelijk de geschiktheid van de methode te beoordelen voor de wijdere doelstelling, waar het in ons onderzoek om ging. (Na afronding van de tweede beoordeling bleek dat de methodes bij de eerste beoordeling vooral beoordeeld waren vanuit de beperktere doelstelling van de methode zelf. Zie 5.3.1 Eindbeoording papieren methodes, eerste alinea. Zie 5.4, derde alinea.)

Kan die wijdere doelstelling preciezer omschreven worden? Tijdens de eerste beoordeling was die doelstelling nog vrij onduidelijk. Tijdens de ontwikkeling van TAVAN is als doelstelling gekozen: het verbeteren van foute zinnen (zie 6.1.2). Studenten in het hbo kunnen wel zinnen produceren, maar veel van de geproduceerde zinnen bevatten fouten. Studenten moeten in staat zijn de foute zinnen die ze produceren te herschrijven tot correcte zinnen.

Tot welke conclusies leidde de eerste beoordeling? Allereerst leidde de beoordeling tot een rangordening van de methodes zodat duidelijk werd wat kennelijk de beste, beschikbare methode was. Een tweede conclusie van de eerste beoordeling was dat ook de beste, beschikbare methode nog steeds belangrijke bezwaren en tekortkomingen had (zie 5.4).

Basisvaardigheden Spelling

Basisvaardigheden Spelling (BS) scoorde bij de eerste beoordeling als beste papieren methode. Inmiddels is dit boek niet meer verkrijgbaar, maar is er onder de titel Basisvaardigheden Spelling en interpunctie een herziene editie verschenen. Deze nieuwe editie heb ik niet beoordeeld om onnodige verschillen tussen beide beoordelingen te voorkomen. Bij de methode werd een CD-rom meegeleverd, maar inmiddels werkte die niet meer (bij de herziene editie wordt nu internet-ondersteuning gegeven). Voor de beoordeling heb ik me beperkt tot het boek.

De methode is bedoeld voor pabo-studenten die de pabo-taaltoets aan het einde van hun eerste jaar willen halen en bevat ook de stof die in de latere jaren van de pabo beheerst moet worden. Dit is een beperkte en concrete doelstelling. Dat maakt dat de methode heel praktisch en concreet. Voor de schrijfproblemen van hbo-studenten lijkt deze doelstelling me te beperkt. Eerstejaars hbo-studenten maken veel spelfouten, maar nog veel meer andere fouten (zie hoofdstuk 4). Op iedere bladzijde staat links de spellingsregel uitgelegd en staan rechts de oefeningen. Voor iedere oefening is achterin het juiste antwoord te vinden.

Bestaat er empirische evidentie dat deze methode werkt? Als die evidentie bestaat, wordt dat in ieder geval niet vermeld en ook zoeken op internet leverde in dit verband geen systematisch onderzoek. Wel vond ik vier beoordelingen van kopers die positief waren over hun aankoop.

De informatie die deze methode geeft over spellingsregels (de informatiebasis) beoordeel ik op alle punten uit het beoordelingsschema (juistheid, volledigheid, duidelijkheid, relevantie, toegankelijkheid) maximaal positief (5 x 5).

Als oefenboek bevat de methode veel spellingsoefeningen (veelheid: 5), die niet te moeilijk zijn (gemakkelijkheid: 5). Bij alle behandelde spellingsproblemen worden ook oefeningen

gegeven, de volledigheid van de oefeningen is vanuit de beperkte doelstelling van het boek zelf goed. Om hbo-studenten beter te leren schrijven, zijn ze te beperkt (2). De relevantie van de oefeningen om de behandelde stof te leren toepassen is vanuit de beperkte doelstelling goed, maar doordat het vaak om invuloefeningen gaat, lijken ze voor hbo-studenten niet erg realistisch. Je wordt al bij voorbaat geattendeerd op het probleem (1). De oefeningen zijn geordend naar spellingsregel en blijven relatief eenvoudig. Dat maakt ze weinig realistisch en weinig effectief. De geordendheid beoordeel ik daarom als slecht (1).

Van iedere oefening is achterin het antwoord te vinden. Een student kan dus zichzelf checken. De veelheid van de feedback is daarmee prima (5) evenals de betrouwbaarheid (5) en de duidelijkheid (5). Zelf nakijken bij een papieren methode kost nogal wat tijd en wil eigenlijk alleen als eerst alle opgaven van een oefening gemaakt zijn. Verder is het eenvoudig het antwoord (onbedoeld) te zien zonder eerst daadwerkelijk antwoord gegeven te hebben. De snelheid en de afhankelijkheid van de feedback scoren daarmee beide laag (2 x 1). Wanneer je het boek van begin tot eind doorwerkt, is voortdurend duidelijk waar je bent en wat je nog moet doen. De voortgangsinformatie is dan goed (5).

Mijn eindoordeel: als informatiebasis 5, als oefenboek 2.8 en als feedbackmiddel 3.7. Deze gemiddelden zijn in Figuur 1 weergegeven als de linker eindpunten van de drie lijnen. Het rechtereindpunt is de gemiddelde score van CambiumNed (CN). In Figuur 2 zijn de gemiddelde scores van de eerste beoordeling voor beide methodes weergegeven.

Ik ben vooral positief over de uitleg in BS. De uitleg is beknopt en ter zake. Er zitten veel oefeningen in en bij alle oefeningen is er feedback mogelijk. Dit is op papier ongeveer wat er maximaal mogelijk lijkt. Voor iemand die gemotiveerd is en problemen met spelling heeft, lijkt dit me een prima boek. Voor de doorsnee hbo-student die problemen heeft met spelling, lijkt BS me een brug te ver door de vereiste zelfstudie.

Als methode om hbo-studenten in 20 uur beter te leren schrijven, zie ik een aantal problemen. Mijn oordeel over de informatie (uitleg) is zeer positief, maar wanneer studenten moeite hebben met het herschrijven van foute zinnen, moet je volgens het ABC-model vooral oefenen met het herschrijven van foute zinnen en daar ook van uitgaan. Deze methode gaat in de eerste plaats van spellingsregels uit, waarna de opgaven erbij gemaakt zijn. De vraag is dus of die theorie wel zo nodig is en zo centraal moet staan.

Wat dan overblijft zijn de opgaven en de feedback. Beide scoorden lang niet maximaal. De opgaven zijn voor de wijdere doelstelling duidelijk te beperkt. Doordat de feedback opgezocht moet worden, werkt dat traag en krijg je gemakkelijk dat per ongeluk een volgend antwoord te zien. Een digitale methode zou op dit punt meer mogelijkheden bieden.

CambiumNed

Het digitale programma dat bij de eerste beoordeling het hoogst beoordeeld werd, Nedercom, zou volgens de site van de uitgever onmiddellijk beschikbaar zijn, maar werd in werkelijkheid niet geleverd. De site CabiumNed (CN) scoorde bij de eerste beoordeling een gedeelde tweede plaats en bleek online en gratis beschikbaar te zijn.

Het probleem met CN is dat het wat een chaos is. Het bevat veel uiteenlopende oefeningen, informatie, links en afleidende en storende reclame. CN is meer een grote collectie materiaal en oefeningen, dan een systematische methode. Het lijkt daarom weinig zinvol te vragen of er empirische evidentie is voor de effectiviteit, omdat onduidelijk is waar die vraag dan precies betrekking op heeft.

Wanneer ik het beoordelingsschema volg, is de juistheid van de informatie wel in orde (5). Maar kun je in dit geval spreken van volledigheid als vaak onduidelijk is, waar iets staat en waar je moet zoeken (2)? De uitleg die er staat, is vaak zo beknopt en staat zo tussen allerhande afleidende informatie, dat de uitleg onduidelijk wordt (2). Door al die afleidende informatie beoordeel ik de relevantie als matig (2) en de toegankelijkheid ook (2): het is volstrekt onduidelijk waar je moet zoeken. Als informatiebasis valt CN niet echt aan te raden. Dat sluit niet uit dat er misschien soms ook nuttige en bruikbare stukjes uitleg op deze site te vinden zijn.

CN bevat zeer veel oefeningen (5), maar ook nu ontbreekt een duidelijke structuur. De geordendheid is daardoor minimaal (1). Bij een boek is er nog altijd de lineaire structuur van het boek, maar ook dat ontbreekt in dit geval. De opdrachten liggen soms op een simpel niveau, en andere keren op een niveau dat ver uitgaat boven hbo-niveau. De gemakkelijkheid heb ik daarom beoordeeld als matig (3). Zijn deze opdrachten relevant? Wel om specifieke taalregels in te oefenen, amper als het doel is om te leren foute zinnen te corrigeren (2). Dekken de oefeningen het gehele gebied dat beheerst moet worden door hbo-studenten, de volledigheid? Nee (2). Als oefenboek vind ik CN voor hbo-studenten te ongestructureerd

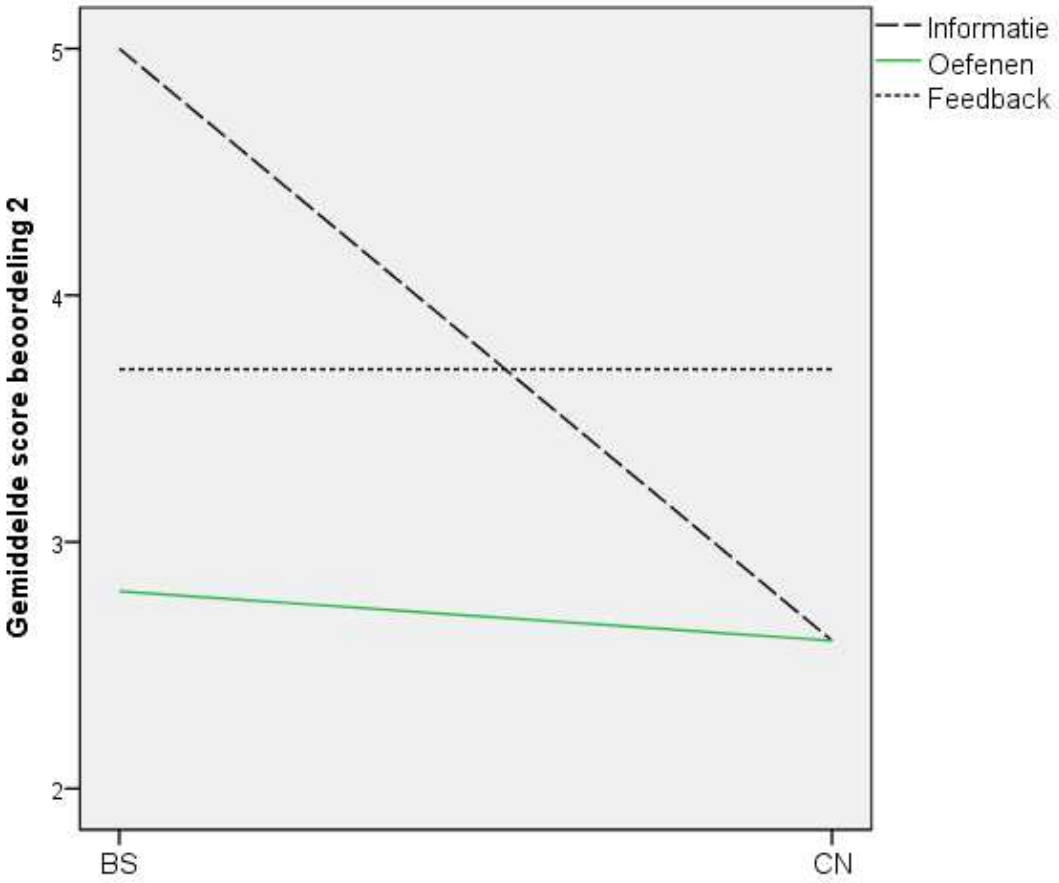
en weinig realistisch. Voor een zeer gemotiveerde zelfstudie-student bevat het echter een schat aan oefenmateriaal.

CN geeft op iedere oefening feedback (veelheid: 5) en die feedback klopt ook vrijwel altijd (betrouwbaarheid: 5). Verder is die feedback wel duidelijk als het antwoord goed is, maar na een herhaald fout antwoord wordt het juiste antwoord niet weggegeven zodat men in dat geval niet verder komt (3). Het voordeel hiervan is dat men het goede antwoord alleen krijgt door het zelf te vinden, de positieve feedback is volledig afhankelijk van het antwoord (5). Vaak is het de bedoeling dat een oefening van bijvoorbeeld 15 zinnen eerst volledig gemaakt wordt, voordat er feedback gevraagd wordt. De snelheid van de feedback is daardoor matig (3). Doordat iedere structuur ontbreekt, ontbreekt alle voortgangsinformatie (1).

Mijn eindoordeel: als informatiebasis 2.6, als oefenboek 2.6 en als feedbackmiddel 3.7. Deze gemiddelde scores zijn als de rechter lijnpunten in Figuur 1 terug te vinden. Als informatiebasis is CN volgens mij minder geslaagd dan BS dat op dit punt zeer goed (5) scoorde. Ook bij de eerste beoordeling werd een groot verschil gevonden in het voordeel van BS (4.4 tegen 3.4) tussen beide methodes op dit punt. Voor beide beoordelingen loopt de informatielijn schuin naar beneden. Bij beide beoordeling werd de informatie in BS hoger gewaardeerd.

De oefeningen en de feedback samen van beide methodes beoordeel ik ongeveer even hoog (BS: 2.8 en 3.7; CN: 2.6 en 3.7), wat ook bij de eerste beoordeling het geval was (BS: 4.0 en 3.5; CN: 4.0 en 3.5). Zo wel in Figuur 1 als in Figuur 2 lopen de oefenlijn en de feedbacklijn vrijwel horizontaal. De methodes verschillen op deze punten bij beoordelingen niet wezenlijk van elkaar.

In beide gevallen zijn de oefeningen volgens mij niet wat ze voor hbo-studenten zouden moeten zijn. Bij de eerste beoordeling werden de oefeningen positiever beoordeeld dan de feedback doordat voor die beoordeling werd uitgegaan van de doelstelling van de methode. Bij de tweede beoordeling is uitgegaan van de wijdere doelstelling, welke opgaven nodig zijn om bij hbo-studenten de schrijfvaardigheid te verbeteren. In Figuur 1 valt dit te zien doordat de oefenlijn de laagste horizontale lijn is. In Figuur 2 is dit de hoogste horizontale lijn. Doordat op dit punt werd uitgegaan van verschillende doelstellingen verschillen ook de toegekende scores belangrijk.

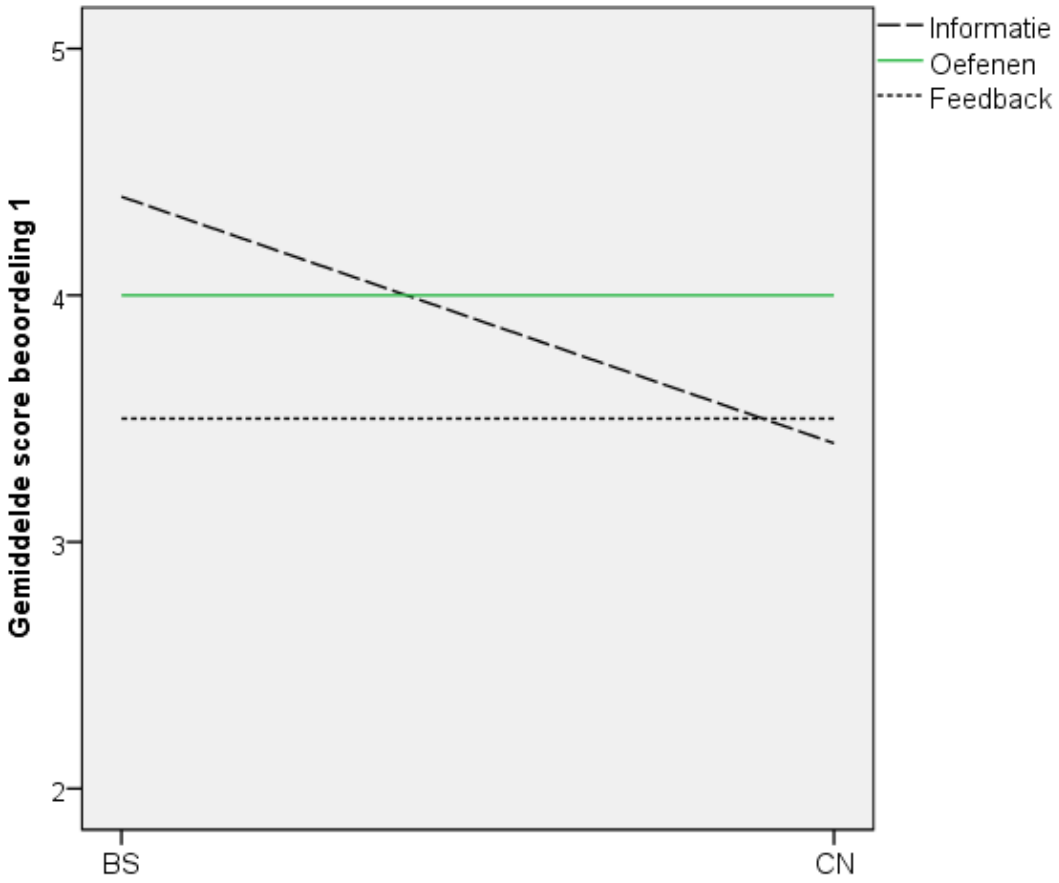


Figuur 1 De gemiddelde scores van beoordeling 2 voor Basisvaardigheden Spelling (BS) en CambiumNed (CN). Beide methodes scoren vergelijkbaar behalve op Informatie.

Er zitten veel oefeningen in CN, maar een duidelijke structuur ontbreekt. Een digitale methode zou op het punt van feedback volgens mij hoger moeten kunnen scoren dan het papieren BS, maar CN scoort op dit punt bij mij even hoog als BS. Dit was ook bij de eerste beoordeling het geval.

Conclusies en discussie

Op grond van het beoordelingsschema scoorde BS het hoogst bij de tweede beoordeling. Strikt genomen gaat het echter vooral om het oefenen (de opgaven) en de feedback en scoren beide methodes op die punten vrijwel gelijk. Mijn persoonlijke reden om dan toch voor BS te kiezen zou de duidelijke structuur zijn.



Figuur 2 De gemiddelde scores van beoordeling 1 voor Basisvaardigheden Spelling (BS) en CambiumNed (CN). Beide methodes scoren vergelijkbaar behalve op Informatie.

Ook bij de eerste beoordeling scoorde BS belangrijk beter dan CN doordat de theorie/uitleg hoger werd beoordeeld. Ook toen scoorden beide methodes gelijk op de oefeningen en de feedback. Beide beoordelingen leveren daarmee ten aanzien van de vraag naar de beste methode, dezelfde uitkomst. In dit geval zijn de verschillen tussen beide methodes BS en CN klein (de twee beste methodes zijn voor zover beschikbaar geselecteerd) waardoor de kans op uiteenlopende uitkomsten aanzienlijk was. Ondanks die kleine verschillen tussen de methodes leveren beide beoordelingen voor deze vraag dezelfde uitkomsten.

Wanneer we uitgaan van BS als beste, beschikbare methode is de volgende vraag die de tweede beoordeling moet beantwoorden of we mogen verwachten dat BS effectief zal zijn om hbo-studenten in 20 uur bij te spijkeren op schrijfvaardigheidsgebied.

Uitgaande van het ABC-model moeten we dan vooral focussen op de opgaven en de feedback. De opgaven in BS zijn beperkt tot spellingsproblemen. Verder zijn de opgaven per spellingsregel geordend en vaak niet al te moeilijk. De opgaven staan daarmee ver af van de zinnen die studenten zelf formuleren en moeten kunnen corrigeren. Een tweede probleem is dat de feedback bij een papieren methode niet optimaal werkt. Op basis van de opgaven en de feedback wordt daarmee verwacht dat BS niet optimaal zal werken.

Bij de eerste beoordeling werd een soortgelijke conclusie getrokken voor Nedercom dat bij die beoordeling nog iets hoger scoorde dan BS. Ook op dit punt leveren beide beoordelingen daarmee dezelfde conclusie op.

Hieruit mag niet afgeleid worden dat een tweede beoordeling op basis van het beoordelingsschema altijd hetzelfde resultaat zal opleveren als een eerste beoordeling. De conclusie moet eerder tegenovergesteld zijn: het beoordelingsschema heeft een eerste, methodologisch gezien niet al te strenge, check op de betrouwbaarheid doorstaan. Verder is betrouwbaarheid nog geen garantie voor validiteit.

Mogelijke verbeterpunten voor het beoordelingsschema

Hoewel het beoordelingsschema bij een tweede beoordeling door een andere beoordelaar van twee methodes leidt tot dezelfde conclusies voor beide hoofdvragen, komen bij deze tweede beoordeling ook een aantal beperkingen van het beoordelingsschema naar voren, zoals hieronder zal blijken.

Gesteld dat de opgaven van BS zouden worden aangepast en dat we vervolgens in verband met de feedback kozen voor een digitale variant. In feite komen we dan in de buurt van een TAVAN-achtig programma. Dat zou twee belangrijke verdere pluspunten kunnen opleveren.

De aanbieding van de opgaven kan dan geautomatiseerd worden, wat structurerend zou kunnen werken voor het leerproces. Juist op het punt van een gestructureerde aanbieding van opgaven en leerstof scoorde CN zo slecht. Geautomatiseerde aanbieding biedt verder bijvoorbeeld de mogelijkheid verschillende studenten verschillende opgaven aan te bieden of de volgorde per student anders te kiezen, waardoor afkijken en onbedoeld samenwerken kan worden tegengegaan. Een andere optie is de moeilijkheid van de opgaven te laten afhangen van het niveau van de student.

Een tweede belangrijk pluspunt is in het geval van digitalisering dat de vorderingen van de studenten bij gebruik van een online-programma op een centrale server kunnen worden bijgehouden. Wanneer een papieren methode als BS niet geschikt is voor zelfstudie bij eerstejaars hbo-studenten, zoals we eerder veronderstelden, biedt digitalisering in beginsel de mogelijkheid tot een begeleide vorm van zelfstudie.

Beide punten, automatische aanbieding van opgaven en registratie van resultaten, zijn niet opgenomen in het beoordelingsschema doordat dit indertijd primair bedoeld was studieteksten te beoordelen waarbij werd uitgegaan van zelfstudie. Het idee dat studeren misschien beter via de computer kon, was toen nog vooral een idee.

Verder was op dat moment ook nog niet duidelijk dat registratie van de resultaten belangrijk was. De veronderstelling was dat het voor een student zo belonend was om het goede antwoord te geven, dat studenten geen verdere aanmoediging nodig zouden hebben. Na bijna drie jaar ervaring met TAVAN lijkt deze veronderstelling wat erg optimistisch. Een verdere toepassing is dat de resultaten van het doorwerken gebruikt kunnen worden om de moeilijkheid van de opgaven te checken. Al te moeilijke opgaven werken frustrerend.

Een ander punt waarop het beoordelingsschema in feite afwijkt van het ABC-leermodel is de hoofdcategorie Informatiebasis. In het ABC-leermodel wordt de benodigde informatie gezien als een onderdeel van de situatie, als een Antecedent. In het beoordelingsschema is daarvan afgeweken doordat studieboeken sterk informatiegericht zijn. Een beoordelingsschema voor studieboeken waarin de opgaven en de feedback centraal zouden staan, leek indertijd een brug te ver. Om die reden werd informatie als eerste hoofdcategorie opgenomen. In een methode als TAVAN staat echter niet meer de informatie centraal, maar staan de opgaven voorop. Pas nadat een opgave niet gelukt is, wordt informatie gegeven over hoe het dan wel had gemoeten. De uitleg die het programma op deze manier bevat, zit gekoppeld aan de feedback in de vorm van het docentantwoord. Voor zover het programma informatie en uitleg bevat, is dit dus een zeer beperkte vorm van uitleg die pas als laatste bij een item wordt gegeven. Uitgaande van dit punt zou het beoordelingsschema mogelijk beperkt kunnen worden tot twee hoofdcategorieën: Oefenboek en Feedbackmiddel.

Het punt geordendheid van de opgaven was wel in het beoordelingsschema opgenomen en werd nog omschreven als oplopend in moeilijkheidsgraad, omdat het beoordelingsschema

uitging van een studieboek waarin het materiaal lineair geordend is. Bij zowel BS als CN blijkt echter dat de opgaven geordend zijn naar taalregel. Het gevolg is dat de clou van de opgave als het ware al bij voorbaat weggegeven wordt, waardoor de opgaven veel minder effectief worden.

Een belangrijk punt dat al eerdere aan de orde kwam, is het doelstellingen-probleem. Moet een methode beoordeeld worden op basis van de doelstelling van de stof of moet een methode beoordeeld worden op basis van de doelstelling die men wil of moet realiseren bij studenten. Omdat de bedoeling van het beoordelingsschema is de beste methode voor een specifieke stuk onderwijs te kiezen, lijkt te moeten worden uitgegaan van het laatste. De methode is vanuit de doelstelling van de methode zelf misschien perfect, maar voor het doel van de docent of onderwijsconstructeur niet en dient dan op het desbetreffende punt laag beoordeeld te worden. In ieder geval bleek uit het verschil in de beoordelingen bij relevantie en volledigheid van de opgaven dat het beoordelingsschema op deze punten niet duidelijk was.

D A N K W O O R D E N

Dankwoord

Anouk van Eerden

Promoveren wordt wel voorgesteld als een solitaire worsteling, maar dat is niet hoe ik het ervaren heb. Onderzoek doen vergt tijd en is daarmee een kwestie van geld. Een dag per week werd gefinancierd door het Instituut voor Marketing Management van de Hanzehogeschool Groningen waaraan ik als docent Nederlands verbonden ben, een tweede dag kreeg ik eveneens van de Hanzehogeschool en een derde dag van de Rijksuniversiteit Groningen. Dit leverde een onmisbare basis, maar was niet toereikend voor het hier gepresenteerde onderzoek. Ik heb de investering van veel eigen tijd niet als een obstakel ervaren, misschien mede daardoor heb ik genoten van deze periode. Ik ben heel erkentelijk voor de gelegenheid die ik kreeg om te promoveren.

Onderzoek doen is ook een kwestie van andere mensen die door hun hulp en bereidwilligheid het onderzoek mogelijk maken. Zo zag Frans Donders, de voormalige dean van het Instituut voor Marketing Management, onmiddellijk de mogelijke waarde van het onderzoek voor de onderwijspraktijk, wat voor de start ervan bepalend is geweest. De huidige dean Paul Ganzeboom heeft met mij meegedacht toen mijn promotietraject een onverwachte wending nam. Mijn teamleider Hanneke Barents zorgde ervoor dat mijn onderwijsverplichtingen daadwerkelijk tot één dag gereduceerd werden, zodat ik de rust had om aan het onderzoek te werken. De effectmeting van het nieuwe schrijfvaardigheidsprogramma bracht extra werk met zich mee voor verschillende collega's. Sonja van Dijk dacht intensief mee over de opzet van het experiment. Cisca Eskes wist inlogproblemen op te lossen. Van Jacqueline Boersema en Saskia Boer kreeg ik alle informatie die ik nodig had over de eerstejaarsstudenten en Jacqueline Heikema plande geschikte uren in het rooster.

Bij mijn onderzoek naar het effect van fouten werd ik eveneens door verschillende mensen geholpen. Op voorspraak van Marjolijn Verspoor gaven Hilde Hacquebord, Ann Hoag en Remco Knooihuizen mij tijdens hun colleges ruimschoots de gelegenheid om teksten te laten beoordelen door universitaire studenten. Voor hen betekende dat een irritante onderbreking, maar voor mij was dat van grote waarde.

Dr. Jeannette Doornenbal, lector Integraal Jeugdbeleid aan de Hanzehogeschool Groningen, heeft bijgedragen aan de eerste drie deelstudies. Zij keek vanuit een ander perspectief naar het onderzoek en signaleerde daardoor zaken die ik miste. Van essentieel belang voor de eerste drie deelstudies was professor Greetje van der Werf. Haar kritische houding tegenover niet goed onderzochte onderwijsvernieuwingen vormde een gemeenschappelijke basis. Voor haar stimulerende en zinvolle commentaar dank ik haar.

Onze promotor, professor Kees de Bot, heeft met veel wijsheid deze dubbelpromotie naar een goed einde geleid. Hij bleek bovendien bijzonder aardig te zijn. Kees was intensief betrokken bij de laatste drie deelstudies, waarbij hij ons vrijliet en tegelijk wel degelijk stuurde. Zijn aandeel bij de afronding van het onderzoek was van cruciale waarde. Ik ken niemand die zo snel en efficiënt kan werken als Kees.

Ook mijn kinderen hebben een bijdrage geleverd. Sara had de neiging om dieptegesprekken met mij te willen voeren op juist die momenten dat ik middenin een lastig onderdeel zat en Ana wist mij op geheel eigen wijze van het onderzoek af te leiden. Heel leuk vind ik het dat zij beiden paranimf willen zijn. Pim droeg bij door voor het online-programma de juiste programmeertaal op te sporen en de webserver te regelen. Als dank mocht hij onder het eten naar eindeloze conversaties over het onderzoek luisteren, wat hij nogal irritant vond.

Onderzoek doen is ten slotte een proces waarin de onderzoeker zelf een rol speelt. De uitvoering van het onderzoek was een reis van beginsituatie naar einddoel. Dankzij de samenwerking met mijn levensgezel Mik was deze reis geen eenzaam avontuur, maar het samen met hem verkennen van onbekend terrein. Het einddoel moest duidelijk zijn en uiteindelijk bereikt worden, hoewel de weg waarlangs dat moest gebeuren aan het begin van het onderzoek niet altijd helder was. Op een aantal punten was de kaart onduidelijk en moest een weg gezocht worden, wat de reis spannend en interessant maakte. Zonder de mensen langs de route die hun toestemming of hulp gaven, was het einddoel nooit bereikt. Iedereen die dit onderzoek mogelijk maakte, dank ik zeer voor zijn bijdrage.

Dankwoord

Mik van Es

Ik wil allereerst opmerken, dat het onderzoek dat in dit proefschrift beschreven is, alleen mogelijk was dankzij de bereidheid van de Hanzehogeschool Groningen de aanvraag voor een promotiebeurs van mijn mede-auteur te honoreren. Verder was dit onderzoek en de verdere ontwikkeling van het TAVAN-programma alleen mogelijk dankzij de bereidheid van het Instituut voor Marketing Management van de Hanzehogeschool Groningen om TAVAN uit te proberen en in te voeren.

Een belangrijk uitgangspunt bij het onderzoek was dat waar mogelijk gekwantificeerd moet worden. Wanneer ik dat probeer te doen voor de totale tijd die het onderzoek mij gekost heeft (exclusief de tijd benodigd voor het ontwikkelen van het online-programma), kom ik uit op een indrukwekkend aantal uren. Bij de start van het onderzoek was dat niet duidelijk. Op dat moment lag er alleen een interessante vraagstelling en een verzoek van Anouk voor methodologische en statistische ondersteuning bij de uitvoering van het onderzoek. Tegelijkertijd was op dat moment al wel duidelijk dat het totale onderzoek belangrijk omvangrijker was dan het beschikbare budget aan tijd toeliet. Vervolgens bleken de resultaten die gevonden werden vaak zo interessant te zijn, dat het onderzoek voortdurend groter groeide. Door Anouk raakte ik in dit gigantische avontuur betrokken. Hoewel de totale investering aan tijd fors was, ben ik met het resultaat zeer tevreden.

Samen een groot onderzoek uitvoeren en samen een proefschrift schrijven, is vragen om problemen. Er zijn dus goede redenen aan te voeren om dat liever niet te doen. Samen een proefschrift schrijven heeft echter ook belangrijke voordelen. Het totale onderzoek kon daardoor veel omvangrijker worden en veel dieper gaan dan anders mogelijk was geweest. Door het samenwerken is het mogelijk de sterke punten van beide partners te combineren en daarmee de zwakke punten van ieder afzonderlijk af te dekken. Een belangrijk voordeel van samenwerken is het gezamenlijk kunnen overleggen en het elkaar in de rails houden. Hoewel ik ook de problemen zie van samenwerken, denk ik, lettend op de uitkomsten, dat het misschien vaker gedaan zou moeten worden.

Mijn geklaag over de grote investering aan tijd die dit onderzoek mij kostte, moet ik relativeren door te vermelden dat ik gedurende de vele jaren van mijn aanstelling bij de Faculteit

der Letteren van de RUG ongeveer vierduizend uur aan onderzoekstijd kreeg om te promoveren in de vorm van een halve dag per week. Deze tijd heb ik in beginsel besteed aan een onderzoek naar leesvaardigheid. In de praktijk plachten ook onderwijsverplichtingen vaak een aanslag te doen op deze uren. Hoewel dat onderzoek niet resulteerde in een promotie, leverde het werken eraan wel een stuk ervaring voor mijn inbreng bij het in dit proefschrift beschreven onderzoek.

Professor Greetje van der Werf speelde een belangrijke rol in de eerste fase van het onderzoek. Allereerst was haar kritische instelling tegenover niet goed onderzochte onderwijsvernieuwingen een belangrijke voedingsbodem. Ten tweede zorgde haar positieve instelling tegenover de resultaten, dat de voortgang optimaal was.

Nadat het onderzoek in zwaar weer terechtgekomen was, constateerde dr. Marjolijn Verspoor dat het zonde was goed onderzoek ongepubliceerd te laten liggen. Professor Kees de Bot toonde zich vervolgens bereid het onderzoek vlot te trekken. Ik ben hem daar zeer erkentelijk voor. Dankzij zijn begeleiding lukte het in iets meer dan één jaar tijd drie nieuwe deelonderzoeken te realiseren, zodat het aantal deelstudies verdubbelde. Een onvoorzien probleem dat hierdoor ontstond, was dat de hoeveelheid informatie voor de lezer te groot dreigde te worden.

De leden van de beoordelingscommissie verdienen dank voor het doorwerken van het manuscript. Van twee leden van de beoordelingscommissie ontvingen wij commentaar op de eerste versie van het manuscript. Het verwerken van dit commentaar kostte soms veel tijd, maar resulteerde in een aantal belangrijke verbeteringen en aanvullingen.

Een aantal mensen was niet rechtstreeks betrokken bij het onderzoek, maar is indirect wel van groot belang geweest.

In dit verband wil ik allereerst meester De Haan noemen, die me in klas 6 van wat toen de 'lagere school' heette, tekstbegrip heeft bijgebracht. Behalve dat ik daar mijn leven lang plezier van heb gehad, heeft dat ook in het onderzoek een belangrijke rol gespeeld bij het doornemen van de literatuur en bij het schrijven van het manuscript.

Professor Ivo Molenaar verdient erkenning, omdat ik dankzij zijn inzet geleerd heb statistiek te fileren. Ik herinner me in dit verband nog zijn artikel "Ik word nog eens ziek van de statistiek" dat veel concrete voorbeelden gaf van zaken die je vooral niet moest doen.

Mijn eerste introductie in de psychometrie dank ik vooral aan mijn studievriend Frank Brokken, aan het boek van Nunnally 'Psychometric Theory' (1967) en aan het heldere stencil met opgaven van Jos ten Berge, die later hoogleraar zou worden.

Dat meten meer inhoudt dan alleen correlaties tussen variabelen zoals psychometrici weleens geneigd zijn te denken, leerde ik via het werk van Bridgman en Einstein. Aan de ene kant was er de nieuwe interpretatie van het equivalentieprincipe. Als zware massa precies gelijk is aan trage massa, is dat wel wat erg toevallig. Psychometrisch gezien is hier echter slechts sprake van één factor die tweemaal gemeten is. Aan de andere kant was er het klokkenexperiment (Hafele-Keating experiment) waarbij twee klokken na verloop van tijd verschillend aanwijzen. Omdat twee soortgelijke meetinstrumenten dezelfde uitkomsten moeten leveren, moet ook daar een verklaring voor zijn (Macdonald, 2013). Psychometrisch gezien zou dit verschil echter afgedaan zijn als meetonbetrouwbaarheid. Beide voorbeelden laten zien dat het onverstandig is alleen naar de meetuitkomsten te kijken en dat ook de meetmethode relevant kan zijn en de omstandigheden waaronder gemeten wordt.

In het onderzoek speelt een op oefenen gerichte aanpak van het onderwijs (het ABC-model: opdracht - antwoord - feedback) een belangrijke rol. De grote effecten van die benadering leerde ik voor het eerst kennen op 11-jarige leeftijd via meester De Haan. Iedere dag beantwoordden we in een groep van vier leerlingen een serie vragen over een tekst op papier en werden aan het einde van de schooldag de antwoorden met ons doorgenomen. Via een soortgelijke benadering leerde ik bij professor Ivo Molenaar in een kleine groep studenten 'voortgezette' statistiek. In een latere opzet voor een ander studie-onderdeel werkte hij met feedback-items waardoor de feedback directer werd en het vele nakijkwerk verdween. Van beide docenten leerde ik aan de ene kant een belangrijke vaardigheid, terwijl aan de andere kant beide docenten een oefenbenadering hanteerden als onderwijsmethode.

Een methode die hier sterk op leek, leerde ik kennen via mijn studie onderwijspsychologie: geprogrammeerde instructie volgens Skinner. Na eerst uitvoerig geëxperimenteerd te hebben met leermachines (Skinner, 1958), kwam hij op het idee dat hetzelfde ook in boekvorm gerealiseerd kon worden. 'The Analysis of Behavior' (Holland & Skinner, 1961) vormde in dit verband het grote voorbeeld. De volledige leerstof was uitgewerkt in korte items waar de student het juiste woord moest invullen. Door de bladzij om te slaan zag je het juiste antwoord staan en kreeg je de volgende vraag.

Een eerste gelegenheid om zelf ervaring op te doen met een oefenbenadering deed zich voor rond 1978. Aart Velthuijsen verscheen op het COWO (Centrum Onderzoek Wetenschappelijk Onderwijs van de Universiteit van Amsterdam) en verzocht om onderwijskundige assistentie bij het opzetten van een cursus psychologie voor pedagogiek-studenten. In het werkboek dat hij bij het studieboek maakte, werden onder andere per hoofdstuk een serie feedback-items opgenomen. Vanuit het COWO peilden we de meningen van de studenten over de verschillende onderdelen van het werkboek. Over de feedback-items was men zeer positief (Van Es, Velthuijsen & Neervoort, 1980).

In het onderzoek speelde het TAVAN online-programma een belangrijke rol. Een aantal mensen en instanties speelde bij de ontwikkeling van het programma en de diverse voorlopers daarvan (Lesmaker en Oefenmachine) een rol.

Ter wille van de overzichtelijkheid verdeel ik de ontwikkeling in vijf fasen. In fase 1, de startfase, werd met de ontwikkeling van het programma begonnen. In fase 2 werd getracht het software- en courseware-probleem op te lossen. In fase 3 werd het user-interface-probleem opgelost. In fase 4 werd de methode in de praktijk ingezet en uitgeprobeerd. Fase 4 eindigt met het inzetten van het TAVAN online-programma voor de eerste keer (TAVAN1) en het onderzoeken van de effectiviteit daarvan zoals in het proefschrift beschreven (hoofdstuk 6) is. In fase 5 wordt het programma verder toegepast, ontwikkeld en geoptimaliseerd (TAVAN2 en TAVAN3). De resultaten van TAVAN2 vormden mede de basis voor het in hoofdstuk 8 beschreven onderzoek.

De eerste aanzet tot de ontwikkeling van dit programma (fase 1) vond plaats in 1981. In dat jaar verscheen de eerste IBM pc en begonnen microcomputers betaalbaar te worden. Een tweede belangrijke factor was de nakende mislukking van het PLATO-IV project en het toen al afgesloten PLATO proefproject van de Universiteit van Amsterdam dat op het COWO werd uitgevoerd (Camstra, Van Dijk & Van der Avoird, 1979). Belangrijk was vermoedelijk ook een bepaalde onderlinge rivaliteit. De werkgroep Curriculumontwikkeling waar ik in zat, placht in de praktijk vaak overhoop te liggen met de 'Plato-boys'. Ik had er daardoor geen problemen mee de zwakheden in PLATO-IV te zien en indien dat nut had, te benoemen. Op hun beurt zagen de collega's die zich verwant voelden met het Plato proefproject vooral de zwakheden en fouten in het microcomputer-plan dat ik opeens voorlegde. Ik kondigde aan een cursus 'educatief programmeren' te gaan volgen die uitging van micro-

computers en Basic en schetste in een korte notitie mijn plan. Mijn idee was dat microcomputers gemeengoed zouden worden en dat het een belangrijke eerste stap zou zijn als we erin slaagden met succes over te stappen van papieren geprogrammeerde instructie naar digitale geprogrammeerde instructie. De voordelen van een oefenbenadering zouden dan gecombineerd worden met de voordelen van een computer. Of een oefenbenadering via de computer echt goed zou werken, viel alleen vast te stellen door het uit te proberen, maar om het uit te proberen moest er eerst een goed werkend programma komen.

Dat idee werd niet bijzonder enthousiast ontvangen door sommige collega's. Men zag het PLATO-IV systeem als een perfect systeem, wat het in technisch opzicht vermoedelijk ook wel was en de keuze voor een ander en ook nog 'minderwaardig' systeem zag men (bij wijze van spreken) als heulen met de vijand. Dat er dan ook in Basic geprogrammeerd moest worden, was een ander euvel. Alleen Pascal en Tutor (de taal van PLATO-IV) konden gezien worden als respectabele computertalen. Mijn 'foute' keuze werd echter voor een belangrijk deel ingegeven door de informatie die ik uit het verslag van het PLATO proefproject en de mondelinge communicatie daarover had afgeleid.

In totaal zag ik een zestal problemen met het PLATO-IV systeem waarin CDC (Control Data Corporation) toen al meer dan een half miljard dollar (600 miljoen) geïnvesteerd had en waar men druk voor adverteerde (Engelstalige Wikipedia, 24-3-2014). Ten eerste kostte het systeem per studentuur 50 dollar voor de 'connect time'. Ten tweede kostte het ontwikkelen van een lesuur veel te veel tijd (soms wel duizend uur of meer). De kosten van het ontwikkelen van een lesuur konden daardoor oplopen tot driehonderd duizend dollar. Ten derde bleek de effectiviteit van het systeem in onderwijskundig opzicht onduidelijk te zijn. Het was niet duidelijk of het systeem effectiever was dan traditioneel onderwijs. Een vierde probleem was dat de ontwikkelaars en docenten die bij PLATO-IV betrokken waren, nogal enthousiast waren over het systeem, zo dat ze mijns inziens niet meer kritisch keken naar het functioneren van het systeem. Als vijfde probleem zag ik dat men niet expliciet voor een oefenbenadering gekozen had als onderwijsmethode, waardoor volgens mij het belangrijkste effecitiviteitsvoordeel van de computer niet werd benut. Ten slotte ging men niet uit van een geleidelijke benadering waarbij eerst een prototype gebouwd werd dat grondig getest werd, maar leek men de oplossing vooral te zoeken in een snelle, grootschalige invoering. Dankzij PLATO-IV leek daardoor vrij duidelijk hoe het in ieder geval niet moest.

Uitgaande van deze problemen streefde ik naar een kostprijs van 1 à 2 gulden per studentuur. Voor het tweede probleem was het doel te kunnen volstaan met 10 uur docenttijd voor het ontwikkelen van een lesuur door uit te gaan van een gestructureerde aanpak. Voor het derde en vijfde probleem wilde ik volledig focussen op een oefenbenadering, omdat ik dacht te weten dat die uitermate effectief kon zijn. Met betrekking tot het vierde probleem nam ik me voor me niet te veel laten meeslepen door het verleidelijke van de hardware, maar het feitelijke doel van het project in het oog te houden. Met betrekking tot het zesde probleem koos ik voor een evolutionaire aanpak, uitgaande van een te construeren prototype. Ik had gemerkt en geleerd dat een gestructureerde benadering waarbij de software stap voor stap werd opgebouwd en getest, belangrijk beter werkte dan de 'alles moet in één keer perfect werken' benadering (Dijkstra, 1969). In de praktijk bleek het echter toch iets minder simpel dan aanvankelijk gedacht.

In fase 2 begon de ontwikkeling van het programma met de cursus die aangeboden werd door een afdeling onderwijs van de Gemeente Haarlem. In eerste instantie ontwikkelde ik het programma voor de microcomputer van de cursus: een Sharp MZ-80K. Doordat deze machine maar eenmaal per week beschikbaar was, schoot dit niet erg op. De aanschaf van een homecomputer, een TI-99/4A was bedoeld dat probleem op te lossen. Het herschrijven van het programma kostte echter veel tijd. Op dit punt toonde Jan Willem Beek zich bereid het programma voor de Sharp te herschrijven voor de TI-99/4A en de eerste versie te voltooien.

Bij het testen van het programma werden nu twee problemen zichtbaar: de hardware was eigenlijk te beperkt en het user-interface (de interactie tussen het computerprogramma en de gebruiker) was niet optimaal en gaf voortdurend problemen. Fase 3 was begonnen. Tegen de tijd dat de eindversie van het programma klaar was en als Lesmaker op de markt gebracht werd, was de TI-99/4A volledig verouderd.

Samen met Adriaan (Jaan) Dijkhuizen was ik op initiatief van uitgeverij Stark-Texel al eerder gestart met de ontwikkeling van een versie van Lesmaker voor MSX computers (Van Es & Dijkhuizen, 1987). Later volgde nog een versie voor DOS computers (Van Es & Dijkhuizen, 1988). Aan een versie voor de C-64 werd wel begonnen, maar deze kwam nooit af. Om de problemen met het user-interface op te lossen voor DOS computers, werd later omgeschakeld naar de hogere orde taal OOPS van Theo Schijf (verschenen in 1987). OOPS was specifiek bedoeld voor educatief programmeren, platform onafhankelijk en daarbij vormge-

geven als een normale programmeertaal waarbij ook eigen commando's konden worden geschreven en toegevoegd. Dit leidde tot Oefenmachine waarin de interpreter gecombineerd was met een eigen editor en een menusysteem zodat het invoeren van de lesstof voor een deel werd geautomatiseerd.

Aan het einde van fase 3 waren een aantal zaken duidelijk. Allereerst bleek in de praktijk de hardware nog steeds beperkingen met zich mee te brengen. Computers waren niet of te weinig beschikbaar of waren te traag of tot te weinig in staat. Het software/courseware-probleem was aan het einde van deze fase opgelost. Het programmeren van een 'engine' of een interpreter die de les uitlas, gaf geen echte problemen meer en ook het ontwikkelen van een computerles verliep door de volledig gestructureerde aanpak in beginsel binnen de 10-uur-per-lesuur-norm.

Ook de problemen met het user-interface waren aan het einde van deze periode opgelost: het was duidelijk aan welke regels het user-interface moet voldoen. Bij de oplossing van het user-interface probleem was allereerst het ABC-model en het denken van Skinner over geprogrammeerde instructie en leermachines van belang. Verder was de werkwijze van Microsoft bij de ontwikkeling van Word die door Bill Gates publiek werd gemaakt, van belang.

De lange serie projecten in fase 3 leidde er uiteindelijk toe dat de computer een stuk educatief gereedschap werd waarvan de mogelijkheden precies bekend waren. Achteraf gezien zou men kunnen zeggen dat ook in dit geval een oefenbenadering bleek te werken.

Een nieuw (en achtste) probleem werd aan het einde van fase 3 duidelijk. Het heeft weinig zin een goed werkend computerprogramma te ontwikkelen zonder dat het praktisch wordt toegepast. Het doel moest daarmee niet langer zijn een computerprogramma of lesstof te construeren, maar een onderwijsprobleem op te lossen. Fase 4 was daarmee begonnen.

Een eerste mogelijkheid in deze richting deed zich voor bij het studie-onderdeel Methodologie en Statistiek I in 1995-1996. Marleen Kuiper (Kuiper, 1996) verzorgde in het kader van haar scriptie-onderzoek drie practica (2 uur, 2x10 opgaven per practum) voor in totaal 106 studenten. In feite werden er per practicum 10 vragen gegeven en 10 soortgelijke controlevragen. Het idee was dat studenten op de soortgelijke controlevraag verderop in de les, beter zouden scoren. De gemeten vooruitgang was echter vrijwel nul. Wel waren de studen-

ten zeer positief over deze vorm van onderwijs. Bij analyse achteraf bleek die minimale vooruitgang wel verklaarbaar met het ABC-model, maar niet met een cognitief model. De informatie die men kreeg na een fout, werd niet benut zoals het ABC-model ook voorspelt. Vanuit een cognitief model viel dit echter lastig te begrijpen, omdat de studenten inmiddels wel de benodigde kennis hadden gekregen. Het resultaat van deze eerste praktijkproef was echter niet dusdanig dat het verdere testen met kracht ter hand werd genomen.

Een tweede mogelijkheid deed zich voor toen professor John Nerbonne verzocht de opdrachten en de handleiding voor het practicum Statistiek I te herzien. Voor dit practicum ging ik nog uit van een volledig papieren opzet die als pdf via de monitor kon worden opgevraagd. John regelde het echter zo dat het juiste antwoord in het html-document met een muisklik in de browser verscheen, zodat het lastige opzoeken van het antwoord werd ondervangen. Het practicum (http://www.let.rug.nl/nerbonne/teach/Statistiek-I/practica/) is in deze vorm nog steeds beschikbaar. Uit de manier waarop de practica verliepen in verhouding met daarvoor, bleek het systeem goed te werken. Ook de toetsresultaten leken dat te bevestigen. Deze ervaring opgedaan bij Statistiek I vormde daarmee een belangrijke eerste bevestiging van het idee dat een gecomputeriseerde oefenbenadering goed kan werken.

Het probleem dat dankzij Anouk als uitgangspunt voor het promotie-onderzoek werd gekozen, leidde vervolgens via de route die in het proefschrift beschreven is, tot het inzicht dat een op internet gebaseerd computerprogramma de enig mogelijke oplossing was voor het gesignaleerde onderwijsprobleem. Een gevolg van die conclusie was dat er in een zeer laat stadium een volledig nieuw computerprogramma geconstrueerd en getest moest worden: (de eerste versie van) het TAVAN online-programma.

Bij de realisatie van het nieuwe programma leverde Pim van Es een cruciale bijdrage door de krachtige en flexibele 'programmeerbare web-applicatie server' Run Basic van Carl Gundel aan te bevelen en vervolgens de problemen met de server voor zijn rekening te nemen.

Door deze derde toepassing in de praktijk ontstond voor het eerst de mogelijkheid het effect van een oefenbenadering via een pretest-posttest-design in combinatie met een controlegroep vast te stellen. Hiermee werd het oorspronkelijke doel van het project dat in 1981 begonnen werd, namelijk onderzoeken of een gecomputeriseerde oefenbenadering effectiever zou zijn dan traditioneel onderwijs, gerealiseerd.

Fase 5 begon met TAVAN2, de tweede maal dat het TAVAN online-programma in de onderwijspraktijk toegepast werd. Hier werd een nieuw (en negende) probleem duidelijk dat ook al waarneembaar was bij TAVAN1: studenten hebben moeite de motivatie op te brengen om de computerlessen door te werken en hebben daar steun en structuur bij nodig. Verder verschoof de aandacht in deze fase naar het optimaliseren, het implementeren en het opschalen van het programma naar grotere aantallen studenten.

Om het grote belang van de afzonderlijke bijdragen duidelijk te maken, ben ik zo vrij het TAVAN online-programma in historisch perspectief te plaatsen.

De eerste leermachine werd gedemonstreerd door Sidney Pressey in 1924 en was in feite alleen nog maar een machine om (op een afzonderlijk papier gedrukte) testvragen te laten beantwoorden zonder dat de leerling feedback kreeg. Pas enkele jaren later kwam Pressey met een versie die feedback gaf (Mirande, 2006, p. 5). De mogelijkheden waren daarmee zeer beperkt.

Skinner introduceerde voor leermachines in 1958 het principe dat de student het antwoord zelf moest formuleren (in plaats van alleen het juiste antwoord kiezen) en ging in 1961 over op boekvorm (Holland & Skinner, 1961). Dat waren -- na bijna veertig jaar -- twee belangrijke verbeteringen en doorbraken. Skinner streefde verder naar niet te moeilijke opgaven en naar een goede opbouw van de opgaven.

De eerste serieuze toepassingen van de computer in het onderwijs dateren uit het begin van de jaren 60. Rond 1980 liep het PLATO-IV systeem ondanks investeringen van meer dan een half miljard dollar vast. Na twintig jaar was men in feite terug bij af.

Inmiddels schrijven we 2014. TAVAN wordt nu voor het derde jaar gegeven. Ruwweg 400 studenten herschrijven ieder meer dan 1500 zinnen. In totaal worden per jaar ongeveer 600 duizend zinnen herschreven. Wanneer deze zinnen door docenten zouden worden beoordeeld, zou dat per klas per lesuur ongeveer tien uur geestdodend nakijkwerk opleveren. In totaal zou de hoeveelheid werk overeenkomen met twee volledige aanstellingen gedurende het gehele studiejaar. Ook bij die grote inspanning zou de feedback voor de studenten echter te laat komen, zodat het totale leersysteem niet effectief zou zijn.

Vergeleken met de geprogrammeerde instructie van Holland en Skinner uit 1961 is het online-programma een duidelijke vooruitgang. Het werkt belangrijk beter en het kan belangrijk meer. Wat we nu realiseren met de computer is met papier niet meer realiseerbaar.

Ten opzichte van het PLATO-IV systeem zijn de drie hoofdbezwaren weggenomen. Ten eerste zijn de kosten van een uur met de computer werken niet langer een probleem. Ten tweede is de tijd die het kost een les te realiseren ook niet langer een probleem en goed op te brengen. Ten derde is het systeem nu belangrijk effectiever dan traditioneel onderwijs. (Men zou zelfs kunnen stellen dat het systeem meer dan honderd of duizend maal zo effectief is als het traditionele onderwijs. Honderd maal nul, de vooruitgang gemeten bij het traditionele programma, is immers nog steeds nul.)

In 2007 toen mijn mede-auteur haar onderzoeksvoorstel indiende voor het onderzoek dat in dit proefschrift beschreven wordt, leek het idee dat je met succes schrijfonderwijs zou kunnen geven via de computer niet realistisch. Misschien was het mogelijk met de computer leesvaardigheid te trainen, maar schrijfvaardigheid leek een brug te ver. Uit ons proefschrift blijkt dat die brug inmiddels (in 2010-2011) gepasseerd is en dat de ontwikkeling al weer iets verder is.

Tegelijkertijd is duidelijk, dat er nog veel niet duidelijk is. Er moet nog veel uitgezocht worden. Er kan nog veel verbeterd worden. Het is allemaal nog lang niet optimaal. Kortom: er is een begin gemaakt, maar er is (hopelijk) nog een lange weg te gaan.

Ik dank een ieder die, in welke vorm dan ook, een positieve bijdrage heeft geleverd aan deze fascinerende ontwikkeling.

Executive and Extended Summary

Measurement and Maximization of Basic Writing Skill of First-Year Students in Higher Education

A. Van Eerden

M. Van Es

University of Groningen

May 2014

Executive Summary

This study shows that basic writing skill is an important measure to predict dropout and to indicate the level of students. Three new methods are presented to measure basic writing skill. By using confirmed errors per A4 (500 words) it was possible to show that the basic writing skill of Dutch first-year students is far from ideal. The newly developed TAVAN program succeeded in ten 2-hour lessons to reduce the number of errors with 20%.

Many reports are available about the insufficient writing skill of first-year students. The objective of this study was to find ways to measure basic writing skill in a reliable way and to remediate insufficiencies.

Basic writing skill was defined as the number of confirmed errors (errors signalled by at least two independent expert-raters) per A4 (500 words). This measure showed to be highly reliable. First-year students of higher professional education made on average 81 confirmed errors per A4. First-year university students made on average 42 confirmed errors. The Dutch educational system seems to be not very effective in this respect, because the ideal value of 0 errors per A4 seems to be far off.

Basic writing skill could also reliably be measured as the number of signalled errors per A4 by a single expert rater. The correlation between signalled errors per A4 and confirmed errors per A4 was very high (0.93). Results for both measurement methods will therefore be mostly the same.

Seventeen paper methods and nine digital methods to remediate basic writing skill were rated on expected effectiveness. No method had clear objectives, no method had been shown to be effective. Focusing on the assignments and the feedback provisions of each method none was rated to be fully satisfactory.

The newly developed method TAVAN reduced the number of errors between pretest and posttest writings with 20% in ten two hour lessons. The control group taught by the traditional teaching method did not improve. TAVAN proved to be very effective in reducing the number of errors with an effect size of 1.1 standard deviation. The TAVAN

online program is not based on multiple choice questions, but actually asks students to correct and to rewrite the presented faulty sentence(s).

How serious are errors in texts? Using an experimental design and three texts written by students we found that two different versions of these texts without errors were rated by readers as 48 while the original texts were rated as 30. An effect size of 1.4 standard deviation. This result shows that errors in texts can have a big impact on readers. It is, therefore, important for students to learn to write with as few errors as possible.

Our study delivered in all three new methods to measure (basic) writing skill:

1. the number of (confirmed) errors per A4 (500 words);
2. the score in the TAVAN online program;
3. the score of the student as a holistic rater of texts (quality of holistic rating).

All three new methods and the traditional measure of writing skill, the holistic rating, measured to a large extent a common factor: the average correlation between the four variables after correction for attenuation was very high (.87).

The factor common to all four methods can be described as skill in EDC (Error Detection and Correction). While holistic raters might think they are focusing on the content of the texts, their ratings proved to be highly correlated with the number of errors per A4. The correlation after correction for attenuation was -.89.

A reliable measurement method for educational purposes is of limited use if no systematic training method is available to increase the score of students who fail the norm. The TAVAN online program might be the first measurement method for basic writing skill that satisfies this criterion. A second possible method, still to be researched further, might be the quality of holistic rating by students. Items with two short texts (one sentence each) can be presented to students to select the best option. By providing feedback it is expected that students will learn quickly to discriminate between well written and faulty sentences. In this way it might be able to show that close reading provides enough training for students to become better writers.

How much text is needed to assess writing skill? Using the number of errors per A4 (500 words) a short text of a half A4 (250 words) showed to be enough for a reliable assessment. There is no need to let students write long texts to assess their basic writing skill.

Extended Summary

Introduction

Many reports are available about the insufficient writing skill of Dutch first-year students. However, quantitative data that can be interpreted in a simple way are mostly lacking. The objective of this study was to find reliable ways to measure basic writing skill and to develop an effective method to remediate insufficiencies.

Writing skill is difficult to measure because raters hardly agree on the quality of writings by students, the so called 'holistic rating'. Consequently, it is difficult to determine whether a student has sufficient writing skill and whether he has made any progress.

A second, by teachers less accepted method to measure writing skill is the use of dedicated objective tests. These tests have been shown to be reliable and valid, but are of limited value for instructional purposes because students don't have to write and the measure cannot be used for writings. Another problem of these tests is that the score is difficult to interpret in a meaningful way, because the score is dependent on the difficulty of the test.

Study of errors

The first part of the study, the study of errors, tried to solve the measurement problem by using the number of errors made by students in their writings. This newly developed method used independent expert raters, who were instructed to underline, number and describe all errors present in the texts. Although raters did not always agree on single errors, they highly agreed on the number of errors per A4 (500 words).

Agreement between raters was measured by calculating the correlation between the number of errors per A4 in a series of texts. It was necessary to take into account the length of the writings, because students writing longer texts showed to make more errors than students writing shorter texts. After correction for length (number of words) the relation reversed: students writing longer texts produced less errors per A4. For the agreement between four expert raters, a mean correlation of .85 was found. This value (the reliability of a single rater) can be considered very high.

Despite this high agreement raters still differed in their mean and standard deviation. Some raters were more critical than others and signalled more errors. Therefore, it is still not easy to interpret the numbers of errors signalled by different raters. In order to solve this problem a second measurement method was developed: confirmed errors. Confirmed errors are errors that have been signalled by at least two independent raters. Therefore, the existence of a confirmed error is hard to doubt. Raters might indicate as many errors as they like, but if these errors are not also signalled by another independent rater, these errors will not be counted as confirmed.

When only the confirmed errors of the raters were used, the agreement between the raters about the number of errors per A4 was even higher than when all indicated errors were used. The mean correlation between the raters for confirmed errors per A4 showed to be . 93. The combined reliability of the four raters (rater alpha) was .98, almost perfectly reliable.

Although the numbers of confirmed errors can be interpreted more easily than errors signalled by single raters (confirmed errors have a meaningful zero value) confirmed errors did not lead to substantially different outcomes. The correlation between confirmed errors and signalled errors per A4 showed to be .93, which can be considered very high.

How many confirmed errors could be found in the texts of the first-year students? The sample of 30 texts (20 texts from first-year students in higher professional education and 10 from first-year university students) was checked for confirmed errors using four expert raters. First-year university students made on average 42 confirmed errors in one A4 text (500 words); first-year students of higher professional education made on average 81 confirmed errors. (SDs respectively 16 and 41). Compared to the ideal (0 errors per A4) both groups seem to be far off.

The study of errors also produced an overview of the types of errors produced by first-year students. The most frequent error types were: 'Wrong word', 'Faulty sentence', 'Punctuation', ‘Unnecessary word/words', 'Paragraph', 'Preposition', 'Spelling error' and 'Missing word'. Together, these eight categories made up for 75% of all confirmed errors made by first-year students. D/t errors in Dutch verbs, a notorious problem in the Dutch language, were indicated by every rater whenever these errors occurred. This was, however, contrary to expectation not very often (less than 2%).

Assessment of available writing methods

How effective were the available methods to increase basic writing skill? The second study tried to answer this question by assessing the writing methods that were available for teaching first-year students. Seventeen paper methods and nine digital methods were assessed. For the assessment a scheme based on the ABC-model (assignment, answer, feedback) was used. The method had to provide enough assignments of the right type and to deliver quick and immediate feedback.

Every assessed writing method missed a clear objective. No empirical research was available to show the effectiveness of any method. The feedback of the digital methods was rated as better than the feedback of the paper methods. For both types of methods the number of assignments and their order was rated the same on average and not fully satisfactory. The highest rated digital program, Nedercom, scored favourably on feedback, but moderately on the number of assignments and the order of these. Even the best paper and best digital method appeared to have important flaws.

A common problem with all methods was that the methods were aimed at all possible writing problems but, in general, not at the errors students actually made. 'Wrong word' has been indicated as the most frequent type of error in the study of errors, but this type was hardly dealt with. Other frequently occurring errors, such as 'Faulty sentence', 'Unnecessary word/words', 'Paragraph', 'Preposition' and 'Missing word' were also hardly practised, if at all, in the assessed methods. Combined, however, these types of errors made up three quarters of all confirmed errors students made.

Effect of the TAVAN writing program

The newly developed TAVAN program was tested on a class of first-year students in higher professional education. The final writings of the TAVAN group contained 19 errors less per A4 (500 words) than the writings at the start of the program. This meant a reduction of 20.5%. The control group that was taught by the traditional writing program did not improve. The difference between the experimental group and the control group regarding the reduction in number of errors, showed to be 1.1 standard deviation. This can be considered a large effect.

Basic writing skill was measured in this part of the study by asking students to rewrite texts containing errors and flaws and by asking students to rewrite sentences in the online program. The latter method showed to be extremely reliable, and correlated very highly with the number of errors per A4 in the combined pre and post writing and did not require human raters. Therefore, the online program does not only constitute a tool to improve basic writing skill, it also offers a simple, reliable and valid way to measure basic writing skill because students actually have to (re)write.

Basic writing skill as a measure showed to be more than just the number of errors per A4. The study of errors showed that students who made less errors per A4, wrote longer texts. The study into the effectiveness of the new program demonstrated that students with good basic writing skill needed less time to (re)write (the online program also measured the time), had gained a higher degree in their previous education, self assessed their writing skill more positively and showed less dropout. These results seem to indicate that basic writing skill is an important predictor for the level a student can reach.

The self assessment of writing ability by the students correlated rather highly with their measured writing ability ($r=-.67$, $p= .000$) at pretest, but showed to be not a valid measure to measure progress in writing skill. Students in the traditional program rated their own writing ability at the posttest significantly higher than at the pretest ($p=.01$), but did not increase according to their measured ability. The TAVAN students, however, increased in measured ability, but their self assessed writing ability did not increase ($p=.36$).

The attitude about writing (with items as: I like writing, The writing of a paper has to be quick) was measured at pretest and at posttest with an alpha reliability of .68. This measure proved to be not significantly correlated to basic writing skill.

The TAVAN students became better writers but did not increase in writing attitude. We could show, however, that the better than average students in the online program decreased in writing attitude, while the students who did worse than average increased in writing attitude ($r=.55$, $p=.02$, 2-tailed, correlation between TAVAN score and the difference in attitude). This change in attitude, however, was not correlated to the gain in measured basic writing skill (.00).

All in all the idea that the attitude about writing would be an important predictor of basic writing skill, was not confirmed. Basic writing ability seems to be mostly a matter of skill and not a matter of attitude or motivation.

Characteristics of the TAVAN method

Why did the TAVAN method manage to reduce the number of errors per A4 in ten two hour sessions, while the traditional teaching method did not and while the available methods did not convince in this respect? There is no sure answer to this question, but several issues are significantly different about TAVAN when compared to traditional writing skills instruction.

1. TAVAN has been developed with an explicit and measurable objective in mind: students should make less errors per A4.
2. TAVAN does not focus on the planning and the process of writing, but instead focuses only on editing and rewriting.
3. TAVAN is not based on (by the teacher) assumed errors, but is based on a list of errors students actually made.
4. The instructional method is fully based on the ABC-model (instruction is seen as a sequence of assignments and feedback) instead of the traditional lecturing model.
5. Instead of few large and vague assignments, many (1000) small (about 20 seconds) and specific assignments (rewrite this faulty sentence) are used in the online program.
6. The online program provides immediate and clear feedback.
7. The online program structures the learning situation by presenting automatically the next assignment (self paced) and keeps track of the results of the student.
8. The lecturer does not lecture, but acts as the manager of the learning system and the students.

In this study the online program was used only to increase basic writing skill. The same educational approach based on frequent testing and direct feedback (the ABC-model) instead of lecturing, might also be useful for the training of other skills.

Effect of errors in texts

How serious is an error? It is not possible to write without making errors, so why bother? Three original texts of students in which the errors were still present were rated by readers on a scale 0 to 100. On average these texts scored 30, after removal of the errors the same texts scored 48. A difference of 1.4 SD. This result shows that errors in texts can have a big impact on readers.

We also found a very high correlation (-0.89) after correction for attenuation between the number of errors per A4 in the texts and the holistic ratings of these 48 texts. This result indicates that the number of errors per 500 words in a text determine the holistic rating to a very large extent.

A third result of this part of the study was that students could be used as holistic raters. The holistic rating of the expert raters and of the students correlated after correction for attenuation 0.99 for 26 texts. The average correlation between student raters was 0.22, while the expert raters correlated 0.65. The student raters were on average much less reliable than the expert raters.

It was possible to use these ratings by the students to score the quality of their rating. This measure was not very reliable yet (0.42), but it was possible to show that students who are better writers are also better raters (r=0.31, p=0.041, 2-tailed, 44 texts).

Measurement and training of basic writing skill

Our study delivered in all three new methods to measure (basic) writing skill:

1. the number of (confirmed) errors per A4 (500 words) in a text;
2. the score in the TAVAN online program;
3. the score of the student as a holistic rater of texts (quality of holistic rating).

By using data of TAVAN2 (the second time TAVAN was offered to students) it was possible to show that all three new methods and the traditional measure of writing skill, the holistic rating of the texts written by the students, measured to a large extent one common factor, because the average correlation between the four variables after correction for attenuation and reversing the first variable showed to be very high (.87).

The factor common to all four different methods can be described as skill in Error Detection and Correction (EDC). While holistic raters might think they are focusing on the content of the texts, their ratings proved to be highly correlated with the number of errors per A4. The correlation for both variables after correction for attenuation was -0.89.

The high correlations we found between the holistic rating, the objectively measured TAVAN-score and the number of errors per A4 (500 words) also explains why objective tests to measure writing skill and holistic rating are highy correlated. Both methods measure mostly skill in EDC (error detection and correction).

To be useful for educational purposes a measure should not only be reliable and valid, there should also be a structured method by which the score can be improved. The TAVAN online program seems to be the first reliable measure for basic writing skill that is also shown to be effective as a means for training.

A second option in this respect seems to use the score for the rating of texts (the quality of the holistic ratings) as a means for (computerized) training. By using items consisting of two sentences, the task for the student would be to select the best sentence. If students really would become better writers by training on this kind of task, this would show that writing can also be learned by close reading. Because reading and writing are highly correlated, such an outcome seems probable.

Short text sufficient?

By splitting up the 48 texts of the students in two halves of equal length (number of words) we found that 250 words were sufficient to rate the number of errors per A4 reliable (topic reliability with perfect reliable raters: 0.97). A short text of about 250 words will therefore be mostly sufficient to measure basic writing skill. Using only one expert rater and half of the total text length we found a reliability of 0.82.

Using holistic rating, however, mostly a number of different texts are needed to reach the same topic reliability (Godshalk et al., 1966). How can this big difference in topic reliability between both methods be explained? Our explanation is that holistic raters are influenced by the content of the writings in a way comparable to the Stroop effect. The content of the text will interfer with the rating of the language use. This effect on the rating proves to be not reliable variance and will, using several texts, cancel out. When raters are instructed to signal (only) the errors in the text, it becomes easier to focus on the language used. The reliability of the ratings therefore increases. The well known topic unreliability of holistic ratings is the only available evidence to assume that basic writing skill of students can vary sharply from topic to topic. The limited topic reliability of holistic ratings is therefore probably due to the method of rating.

Conclusions and limitations

This study focused on two problems: measuring and maximizing basic writing skill of first-year students in higher professional education. We try to summarize the results in 10 conclusions. For each conclusion we try to mention the most important limitation(s).

1. A short text of about 250 words seems mostly to be sufficient to determine the basic writing skill of students in a reliable way by counting the number of errors per A4 (500 words).

This conclusion is based on the high topic reliability (with perfect rater reliability) we found for the number of errors in texts with a length of about 250 words (.97). Using holistic rating a number of different texts might be needed to get the same result. By using only one expert rater the realized score reliability was .82. This result suggests that basic writing skill can be measured rather easily by counting the number of errors: we do not need a big number of expert raters and we do not need a big number of different texts per student. We do not even need a long text, a short text will mostly be sufficient.

Limitation: the high topic reliability we found, is based on two halves of the same topic. To make sure, a replication is needed based on two different topic texts of the same students.

2. In total there are at present five different methods available to measure basic writing skill that will all deliver the same kind of results.

1. the holistic rating by an expert rater of texts written by the student;
2. the score on an objective test to measure basic writing skill;
3. the number of errors per A4 (500 words) in a text written by the student;
4. the score of a student in the TAVAN online program;
5. the quality (score) of a student as a holistic rater.

The first two methods were already known, the last three methods are the result of this study and were shown in this study to have criterion validity (the measure correlates highly with holistic rating) and construct validity (the measure correlates highly with all other available methods to measure basic writing skill).

Limitation: in our research the score for the quality of a student as a holistic rater was based on only 6 rated texts and therefore not very reliable. We would like to see a replication with a more reliable measure for this method. We would also like to see in this replication another objective test involved then TAVAN.

3. All five measures to measure basic writing skill mostly measure the ability to detect and correct errors in text: the EDC-factor (Error-Detection-and-Correction-factor).
This conclusion means that errors play an important role in the way we rate texts, while we often assume to focus mostly on the content of the text. This conclusion also contradicts the idea that errors in a text are insignificant.

Limitation: as far as we can see, none. This conclusion is based on research of Godshalk et al. (1966) and based on our research. All this research was based on texts written by students. It is perhaps possible that for other kind of texts the relationship is less clear.

4. That all five variables to measure basic writing skill measure mostly the EDC-factor, does not yet mean that a second non-EDC-factor could not be possible.
It might be that we never found such a (probably content related) factor, because we assumed holistic rating was this factor, so we never searched any further.

Limitation: at the moment there seems to be no evidence for a content related factor.

5. Errors show to have a very big negative effect on the ratings of readers of the texts.
This result explains why all known measures to measure basic writing skill are so heavily loaded with the EDC-factor. Why people are influenced so much by errors in (the language of) texts is not yet clear. (It might be that readers infer the status of the author by rating the number of errors in the text.)

Limitation: none. This result was found by using an experimental design, there seems to be little room for doubt in this respect.

6. Of the five current methods to measure basic writing skill only (the score of) the TAVAN online program has been shown to be useful as a means to increase basic writing skill.

Limitation: none. This result was found using an experimental design.

7. A second measurement method that might be useful for training purposes could be the quality of the holistic rating by the student.
By using an online program it is possible to present items consisting of two sentences to the students and let them choose the best option. By providing feedback students will quickly

learn to discriminate between good and bad sentences. If this approach results in less errors per A4, this would prove that students can become better writers by just reading closely.

Limitation: at this moment this is just an option. This has to be researched further.

8. The newly developed TAVAN program has been shown to be very effective to increase the basic writing skill of first year students of higher professional education.
The 20 hours of TAVAN resulted in 20% less errors. The size of the effect showed to be 1.1 SD.

This big effect seems to indicate that basic writing skill is indeed a skill that has to be learned by practice. Or as Gandhi put it: "An ounce of practice is worth more than tons of preaching."

Limitation: what is not yet clear at the moment, is the optimal fine tuning of TAVAN, the best way to implement TAVAN and the problems associated with scaling up to a large number of students.

9. Most existing methods used in writing education are probably not effective at all to increase basic writing skill.
The study of errors showed averages of 41 and 81 confirmed errors per A4. Numbers of errors that are only possible if the Dutch educational system is hardly effective in this regard. The rating of existing methods to remediate basic writing level delivered little evidence to think differently. The study about the effectivenss of TAVAN showed that the traditional teaching program did not result in any effect at all on basic writing skill. All in all there seems to be much evidence to assume that traditional methods for teaching writing are hardly effective, while convincing evidence that these methods are effective, seems to be mostly lacking.

If conclusion 8 is true, students need a lot of practice to increase basic writing skill. Nowadays, popular educational methods seem to focus mainly on lecturing and communication of information. If we just tell students what to do, they can do it, seems to be the leading educational principle. Probably, however, few people learned to swim by reading a booklet.

At the same time, providing a class of about 30 students with effective writing practice, is almost impossible without the availability of an online computerprogam like TAVAN.

Limitation: we should not rule out in advance any educational method, but we should insist on convincing evidence that the method is effective.

10. A method like TAVAN is probably also effective for other subjects.

The TAVAN method makes it possible to realize a practice approach (the ABC-model: assignment, answer, feedback) by solving the many practical problems associated with such an approach. Using mainly assignments as a means of communication with students will often lead to a high workload and strict deadlines for the teacher involved. At the same time, feedback to the students will often be slow and too global.

Limitation: in practice it might be not yet that easy to implement an approach like TAVAN. If implementation succeeds, the effectiveness of the method to increase the scores of the students has to be researched.

Future research

We mentioned already some possibilities for future research. We describe three other options for future research.

-- The method of confirmed errors is very time consuming. Also, highly motivated expert raters are needed. TAVAN on the other hand delivers with minimal costs a very reliable score. So, it would be nice if the TAVAN score of a certain item pool and the number of confirmed errors per A4 became related to each other. In that way it is possible to convert that TAVAN score to confirmed errors per A4.

-- On a national level it seems important to monitor the level of basic writing skill.

-- Not every error has the same impact. In the study about the effect of errors the two errorless versions differed in the number of corrected errors, but did not differ significantly in the effect on the readers. It seems important to know why some errors are important, while others seem to be not important at all.

We like to end by repeating an option for research we already mentioned before, because this option seems to open up a window to unexpected new possibilities.

It seems important to have a second method available next to TAVAN by which we not only can measure basic writing skill, but that can also be used to train students and to increase their basic writing skil score. Quality of holistic rating can be modified in such a way that the items can be presented by an online computerprogram like TAVAN for training. By providing direct feedback students will probably learn quickly to discriminate between good and bad sentences. The big question still to be answered is: will these students become better writers, that is, will they make less errors in their texts than before? If so, this would mean that students can become better writers just by reading closely.