2 Facetten van schrijfvaardigheidsmetingen: taken, beoordelingsaspecten en beoordelingsmethoden

R. Schoonen

Inleiding

De operationalisatie van schrijfvaardigheid stelt een onderzoeker vaak voor grote problemen. Van den Bergh, De Glopper en Schoonen (1988) demonstreerden dat, voor een valide en betrouwbare schrijfvaardigheidsmeting vele schrijfopdrachten aan de proefpersoon afgenomen dienen te worden, in sommige gevallen meer dan 20. Ook in een recent literatuuroverzicht van Huot (1990) wordt nog eens onderstreept dat het meten van schrijfvaardigheid middels schrijfopdrachten lastig is door de vele facetten waarmee men rekening dient te houden. Schrijfprestaties op schrijfopdrachten zijn vaak dermate specifiek dat ze weinig lijken te zeggen over de kwaliteit van een volgende schrijfprestatie van dezelfde proefpersoon op een nieuwe schrijfopdracht. Deze geringe generaliseerbaarheid van scores maakt schrijfvaardigheidsmetingen met (open) schrijfopdrachten problematisch, de aantrekkelijkheid van schrijfopdrachten ten spijt. Of zoals Cooper (1984) opmerkt: onderzoekers die beweren dat open schrijfopdrachten vanwege hun validiteit te allen tijde de voorkeur verdienen (boven bij voorbeeld een meerkeuzetoets) stellen ten onrechte validiteit zonder meer gelijk aan 'face validity'. Schrijfvaardigheidsscores ontleend aan door schrijvers gemaakte schrijfopdrachten zijn niet zo vanzelfsprekend als ze wel lijken.

De beperkte generaliseerbaarheid van schrijfvaardigheidsscores blijkt ook als men verschillende aspecten van een schrijfprestatie, bij voorbeeld Taalgebruik en Inhoud & organisatie, apart scoort. De verschillende aspecten bij een opdracht vertonen vaak een grotere onderlinge correlatie dan de scores voor hetzelfde aspect, maar gemeten aan verschillende opdrachten. Men spreekt in dergelijke gevallen van taakeffecten.

Een deel van de problemen in de schrijfvaardigheidsmeting lijkt veroorzaakt te worden door de schrijfopdracht die men de schrijver voorlegt, maar een ander deel van de problemen heeft te maken met het feit dat de schrijfprestaties die uitgelokt zijn met de schrijfopdrachten beoordeeld moeten worden. Verschillende beoordelingsmethoden, zoals globale, analytische en schaaloordelen en het gebruik van strikte scoringsvoorschriften kunnen tot verschillende waarderingen van de schrijfprestaties leiden (vgl. Schoonen, 1991).

Deze constateringen van problemen zijn niet geheel nieuw (vgl. Godshalk, Swineford & Coffman, 1966; Wesdorp, 1974), maar er lijkt de laatste jaren een verschuiving plaats te vinden van de onderzoeksaandacht van de beoordelingsproblematiek naar de problemen van de schrijfopdrachten, waarbij het gaat om opdrachtkarakteristieken als onderwerp, retorische specificatie in de opdracht en dergelijke.

In deze bijdrage wordt op basis van empirische data ingegaan op enkele facetten van de schrijfvaardigheidsmetingen. Op basis van een aantal schrijfvaardigheidsmetingen

met betrekkelijk open opdrachten wordt getracht de relatieve invloed van *taak (opdracht)*, *beoordelingsmethode* en *aspect van beoordeling* te beschrijven. Ter illustratie wordt gedemonstreerd hoe een meerkeuze-schrijfvaardigheidstoets zich gedraagt ten opzichte van de schrijfopdrachten.

Onderzoeksopzet

De data waaraan we de effecten van de diverse facetten van de schrijfvaardigheidsmetingen evalueren, zijn verzameld in een onderzoek waarin de validiteit van alternatieve operationalisaties van schrijfvaardigheid onderzocht werd (Schoonen, 1991). In dat onderzoek figureerde een controlegroep van leerlingen die vier schrijfopdrachten en twee meerkeuze-toetsen maakten.

De schrijfopdrachten zijn iets gewijzigde versies van opdrachten zoals die in de Voorstudie PPON gebruikt zijn (vgl. Wesdorp, et al., 1986). In die opdrachten kregen de leerlingen een communicatieve situatie voorgelegd waarin het doel en de beoogde lezer(s) van hun schrijfprodukt geëxpliciteerd worden. Tevens bood de opdracht relevante informatie voor de inhoud van de te schrijven tekst. De leerlingen maakten de volgende vier opdrachten (vgl. Schoonen, 1991):

naam	*taalhandeling*	*omschrijving*
PELS	overtuigen	het schrijven van een betogend stukje voor de schoolkrant om de eigen mening te uiten over het doden van dieren om hun pels;
KAMP	beschrijven	het beschrijven van concrete zaken voor de schoolkrant, i.c. een kampeerboerderij;
DOOL	uitleggen	uitleggen in de schoolkrant hoe je een spel speelt, i.c. een doolhofpuzzel maakt;
ROUTE	uitleggen	uitleggen in een brief aan een volwassene hoe zij van het busstation naar de kampeerboerderij kan lopen.

Zesentachtig leerlingen uit groep 8 van de basisschool maakten deze vier opdrachten. De leerlingen waren aselect gekozen uit de totale onderzoekssteekproef van 442 en afkomstig van 22 verschillende scholen.

De aldus verkregen schrijfprestaties werden tweemaal beoordeeld op Inhoud & organisatie en tweemaal op Taalgebruik. Eenmaal werd een score gegeven aan de hand van geschaalde voorbeeld-opstellen en eenmaal aan de hand van scoringsvoorschriften. Bij schaalbeoordeling dient de beoordelaar het te beoordelen opstel te vergelijken met een reeks in kwaliteit oplopende opstellen. De opstellen in de reeks zijn voorzien van scores. Door de vergelijking van de kwaliteit van het te beoordelen opstel met die van de opstellen in de reeks moet de beoordelaar een score toekennen (vgl. Wesdorp et al., 1986; Schoonen, 1991). Bij de strikte scoringsvoorschriften moet men denken aan het tellen van vooraf gespecificeerde inhoudselementen respectievelijk taal(gebruiks)fouten.

Zo beschikten we voor elke leerling over zestien scores voor het maken van schrijfopdrachten: 4 (taken) x 2 (beoordelingsaspecten) x 2 (beoordelingsmethoden)

= 16 scores. Elke score was een *juryoordeel*, het gemiddelde van vijf onafhankelijk werkende beoordelaars. De jurybetrouwbaarheid varieerde van .86 tot .96. Tabel 1 geeft nog eens de niveaus van de facetten van de schrijfvaardigheidsmetingen weer.

Tabel 1: Facetten van schrijfvaardigheidsmetingen: taken, beoordelingsaspecten en beoordelingsmethoden.

Taak (T)	Aspect (A)	Beoordelingsmethode (B)
1. Pels		
2. Kamp	1. Inhoud & organisatie (I&o)	1. voorbeeldopstellen (vbb)
3. Dool	2. Taalgebruik (Tgb)	2. scoringsvoorschrift (sco)
4. Route		

In zekere zin hebben we te maken met zestien operationalisaties van schrijfvaardigheid. Dat zou betekenen dat alle 16 scores te herleiden zijn op één onderliggende factor, namelijk schrijfvaardigheid. Als echter het facet van de beoordelingsaspecten van belang is, zal het beoordeelde aspect (Inhoud & organisatie versus Taalgebruik) zijn eigen typische invloed op de scores hebben. Dit zou betekenen dat er naast de schrijfvaardigheidsfactor ook een Inhoud & organisatie- en een Taalgebruikfactor verondersteld moet worden (het *aspectfacet*).

Een zelfde redenering geldt voor het facet van de taken c.q. communicatieve situaties (Pels, Kamp, Dool en Route). Men kan zich voorstellen dat elke taak of communicatieve situatie specifieke eisen stelt aan de schrijver. In dat geval moet men voor elke taak een aparte factor veronderstellen (het *taakfacet*). Scores ontleend aan de Pels-opdracht zijn dan niet zonder meer vergelijkbaar met scores voor bij voorbeeld de Route-opdracht.

Hetzelfde geldt ten slotte voor de beoordelingsmethoden (geschaalde voorbeeldopstellen versus scoring). Elke beoordelingsmethode zou specifieke kenmerken van de tekst kunnen benadrukken en/of bepaalde voorkeuren of karakteristieken van de jury's in de scores kunnen doen gelden. Men zou dan twee beoordelingsfactoren moeten veronderstellen (het *beoordelingsmethode-facet*). Scores aan de hand van de voorbeeldschalen hebben dan een (gedeeltelijk) andere betekenis dan die aan de hand van de strikte scoringsvoorschriften.

Men kan zelfs zover gaan dat men interacties tussen de verschillende facetten veronderstelt. Bij voorbeeld: een interactie tussen beoordelingsaspect en beoordelingsmethode betekent dat eventuele effecten van de beoordelingsmethoden voor Inhoud & organisatiescores anders zijn dan voor Taalgebruiksscores. De twee beoordelingsmethoden kunnen voor Inhoud & organisatie bij voorbeeld wèl en voor Taalgebruik niet dezelfde tekstkenmerken waarderen.

Mellenbergh, et al., (1979) hebben laten zien hoe een dergelijk zogeheten facetdesign als een lineair model geanalyseerd kan worden. Een schrijfvaardigheidsscore Y is dan te beschrijven als de resultante van verschillende invloeden.

$$Y = S + A + B + T + AB + AT + BT + ABT + E,$$

waarbij S de algemene schrijfvaardigheid is; A, B en T zijn respectievelijk het aspect, de beoordelingsmethode en de taak. Lettercombinaties zijn de interacties tussen de

facetten en E is de onverklaarde variantie in de meting die als ruis (error) bestempeld wordt.

Om de grootte van al deze termen in de vergelijking te kunnen bepalen zou men leerlingen meerdere malen dezelfde schrijfopdrachten moeten laten maken. Wij beschikken slechts over een schrijfprestatie per schrijfopdracht zodat de interactietermen niet te onderscheiden zijn van ruis en hier samenvallen met de errorterm (E). Dit vereenvoudigt het model aanzienlijk.

$$Y = S + A + B + T + E',$$

waarbij E' de aangepaste errorterm is.

Het is uiteraard nog maar de vraag of alle facetten relevant zijn voor de beschrijving van de schrijfvaardigheidsscore Y. Het kan zijn dat de genoemde facetten niet van belang zijn in de operationalisatie van schrijfvaardigheid en dat alle operationalisaties een uiting zijn van alleen de algemene schrijfvaardigheid (Y=S+E'), of dat slechts een enkel facet van belang is, bij voorbeeld het type taak dat men maakt (Y=S+T+E'). Zo zijn er verschillende combinaties van termen denkbaar:

(1) $Y = S + A + B + T + E'$
(2) $Y = S + A + B + E'$
(3) $Y = S + A + T + E'$
(4) $Y = S + B + T + E'$
(5) $Y = S + A + E'$
(6) $Y = S + B + E'$
(7) $Y = S + T + E'$
(8) $Y = S + E'$.

In enkele modelpassingen zijn we nagegaan welk model het best bij onze data past: het meest zuinige model (8), waarin alle scores uitingen zijn van alleen schrijfvaardigheid, het meest complexe model (1), waarin de scores behalve van de schrijfvaardigheid afhankelijk zijn van de taak, het beoordeelde aspect en de beoordelingswijze, of een model tussen de twee uitersten.

De modellen kunnen vergeleken worden middels een toetsende factor-analyse, waarin de verschillende niveaus van de facetten als factor (kunnen) optreden. In tabel 2 wordt geïllustreerd wat het ladingenpatroon is voor het meest complexe model dat wij onderzochten. Eenvoudigere modellen kunnen hieruit afgeleid worden door geen ladingen toe te staan op factoren van facetten die men wil uitsluiten.

Tabel 2: Patroonmatrix voor het complexe model (1). *Vbb* is beoordeling met geschaalde voorbeeldopstellen, *sco* is beoordeling met scoringsvoorschriften, *I&o* is het aspect Inhoud & organisatie, *Tgb* is het aspect Taalgebruik en *S* is algemene schrijfvaardigheid.

Score	S	Pels	Kamp	Dool	Route	I&o	Tgb	Vbb	Sco
Pels vbb I&o	*	*							
Pels vbb Tgb	*	*				*		*	
Pels sco I&o	*	*					*	*	
Pels sco Tgb	*	*				*			*
Kamp vbb I&o	*		*				*		*
Kamp vbb Tgb	*		*			*		*	
Kamp sco I&o	*		*				*	*	
Kamp sco Tgb	*		*			*			*
Dool vbb I&o	*			*			*		*
Dool vbb Tgb	*			*		*		*	
Dool sco I&o	*			*			*	*	
Dool sco Tgb	*			*		*			*
Route vbb I&o	*				*		*		*
Route vbb Tgb	*				*	*		*	
Route sco I&o	*				*		*	*	
Route sco Tgb	*				*	*			*
					*		*		*

* = toegestane factorlading

Van het meest zuinige model tot het complexere model zijn niet alleen varianten denkbaar wat betreft de factoren die in het model verondersteld worden, maar ook wat betreft de correlaties die men tussen de factoren toestaat (Mellenbergh et al., 1979). Omwille van de eenvoud gaan we vooralsnog uit van ongecorreleerde factoren.

De correlatiematrix van de zestien variabelen is geanalyseerd in Lisrel7 (Jöreskog & Sörbom, 1989) volgens de 'generalized least squares' (GLS) schattingsprocedure.

Resultaten

De resultaten van de modelpassing staan in Tabel 3. De passing van de modellen wordt beschreven met twee passingsmaten: 'goodness-of-fit' (gfi) en chi-kwadraat (X^2). Voor de eerste geldt: hoe hoger, hoe beter de passing (maximale waarde 1). De tweede is een maat voor de spanning tussen het model en de data. Hiervoor geldt: hoe lager, hoe beter. Gegeven het aantal vrijheidsgraden (df) kan X^2 als een statistische toetsing van de modellen gebruikt worden. Bij een statistische evaluatie van de modellen dient enige voorzichtigheid betracht te worden, omdat de toetsingsmaat alleen onder bepaalde aannamen juist is (bij voorbeeld een multivariate normaalverdeling van de scores) en hier niet aan alle aannamen voldaan wordt. De modellen worden vergelijkenderwijs bekeken. Het verschil in X^2 tussen twee geneste modellen is op zich weer X^2 verdeeld met een aantal vrijheidsgraden (df) gelijk aan het verschil in vrijheidsgraden tussen de twee modellen.

Tabel 3: De passing van de modellen ter beschrijving van de schrijfvaardigheidsscores. *Df* is het aantal vrijheidsgraden, *p* is de overschrijdingskans voor de verwerping van het model, *gfi* is de 'goodness-of-fit' index.

Model	X^2/df	p	gfi
(8) S	206.24/104	.000	.693
(7) S + T	107.36/ 88	.079	.840
(6) S + B	188.03/ 88	.000	.720
(5) S + A	153.00/ 88	.000	.772
(4) S + T + B	88.96/ 72	.085	.868
(3) S + T + A	63.71/ 72	.746	.905
(1) S + T + A + B	50.39/ 56	.686	.925

Uitgaande van het zuinigste model (8) dat betrekkelijk slecht past, blijkt de toevoeging van het aspect- respectievelijk het taakfacet een duidelijke verbetering van de beschrijving van de prestaties te zijn, in het bijzonder de toevoeging van het taakfacet leidt tot een duidelijke passingsverbetering (model 7). De toevoeging van het beoordelingsmethode-facet geeft betrekkelijk weinig passingsverbetering (model 6 versus 8).

Uitgaande van een model dat algemene schrijfvaardigheid en taakspecifieke variantie veronderstelt (model 7) blijkt de toevoeging van het aspectfacet nog steeds een duidelijke verbetering in de beschrijving van de data (model 3 ten opzichte van 7). Toevoeging van het beoordelingsmethode-facet leidt ook nu niet tot een verbeterde beschrijving (model 4 versus 7).

Het toevoegen van het facet van de beoordelingsmethode aan model 3 (model 1) geeft zoals verwacht kon worden geen verdere verbetering ten opzichte van dit model. Model 3 lijkt de voorkeur te verdienen. In de beschrijving van de scores kunnen we volstaan met een schrijfvaardigheidsfactor met daarbij de aspect- en taakfactoren.

Zoals eerder gesteld kan men nog correlaties tussen de factoren veronderstellen. Als we ons beperken tot correlaties tussen factoren binnen een facet, betekent dat, dat we correlaties tussen de twee beoordelingsaspecten toestaan respectievelijk tussen de vier taakfactoren. Beide uitbreidingen leiden niet tot betere beschrijving van de data ten opzichte van model 3 (X^2/df: 62.74/71 respectievelijk 63.33/66).

Gegeven model (3) kunnen we de variantie-componenten (gekwadrateerde ladingen) berekenen voor de invloed van de verschillende facetten. Hoeveel van de variantie is te herleiden op de S-factor, hoeveel op de taak(T)-factoren en hoeveel op de aspect(A)-factoren. De variantiecomponenten voor de zestien variabelen worden in tabel 4 gerapporteerd[1].

Tabel 4: Variantie-componenten (gekwadrateerde factorladingen) voor de schrijfvaardigheidsscores zoals geschat onder model 3[1]. *S* is algemene schrijfvaardigheid, *I&o* is het aspect Inhoud & organisatie, *Tgb* is het aspect Taalgebruik, *vbb* is beoordeeld met voorbeeldschalen, *sco* is gescoord met strikte scoringsvoorschriften.

Score	S	Pels	Kamp	Dool	Route	I&o	Tgb	Totaal
Pels vbb I&o	.466	.399						
Pels vbb Tgb	.613	.039				.004		.870
Pels sco I&o	.214	.406					.001	.653
Pels sco Tgb	.338	.003				.005		.625
Kamp vbb I&o	.442		.456				.026	.367
Kamp vbb Tgb	.461		.092			.005		.903
Kamp sco I&o	.434		.334				.125	.678
Kamp sco Tgb	.289		.049			.000		.768
Dool vbb I&o	.242			.085			.587	.925
Dool vbb Tgb	.489			.365		.503		.830
Dool sco I&o	.102			.106			.045	.898
Dool sco Tgb	.241			.362		.371		.578
Route vbb I&o	.228				.520		.029	.633
Route vbb Tgb	.719				.102	.048		.795
Route sco I&o	.165				.517		.008	.830
Route sco Tgb	.412				.013	.042		.724
							.099	.524

De factoren verklaren gezamenlijk een redelijke proportie van de variantie in de scores (73%), zodat we mogen aannemen dat de verdeling van de variantie over de factoren inzicht geeft in de interpretatie van de scores.

Idealiter zouden de scores herleid moeten worden op één factor, i.c. de algemene schrijfvaardigheid (S) of eventueel op drie factoren, i.c. de algemene schrijfvaardigheid (S) en de twee beoordeelde tekst-aspecten (I&o en Tgb). Als we de variantie die verklaard wordt door de schrijfvaardigheidsfactor en aspectfactoren opvatten als valide variantie en die, die verklaard wordt door de taakfactoren als invalide variantie, dan blijkt dat de Taalgebruiksscores relatief beter beschreven worden door de algemene-schrijfvaardigheidsfactor (S) dan de Inhoud & organisatiescores. Uitzondering is Kamp-scoring. De Taalgebruiksscores zijn relatief weinig taakspecifiek. Dit is niet zo verwonderlijk als men bedenkt dat een schrijver bij elke schrijfopdracht een andere inhoud moet genereren die telkens in een nieuwe communicatieve context moet figureren, terwijl de schrijver steeds een beroep kan doen op min of meer dezelfde taal(gebruiks)kennis. De (relatief korte) Dool-taak vertoont een enigszins afwijkend patroon.

Voorts valt op dat de aspectfactoren betrekkelijk weinig variantie verklaren. Dit zou gedeeltelijk toegeschreven kunnen worden aan het feit dat het hier om factoren gaat die onderling en met de algemene schrijfvaardigheid ongecorreleerd zijn. Het deel van de variantie dat Inhoud & organisatie- en Taalgebruikscores gemeenschappelijk hebben, is al beschreven in de S-factor.

Discussie

De hier gesignaleerde taakfactoren onderstrepen nog eens de analyses van Van den Bergh et al. (1988). In de literatuur zijn al wel suggesties gedaan om deze taakeffecten te verklaren. Zo zouden de taakeffecten verklaard kunnen worden door de discourse-modus of taalgebruiksfunctie van de tekst. In ons geval gaat het om een argumentatieve, een beschrijvende en twee instruerende teksten. Overigens zou dat betekenen dat de eerdere interpretatie van de taakspecifieke variantie als invalide variantie niet geheel terecht is. De taakspecifieke variantie is alleen invalide voor zover men algemene schrijfvaardigheid pretendeert te meten. De taken kunnen evenwel zeer valide operationalisaties van de vaardigheid 'beschrijven', 'overtuigen' respectievelijk 'uitleggen' zijn. In hoeverre deze interpretatie juist is kunnen wij op basis van onze gegevens niet uitmaken. Opmerkelijk is wel dat taakspecificiteit zich vooral manifesteert in de Inhoud & organisatiescores.

Een andere verklaring (die de voorgaande overigens niet uitsluit) is dat de mate van specificatie van de retorische context in de schrijfopdracht taakeffecten veroorzaakt (Brossell, 1983; Huot, 1990). In ons geval lijkt die verklaring niet waarschijnlijk, want voorzover men die specificaties kan vergelijken zijn de specificaties gelijkwaardig: het doel van de tekst wordt in de opdracht geëxpliciteerd, het publiek is omschreven en de benodigde informatie is uit de opdracht te selecteren.

Een meer technische verklaring is dat er sprake is van gecorreleerde error, d.w.z. dat foutenvariantie binnen taken ten onrechte voor systematische variantie aangezien wordt. Het gaat namelijk steeds om één taak waaraan de scores voor de verschillende aspecten en beoordelingsmethoden ontleend worden (vgl. Werts et al., 1980). Een leerling die toevallig zijn dag niet heeft, zal waarschijnlijk een tekst produceren die op alle aspecten en volgens alle beoordelingsmethoden een lage waardering krijgt, terwijl die leerling bij een volgende taak weer op zijn normale niveau presteert. Dergelijke toevalligheden verhogen (kunstmatig) de geobserveerde correlatie tussen scores binnen een taak ten opzichte van de correlaties tussen taken. Het is niet goed uit te maken in hoeverre gecorreleerde error in onze data een rol speelt. Gedeeltelijk tegen deze verklaring pleit dat alleen de scores voor Inhoud & organisatie taakspecifiek lijken en die voor Taalgebruik niet.

Om de hier genoemde verklaringen voor de taakspecificiteit van schrijfvaardigheidsscores uit te kunnen sluiten dient men meer-experimentele onderzoeksopzetten te kiezen in plaats van de correlationele waarop het onderzoek (incl. onderhavige) gebaseerd is, en/of men dient van leerlingen meerdere prestaties op een taak te verzamelen. Dergelijk onderzoek is ons niet bekend.

Naar aanleiding van deze en andere onderzoeksresultaten (vgl. Van den Bergh et al., 1988) kan men zich afvragen of het gebruik van meerkeuzetoetsen voor de meting van de schrijfvaardigheid toch niet overwogen moet worden (zie ook Cooper, 1984; Van Schooten & De Glopper, 1990). Tegen die achtergrond is het interessant na te gaan hoe twee meerkeuzetoetsen, een voor Inhoud & organisatie en een voor Taalgebruik, zich gedragen ten opzichte van de eerder beschreven factorstructuur van de schrijfopdrachten. Aan de eerdere analyse (model 3) zijn de scores voor de twee meerkeuzetoetsen toegevoegd, waarbij model 3 als een vast gegeven beschouwd werd. Voor de scores op de twee toetsen staan we alleen ladingen toe op de algemene schrijfvaardigheidsfactor (S) en de overeenkomstige aspectfactoren. Het resultaat van deze analyse wordt in Tabel 5 weergegeven.

Tabel 5: Variantie-componenten voor twee meerkeuzetoetsen (gekwadrateerde factorladingen). *S* is algemene schrijfvaardigheid, *I&o* is het aspect Inhoud & organisatie, *Tgb* is het aspect Taalgebruik[2].

Score	S	Pels	Kamp	Dool	Route	I&o	Tgb	Tot.
Inhoud & organisatie	.224	-	-	-	-	.002	-	.226
Taalgebruik	.551	-	-	-	-	-	.021	.572

Tabel 5 laat zien dat de meerkeuze-opdrachten niet wezenlijk onderdoen voor de schrijfopdrachten. De variantie in de Inhoud & organisatiescores wordt voor een relatief klein deel beschreven door de algemene schrijfvaardigheidsfactor tegenover een groter deel van de variantie in de Taalgebruiksscores. De aspectfactoren beschrijven een zeer gering deel van de variantie (vgl. tabel 4). Het enig opvallende verschil tussen de schrijf- en de meerkeuze-opdrachten geldt de Inhoud & organisatiescores. Bij de schrijfopdrachten worden die nog voor een deel door de taakfactoren verklaard, terwijl de Inhoud & organisatiescores voor de meerkeuze-opdracht hier voor een belangrijk deel onverklaard blijven.

Zou men op basis van deze gegevens willen beslissen over het gebruik van meerkeuzetoetsen, dan kan men concluderen dat het gebruik van meerkeuzetoetsen voor de taalgebruiksaspecten van het schrijven voor een belangrijk deel dezelfde informatie oplevert als het gebruik van schrijfopdrachten; het gebruik van meerkeuzetoetsen voor de inhoudelijke en tekstorganisatorische aspecten lijkt voor een belangrijk deel andere informatie op te leveren dan het gebruik van schrijfopdrachten (vgl. Schoonen, 1991).

Noten

1. We hebben geabstraheerd van 'illegale' parameterschattingen (negatieve errorvariantieschattingen). Deze waren in sommige modellen aanwezig, maar steeds binnen acceptabele grenzen, zo ook in model (3). De variantieschattingen zijn gemaakt onder model (3'), waarin twee errortermen op 0 gefixeerd werden. De passing van het model wordt nauwelijks door deze correctie beïnvloed. De passing van model (3') was ^{2}/df: 63.95/ 74, p: .791 en gfi: .905.
2. Het op nemen van een speciale 'taakfactor' voor de meerkeuze-opdrachten in het taakfacet leidde niet tot verbetering van de beschrijving van de data.

Bibliografie

Bergh, H. van den, K. de Glopper & R. Schoonen
1988 Directe metingen van schrijfvaardigheid: validiteit en taakeffecten. In F.H. van Eemeren & R. Grootendorst, *Taalbeheersing* (pp. 370-378). Dordrecht.

Brossell, G.
1983 Rhetorical specification in essay examination topics. *College English*, 45, 165-173.

Cooper, P.L.
1984 *The assessment of writing ability: A review of research*. Princeton, N.J.: Educational Testing Service. (GRE Board Research Report GREB No. 82-15R / ETS Research Report 84-12).

Godshalk, F.I., F. Swineford & W.E. Coffman
1966 *The measurement of writing ability*. New York: College Entrance Examination Board.

Huot, B.
1990 The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, 60, 2, 237-263.

Jöreskog, K.G., & D. Sörbom
1989 *LISREL 7 User's Reference Guide*. Mooresville: Scientific Software, Inc.

Mellenbergh, G.J., H. Kelderman, J.G. Stijlen & E. Zondag
1979 Linear models for the analysis and construction of instruments in a facet design. *Psychological Bulletin*, 86, 4, 766-776.

Schoonen, R.
1991 *De evaluatie van schrijfvaardigheidsmetingen. Een empirische studie naar betrouwbaarheid, validiteit en bruikbaarheid van schrijfvaardigheidsmetingen in de achtste groep van het basisonderwijs*. (Academisch proefschrift). Amsterdam: Universiteit van Amsterdam. (in druk)

Schooten, E. van, & K. de Glopper
1990 De validiteit van meerkeuze-instrumenten voor het meten van schrijfvaardigheid. *Tijdschrift voor Taalbeheersing*, 12, 93-110.

Werts, C.E., H.M. Breland, J. Grandy & D.R. Rock
1980 Using longitudinal data to estimate reliability in the presence of correlated measurement errors. *Educational and Psychological Measurement*, 40, 19-29.

Wesdorp, H.
1974 *Het meten van de produktief-schriftelijke taalvaardigheid*. Amsterdam: Universiteit van Amsterdam/Purmerend (Academisch proefschrift).

Wesdorp, H., H. van den Bergh, D.J. Bos, J.B. Hoeksma, R.J. Oostdam, J. Scheerens & B. Triesscheijn
1986 *De haalbaarheid van periodiek peilingsonderzoek; een voorstudie op het gebied van het taalonderwijs in de lagere school*. Lisse.