Effect van toetsvorm en vraagtype op de moeilijkheid van de afsluitingstoetsen basisvorming
Een toepassing van multiniveau analyse met random kruisclassificatie

H. Kuhlemeier, F. Kleintjes en H. van den Bergh

Inleiding

Toetsontwikkelaars blijken nog nauwelijks in staat te voorspellen hoe goed leerlingen presteren op items en toetsen (Bejar, 1983; Mellenbergh, 1971). Evenmin blijkt het eenvoudig een te moeilijke toets in de gewenste richting bij te stellen (Adams, Carson & Cureton, 1993; Alkema & Huson, 1971; Groen & Moelands, 1989). Mede daardoor zijn grootschalige en arbeidsintensieve proefafnames nodig om de (gewijzigde) moeilijkheid van de items langs empirische weg te bepalen. Bij gebrek aan bruikbare theorie over het effect van toets- en itemkenmerken op de prestaties lijkt het maken van items nog steeds eerder een kunst dan een kunde (Scheuneman & Steinhaus, 1987). Vandaar dat onderzoek naar de relatie tussen toets- en itemkenmerken en de prestaties van leerlingen wenselijk is.

In dit artikel rapporteren wij over een secondaire analyse van de invloed van de toets, de toetsvorm en het vraagtype op de prestaties van leerlingen in de eerste fase van het voortgezet onderwijs. Ten aanzien van de toetsvorm onderscheiden wij 'gewone' schriftelijke toetsen en zogeheten praktijktoetsen. Praktijktoetsen worden gekenmerkt door tenminste enkele van de volgende kenmerken (vgl. Linn & Baker, 1996): de vraag- of probleemstelling is open, de opdracht is realistisch en min of meer authentiek, de opdracht vereist praktisch handelen, de opdracht vereist complexe vaardigheden en hogere denkprocessen, de leerling integreert leerstof uit verschillende vakgebieden en/of de toetsscore berust op observatie van leerlinggedrag en/of beoordeling van leerlingproducten (Sluijter e.a., 1996).

Ten aanzien van het toetskenmerk vraagtype onderscheiden wij gesloten en open vragen. Meerkeuze-items worden nogal eens vereenzelvigd met laag gewaardeerde feitenkennis en open vragen met hoog gewaardeerde hogereorde denkprocessen. Gesloten vragen blijken doorgaans wat gemakkelijker en wat minder betrouwbaar dan vergelijkbare open vragen (o.a. Bennett, Rock & Wang, 1991; Frary, 1985; Kinney & Eurich, 1938; Van den Bergh, 1988).

Behalve van de toetsvorm en het vraagtype zijn de moeilijkheid van toetsen en de prestaties van leerlingen afhankelijk van velerlei andere factoren. De meest belangrijke daarvan is natuurlijk de leerling zelf. Zijn of haar achtergrond, intelligentie, vaardigheid, motivatie en inzet bepalen in hoge mate de hoogte van de toetsscore. Een tweede groep van factoren heeft te maken met de school die de leerling bezoekt en het type opleiding dat hij of zij op die school volgt (Willms, 1992). Vandaar dat wij het effect van toets, toetsvorm en vraagtype in een multiniveau analyse relateren aan dat van de school, het opleidingstype en de leerling.

Het effect van de toetsvorm op de prestaties hoeft niet voor elke school gelijk te zijn. Ten aanzien van de interactie tussen prestaties, toetsvorm en school hebben wij tegengestelde verwachtingen. Enerzijds verwachten wij dat leerlingen uit de 'lagere' opleidingstypen met de praktijktoetsen minder moeite hebben dan hun leeftijdsgenoten uit de 'hogere' opleidingstypen. Leerlingen uit het (i)vbo zullen door hun concreet-praktische instelling mogelijk beter uit de voeten kunnen met opdrachten die praktisch handelen vereisen. Wat ook meespeelt is dat het onderwijzen ervan in het (i)vbo een langere traditie heeft dan in het avo. Anderzijds beogen praktijktoetsen een beroep te doen op complexe vaardigheden en hogere denkprocessen waarbij de leerstof van verschillende vakgebieden geïntegreerd moet worden. Op grond hiervan zouden het juist de leerlingen uit de 'hogere' opleidingstypen zijn voor wie praktijktoetsen relatief goed te doen zijn.

In een onderzoek naar tekstbegrip laat Van den Bergh (1988) zien dat het verschil in de moeilijkheid van open en gesloten vragen mede afhankelijk is van het vaardigheidsniveau van de leerlingen. Vbo-leerlingen bleken naar verhouding meer moeite met open vragen te hebben dan mavo-leerlingen. In dit artikel gaan we tevens na in hoeverre er sprake is van een interactie tussen de gemiddelde prestaties van de school en het vraagtype. Als het effect van gesloten vragen inderdaad afneemt naarmate de school beter presteert, is het vraagtype voor de 'betere' scholen van minder groot belang dan voor de 'minder goede' scholen.

Al met al proberen we een antwoord te geven op de volgende vragen:

1. Wat is het relatieve belang van de toets, de school en de leerling voor de prestaties?
2. Hoe belangrijk zijn de toetsvorm (schriftelijke toetsen en praktijktoetsen) en het vraagtype (gesloten en open vragen) voor de prestaties?
3. In hoeverre varieert het effect van de toetsvorm en het vraagtype van school tot school (en van opleiding tot opleiding)? Zijn er met andere woorden scholen traceerbaar die het relatief 'goed' doen op de praktijktoetsen en andere scholen die met de schriftelijke toetsen juist weer minder moeite hebben? En is de ene school 'gevoeliger' voor de samenstelling van de toets naar vraagtype dan de andere school?
4. Zijn de effecten van de toetsvorm en het vraagtype op hoger presterende scholen/opleidingen even groot als op lager presterende scholen/opleidingen? Is het verschil in moeilijkheid tussen de schriftelijke en praktijktoetsen op scholen met hogere gemiddelde prestaties groter of juist kleiner dan op de lager presterende scholen? En zijn toetsen met veel gesloten vragen op de hoger presterende scholen naar verhouding even goed gemaakt als op de minder hoog presterende scholen?

1 Methode van onderzoek

1.1 Dataset

Gebruik is gemaakt van afnamegegevens van het eerste pakket afsluitingstoetsen basisvorming. De afnames vonden plaats in de schooljaren 1994/95 en 1995/96 bij het zogeheten eerste basisvormingscohort: leerlingen die in 1993 met de basisvorming begonnen. Van de zeventig toetsen zijn er 47 in voldoende mate afgenomen om secundaire analyse te rechtvaardigen. Het betreft 29 schriftelijke toetsen en achttien praktijktoetsen. Het percentage gesloten vragen per toets varieert van nul tot honderd, met een gemiddelde van 28, waarbij de praktijktoetsen alleen open vragen kennen. Van alle toetsen zijn de scores van de leerlingen uitgedrukt op een schaal van nul tot honderd (percentage goed).

In de rapportage aan het Cito gaven de docenten aan welk type opleiding de klas volgde: ivbo, ivbo/vbo, vbo, vbo/avo, mavo, mavo/havo(vwo), havo, havo/vwo of vwo. Om technische redenen[1] zijn alleen de gegevens van de homogene opleidingstypen ivbo, vbo, mavo, havo en vwo gebruikt. Het betreft 931 scholen, 2267 opleidingen (waarvan 294 opleidingen voor het type ivbo, 440 voor vbo, 646 voor mavo, 430 voor havo en 457 voor vwo) en 74.988 leerlingen. Voor verdere gegevens over de afsluitingstoetsen, de dataset, de toetsbetrouwbaarheid en de prestaties van de leerlingen wordt verwezen naar Kuhlemeier, Kleintjes & Kremers (1997).

1.2 Statistische analyse

De gegevens zijn geanalyseerd met behulp van multiniveau analyse (Goldstein, 1995), uitgevoerd met het programma MLwiN (Goldstein, Rasbash, Plewis & Draper, 1998). In de modellering hebben we te maken met zes variabelen die de prestaties van de leerlingen kunnen beïnvloeden. Daarvan zijn er drie random en drie fixed. Toets, school en leerling beschouwen we als random en opleidingstype, toetsvorm en vraagtype als fixed. Toetsen beschouwen we hiermee als uitwisselbaar. Het toetspakket 1994 zien we als een a-selecte steekproef uit de oneindig grote populatie van toetsen die maakbaar zijn bij het domein van de basisvorming. Van belang is tevens dat elke school zijn eigen keuze kon maken uit de 'doos' met afsluitingstoetsen. Voor vrijwel elk vak waren er namelijk meer toetsen ter beschikking gesteld, waarvan de school er per vak één of

meer kon kiezen (Kuhlemeier, Kleintjes & Kremers, 1997). In de schooljaren 1994/95 en 1995/96 waren scholen wettelijk verplicht (vrijwel) elk vak van de basisvorming met tenminste één afsluitingstoets af te sluiten. Afgezien van keuzevakken en non-response zijn vak en school in de dataset dus volledig gekruist. Technisch gesproken zijn leerlingen genest binnen de cellen van de random kruisclassificatie van vak en school. School en toets zijn minder volledig gekruist. Scholen konden immers voor vrijwel elk vak kiezen uit twee à vijf toetsen. Het ontwerp is hier structureel onvolledig.

Tot voor kort konden met behulp van multiniveau analyse alleen zuiver geneste data geanalyseerd worden. Sinds kort is echter de theorie en de programmatuur beschikbaar om random kruisclassificaties in een multiniveau analyse expliciet te modelleren (Goldstein, 1995; Goldstein, Rasbash, Plewis & Draper, 1998; Goldstein & Sammons, 1997; Rasbash & Goldstein, 1994). In onze analyse modelleren we de kruisclassificatie van toets en school. De verschillen tussen leerlingen schatten we op het eerste niveau, waarbij leerlingen genest zijn binnen de cellen van de kruisclassificatie van toets en school op het tweede niveau. Hiermee is de variantie op het tweede niveau de som van de variantie tussen toetsen en de variantie tussen scholen. De totale variantie wordt hiermee opgesplitst in drie delen: verschillen tussen leerlingen (niveau 1), verschillen tussen scholen (niveau 2) en verschillen tussen toetsen (niveau 2).

1.3 Geanalyseerde modellen

Ter beantwoording van de onderzoeksvragen zijn drie modellen gespecificeerd. De eerste onderzoeksvraag wordt beantwoord aan de hand van model 1, de tweede onderzoeksvraag aan de hand van model 2 en de onderzoeksvragen 3 en 4 aan de hand van model 3.

In model 1 zijn leerlingen genest binnen de cellen van de kruisclassificatie van toets en school. De variantie op het tweede niveau is nu de som van de tussen-toetsenvariantie en de tussen-scholenvariantie. De totale variantie wordt opgesplitst in drie componenten: tussen-leerlingen, tussen-scholen en tussen-toetsen. In het fixed part wordt alleen het algemene gemiddelde geschat. De verhouding van de variantiecomponenten geeft informatie over het relatieve belang van de factoren toets, school en leerling voor de prestaties op de afsluitingstoetsen.

In model 2 worden de variabelen toetsvorm en vraagtype als fixed effecten toegevoegd. Voor de toetsvorm is er een dummy die één is als het een praktijktoets betreft en nul als het gaat om een reguliere toets. Voor het vraagtype voegen we het percentage gesloten vragen toe aan de regressievergelijking. Dat percentage varieert van nul voor een toets met uitsluitend open vragen tot honderd voor een toets met alleen gesloten vragen. Het belang van toetsvorm en vraagtype wordt geëvalueerd aan de hand van de grootte van de fixed effecten (ten opzichte van de standaardfout) en het percentage verklaarde tussen-toetsenvariantie.

In model 1 en 2 mocht alleen het gemiddeld prestatieniveau - het intercept - variëren tussen scholen en tussen leerlingen. De gewichten voor de regressie van de prestaties op de toetsvorm en het vraagtype - de hellingshoeken - werden beschouwd als zijnde invariant over scholen en leerlingen. Zoals we in de inleiding uiteenzetten, hoeft dit niet per se het geval te zijn. Vandaar dat we de regressiegewichten voor de toetsvorm en het vraagtype in model 3 laten variëren over scholen en over leerlingen.[2]

1.4 Vergelijking van toetsen

In schooleffectiviteitsonderzoek worden schoolresiduen gebruikt om scholen met elkaar te vergelijken in wat zij toevoegen aan de prestaties van de leerlingen (o.a. Aitkin & Longford, 1986; Goldstein & Spiegelhalter, 1996). Een zelfde procedure kan worden toegepast op de vergelijking tussen toetsen. Ook die kunnen we rangordenen op grond van hetgeen zij toevoegen aan de prestaties boven hetgeen is toe te schrijven aan de school en de leerling. Uiteraard zijn de schattingen van de school- en toetsresiduen niet volledig betrouwbaar. Voor elk residu is er een interval waarbinnen de schatting zich met een bepaalde mate van waarschijnlijkheid bevindt. In deze publicatie maken we gebruik van een door Goldstein en Healy (1995) voorgestelde procedure waarbij de type I fout over alle mogelijke even waarschijnlijke paarsgewijze vergelijkingen .05 be-

draagt. De verschillen tussen de conditionele gemiddelden kunnen direct inzichtelijk worden gemaakt aan de hand van een figuur waarin de residuen en hun betrouwbaarheidsintervallen zijn afgebeeld (zie bijvoorbeeld Figuur 1). De afstand tot de nullijn geeft aan hoezeer het residu voor een toets of school afwijkt van het gemiddelde residu (dat uiteraard gelijk is aan nul). Het lijntje om het gemiddelde representeert het betrouwbaarheidsinterval ter grootte van 1.4 maal de standaardfouten van de residuen. Als de intervallen van twee gemiddelden elkaar niet overlappen, mogen we deze als significant verschillend beschouwen.

2 Resultaten

2.1 Effect van toets, school en leerling

Wat is het relatieve belang van de factoren toets, school en leerling voor de prestaties op de afsluitingstoetsen (eerste onderzoeksvraag)? De uitkomsten van de analyse van het onconditionele model met leerlingen genest binnen de cellen van de kruisclassificatie van toets en school zijn weergegeven in de eerste kolom van Tabel 1 (model 1).

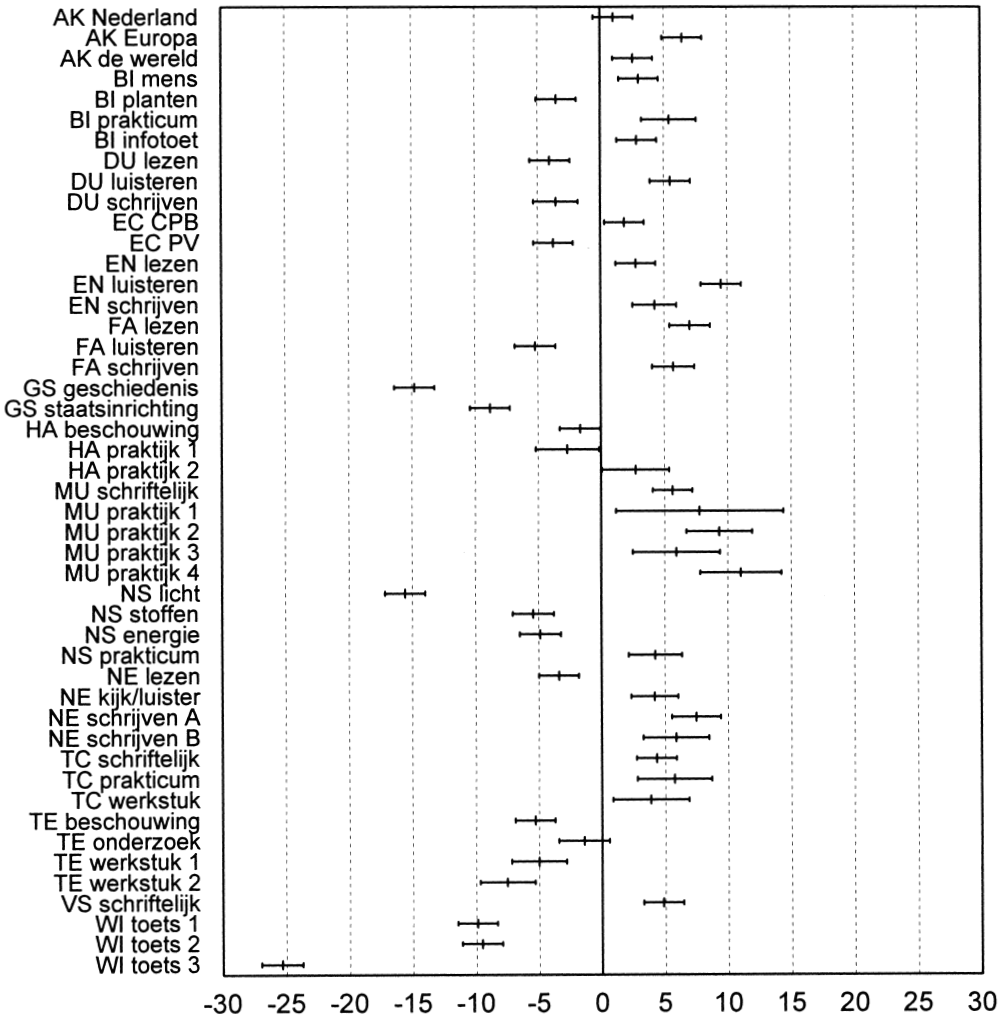
Zoals verwacht trekken de verschillen tussen leerlingen het leeuwendeel van de variantie naar zich toe (282.95). De algemene bijdrage van de school aan de prestaties blijkt aanzienlijk groter dan die van de toets (105.99 versus 54.00)[3]. Van de totale variantie in de prestaties op de afsluitingstoetsen is 12% tussen toetsen, 24% tussen scholen en 64% tussen leerlingen binnen scholen.

In model 1 zijn behalve het algemene gemiddelde geen andere gemiddelden geschat. Toch kunnen we de verschillen tussen individuele toetsen inzichtelijk maken aan de hand van de residuen. Figuur 1 visualiseert de verschillen tussen de 47 toetsen.

Tabel 1

Parameterschattingen voor model 1 (onconditioneel model), model 2 (fixed effect van toetsvorm en vraagtype) en model 3 (fixed en random effect van toetsvorm en vraagvorm) (tussen haakjes: standaardfouten)

	Model 1		Model 2		Model 3	
			Fixed			
Intercept	62.36	(1.14)	57.60	(1.71)	57.58	(1.78)
Praktijktoets			7.96	(2.32)	9.12	(2.55)
Percentage gesloten vragen			.09	(.04)	.09	(.04)
Toets			Random			
Var (intercept)	54.00	(11.45)	42.58	(9.08)	45.81	(9.82)
School						
Var (intercept)	105.99	(5.23)	105.95	(5.23)	34.00	(6.76)
Cov (intercept * praktijktoets)					-67.42	(7.10)
Var (praktijktoets)					108.40	(10.54)
Cov (intercept * perc. gesl. vragen)					-.49	(.04)
Cov (praktijktoets * perc. gesl. vragen)					.34	(.05)
Var (percentage gesl. vragen)					.004	(.0004)
R (intercept * praktijktoets)					-.56	
R (intercept * perc. gesl. vragen)					-.67	
Leerling						
Var (intercept)	282.95	(1.47)	282.95	(1.47)	304.60	(2.70)
Cov (intercept * praktijktoets)					-15.28	(3.73)
Cov (intercept * perc. gesl. vragen)					-1.83	(.08)
Var (percentage gesl. vragen)					.04	(.002)

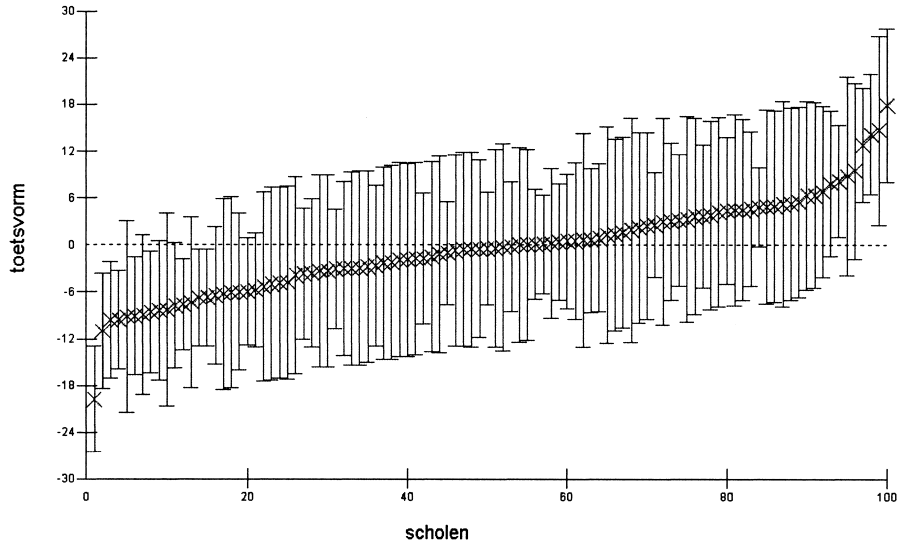


Figuur 1. Toetsresiduen met ± 1.4 se betrouwbaarheidsinterval.

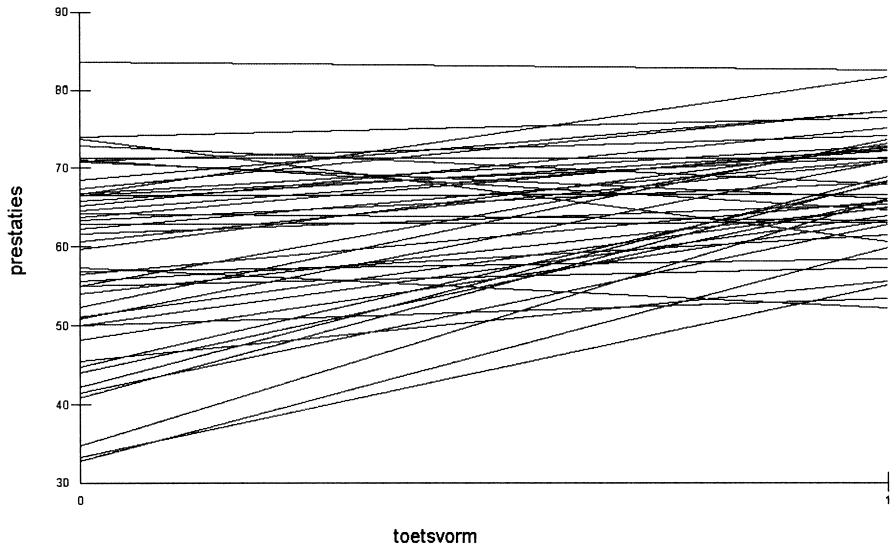
De betrouwbaarheidsintervallen van de toetsresiduen blijken sterk qua grootte te verschillen. Dit hangt deels samen met het verschillend aantal waarnemingen per toets. Zeker voor de praktijktoetsen geldt dat deze aan minder leerlingen zijn voorgelegd dan de schriftelijke toetsen. Er kan alleen een onderscheid worden gemaakt tussen zeer moeilijke en zeer makkelijke toetsen; de overlap tussen de betrouwbaarheidsintervallen is namelijk groot. De derde wiskundetoets is wel goed onderscheidbaar van de overige toetsen. De unieke bijdrage van deze toets aan de prestaties is extreem negatief. Dat wijst erop dat de leerlingen met deze toets erg veel moeite hadden.

2.2 Fixed effect van toetsvorm en vraagtype (model 2)

Hoe belangrijk zijn de toetsvorm en het vraagtype voor de prestaties op de afsluitingstoetsen (tweede onderzoeksvraag)? De tweede kolom van Tabel 1 bevat de uitkomsten van de analyse van het model met toetsvorm en vraagtype in het fixed gedeelte (model 2). Ten gevolge van de toetsvorm en het vraagtype daalt de tussen-toetsenvariantie met 21% (van 54.00 tot 42.58). Op de verschillen tussen scholen en tussen leerlingen zijn er geen noemenswaardige effecten. Hierbij zij aangetekend dat de standaardfout van de tussen-toetsenvariantie vanwege het klei-



Figuur 2. Schoolresiduen voor het regressiegewicht van de toetsvorm met ± 1.4 se betrouwbaarheidsinterval.



Figuur 3. Tussen-scholen regressie van de prestaties op de toetsvorm (0 = schriftelijke toets; 1 = praktijktoets).

ne aantal toetsen erg groot is.

Het regressiegewicht voor de toetsvorm wijkt significant af van nul (op 5%-niveau). Onder constant houding van het percentage gesloten vragen zijn praktijktoetsen in het algemeen beter gemaakt dan de schriftelijke toetsen. Het verschil is bijna acht punten in het voordeel van praktijktoetsen (op een schaal van nul tot honderd).

Het percentage gesloten vragen is eveneens van belang voor de prestaties. Het regressiegewicht bedraagt .09 en verschilt significant van nul. Hoe meer gesloten vragen de toets bevat, hoe beter de toets gemaakt is (gegeven het effect van de toetsvorm). Gemiddeld resulteert het vervangen van tien open vragen door evenzoveel gesloten vragen in een toename van de prestaties met bijna

één punt (gegeven een toets met honderd items en één te behalen punt per item).

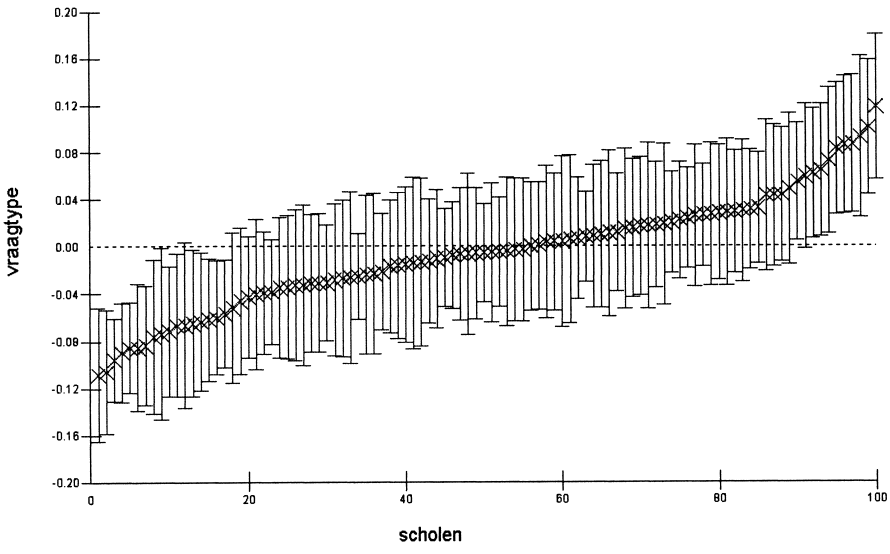
2.3 Random effect van toetsvorm en vraagtype (model 3)

In het derde model mogen de regressiegewichten voor toetsvorm en vraagtype variëren over scholen en over leerlingen (onderzoeksvragen 3 en 4). De derde kolom van Tabel 1 toont de parameterschattingen en hun standaardfouten.

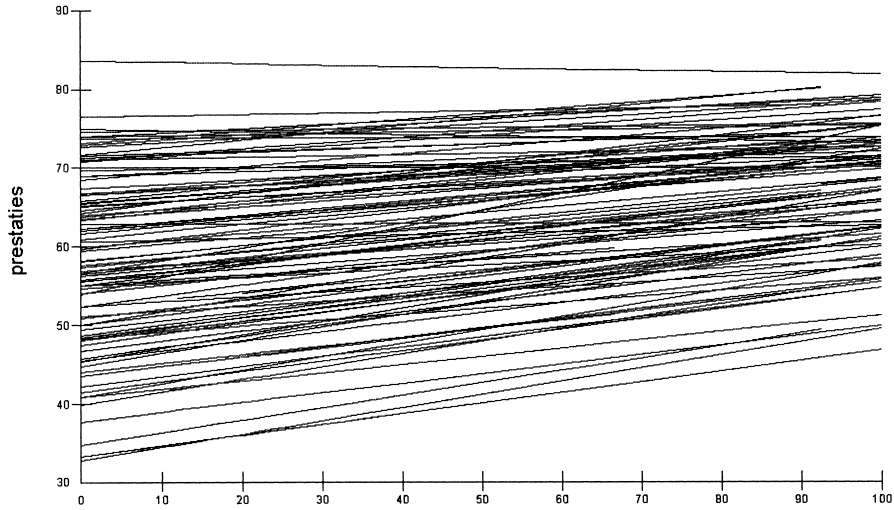
De regressie van de prestaties op de toetsvorm blijkt van school tot school te verschillen (derde onderzoeksvraag). De tussenscholenvariantie bedraagt 108.40 en is significant. Het verschil in moeilijkheid tussen de praktijktoetsen en de reguliere toetsen is op de ene school dus groter dan op de andere school. Van de regressiegewichten van de scholen ligt 90% binnen het bereik van -8.05 en 26.29 ($9.12 \pm 1.65^* \sqrt{108.4}$). Op veel scholen zijn de praktijktoetsen beter gemaakt dan de schriftelijke toetsen, maar op sommige andere scholen is het precies omgekeerd. Figuur 2 toont de schoolresiduen voor een steekproef van honderd van de 931 scholen.

Het verschil in moeilijkheid tussen de beide toetsvormen blijkt op scholen met hoge gemiddelde prestaties kleiner dan op lager presterende scholen (vierde onderzoeksvraag). De covariantie tussen het schoolintercept en het regressiegewicht voor de toetsvorm is namelijk significant en wordt geschat als -67.42 (zie Tabel 1). De overeenkomstige correlatie is -.56. Hoe hoger het prestatieniveau van de school, hoe minder de twee toetsvormen zich qua moeilijkheid van elkaar onderscheiden. Ter illustratie zijn in Figuur 3 de regressielijnen geplot voor een random steekproef van honderd uit alle 931 scholen (waarvan er 47 tenminste één schriftelijke toets en één praktijktoets aan hun leerlingen voorlegden). Figuur 3 laat duidelijk zien dat de regressielijnen voor de lager presterende scholen steiler verlopen dan voor de hoger presterende scholen. Deze figuur laat overigens ook zien dat de schriftelijke toetsen beter discrimineren tussen hoog en laag presterende scholen dan de praktijktoetsen. Voor de schriftelijke toetsen (waarde nul) is de spreiding van de regressielijnen namelijk groter dan voor de praktijktoetsen (waarde één). Voor de praktijktoetsen bedraagt de tussen-scholenvariantie 107.56 [=134.00 + (2*-67.42) + 108.40] versus 134.00 voor de schriftelijke toetsen.

Het gewicht voor de regressie van de prestaties op het percentage gesloten vragen varieert tussen scholen (derde onderzoeks-



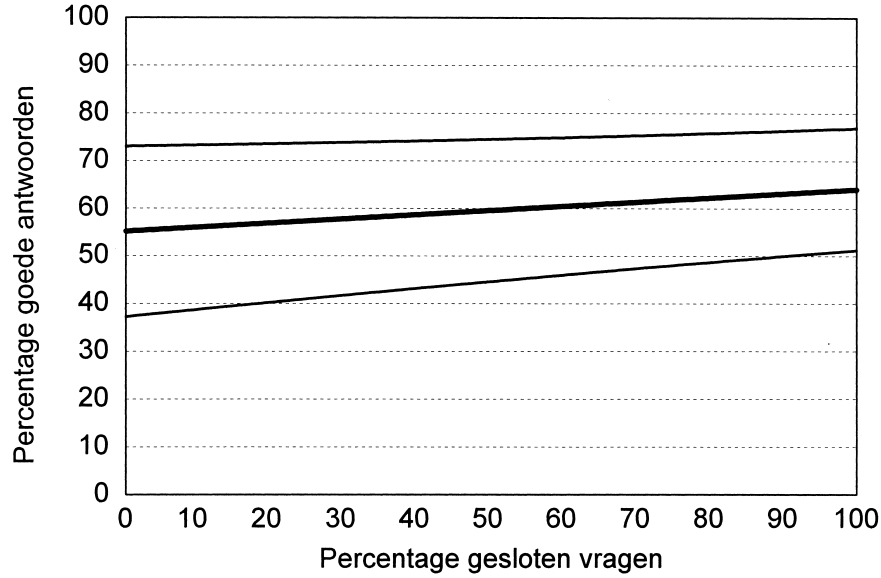
Figuur 4. Schoolresiduen voor het regressiegewicht van het percentage gesloten vragen met ± 1.4 se betrouwbaarheidsinterval.



Figuur 5. Tussen-scholen regressie van de prestaties op het percentage gesloten vragen.

vraag). De tussen-scholenvariantie bedraagt .004 en is significant. Van de regressiegewichten van de scholen ligt 90% tussen de .01 en .19 (.09 ± 1.65 *√.004). Figuur 4 toont de residuen van een random steekproef van honderd van de 931 scholen. Op de ene school is het doorgaans positieve effect van het percentage gesloten vragen groter dan op de andere school.

Op scholen met hoge gemiddelde prestaties lijkt het percentage gesloten vragen van minder groot belang dan op scholen met lage gemiddelde prestaties (vierde onderzoeksvraag). De covariantie tussen het schoolintercept en het regressiegewicht voor het vraagtype is namelijk negatief en significant verschillend van nul (r = -.67). Figuur 5 toont de regressielijnen voor de steekproef van honderd scholen. Voor de lager presterende



Figuur 6. Gemiddeld percentage goed en verschillen tussen scholen als een functie van het percentage gesloten vragen.

Tabel 2

Fixed en random effect van toetsvorm en vraagtype per opleidingstype (tussen haakjes: standaardfouten)

	Opleidingstype									
	IVBO		VBO		MAVO		HAVO		VWO	
					Fixed					
Intercept	31.98	(2.62)	46.15	(2.43)	58.51	(1.96)	67.54	(1.88)	75.51	(1.57)
Praktijk	16.78	(4.71)	13.37	(3.86)	9.08	(2.81)	2.40	(2.75)	-.58	(2.37)
Perc. gesl. vr.	.16	(.06)	.13	(.05)	.09	(.04)	.06	(.04)	.03	(.03)
Toets					Random					
Var(intercept)	99.58	(25.19)	88.35	(20.72)	57.75	(12.65)	52.58	(11.79)	36.75	(8.41)
School										
Var(intercept)	46.91	(5.71)	40.98	(3.83)	31.76	2.58)	22.68	(2.76)	14.59	(1.93)
Cov(int*prakt)	-7.01	(18.82)	-26.31	(9.31)	-23.64	(4.73)	-29.88	(6.08)	-15.61	(5.00)
Var(praktijk)	94.06	(57.54)	145.80	(29.03)	71.86	(12.34)	97.39	(18.48)	113.60	(19.33)
Cov(int*pgesl)	-.24	(.07)	-.23	(.04)	-.19	(.03)	-.20	(.04)	-.12	(.03)
Cov(prak*pgesl)	.06	(.28)	.19	(.12)	.16	(.06)	.32	(.11)	.22	(.10)
Var(perc.gesl.)	.005	(.001)	.004	(.001)	.003	(.001)	.005	(.001)	.003	(.001)
R(int*prak)	-.11		-.34		-.50		-.64		-.38	
R(int*pgesl)	-.51		-.57		-.60		-.62		-.56	
Leerling										
Var(intercept)	217.30	(6.41)	250.50	(4.39)	230.40	(3.63)	197.60	(4.67)	181.50	(4.07)
Cov(int*prakt)	28.37	(25.10)	1.93	(9.21)	11.05	(5.82)	10.85	(6.96)	21.90	(6.60)
Cov(int*pgesl)	-.36	(.21)	-1.32	(.13)	-1.65	(.11)	-1.63	(.14)	-1.55	(.13)
Var(perc.gesl.)	.008	(.005)	.03	(.003)	.038	(.002)	.045	(.003)	.044	(.003)

scholen blijken de regressielijnen inderdaad vaak steiler dan voor de hoger presterende scholen[4].

In model 3 zijn de prestaties uitgedrukt als een kwadratische functie van het regressiegewicht voor het vraagtype (vgl. Van den Bergh & Kuhlemeier, 1997). De tussen-scholenvariantie van het intercept kan derhalve verschillend zijn voor verschillende waarden van het percentage gesloten vragen. Voor elke waarde van het percentage gesloten vragen kan de variantie tussen scholen geschat worden. In dit geval is dat: VAR (tussen scholen)/percentage gesloten vragen = 134.00 + (2 * -.49 * pgesl) + (.04 * pgesl^2). In Figuur 6 zijn het voorspelde percentage goede antwoorden en de prestatieverschillen tussen scholen afgezet tegen het percentage gesloten vragen. De middelste, wat dikkere lijn staat voor het gemiddeld percentage goede antwoorden en met de beide dunnere lijnen is aangegeven binnen welke grenzen zich 90% van de scholen bevindt. De verschillen tussen scholen nemen af naarmate de toets meer gesloten vragen bevat (zie Figuur 6). Voor een toets met honderd procent open vragen bedraagt de tussen-scholenvariantie bijvoorbeeld 163.38 en voor een toets met honderd procent gesloten vragen slechts 84.18. Kennelijk maken toetsen met veel gesloten vragen een minder goed onderscheid tussen scholen dan toetsen met veel open vragen.

2.4 Fixed en random effect van toetsvorm en vraagtype per opleidingstype

De hiervoor geconstateerde fixed en random effecten van toetsvorm en vraagtype zijn aangetoond in de totale responsgroep. Hoog presterende scholen zijn hierbij vooral scholen met havo- en vwo-opleidingen en aan laag presterende scholen zijn vooral ivbo- en vbo-opleidingen verbonden. Dit roept de vraag op naar de generaliseerbaarheid naar de afzonderlijke opleidingstypen. Tabel 2 toont de resultaten van de analyse van model 3 per opleidingstype.

Een eerste algemene constatering is dat de verschillen tussen scholen niet voor elk opleidingstype gelijk zijn. Opvallend zijn de naar verhouding kleine verschillen tussen scholen voor havo- en vwo-opleidingen. Zeker voor havo- en vwo-leerlingen maakt

het voor de prestaties op de afsluitingstoetsen niet zoveel uit aan welke school de havo- of vwo-opleiding verbonden is. Anders gezegd: op grond van de residuen is alleen een onderscheid te maken tussen zeer hoog en zeer laag presterende opleidingen (vgl. Goldstein & Spiegelhalter, 1996). Voor (i)vbo-opleidingen is het belang van de school voor de gemiddelde prestaties van de leerlingen aanzienlijk groter[3].

Een tweede algemene constatering is dat de verschillen tussen toetsen in alle vijf opleidingstypen groter zijn dan die tussen scholen. Hoe een leerling het doet op de afsluitingstoetsen lijkt derhalve sterker afhankelijk van welke toets de docent hem of haar voorlegt dan van de school waaraan de opleiding verbonden is.

Een derde algemene constatering is dat de verschillen tussen toetsen niet voor alle opleidingstypen even groot zijn. In het ivbo en vbo trekt de toets veel meer variantie naar zich toe dan in het havo en vwo. Al met al lijkt het voor havo- en vwo-leerlingen wat minder uit te maken met welke toets de prestaties gemeten zijn en aan welke school de opleiding verbonden is dan voor (i)vbo-leerlingen.

2.5 Toetsvorm per opleidingstype

Eerder zagen we dat praktijktoetsen gemiddeld beter gemaakt zijn dan schriftelijke toetsen (conditioneel op het vraagtype). In de totale responsgroep ging het om een gemiddeld verschil van bijna acht punten (op een schaal van nul tot honderd). Dit gemiddelde verschilt sterk van opleidingstype tot opleidingstype (zie Tabel 2). In het ivbo, vbo, mavo gaat het respectievelijk om zeventien, dertien en negen punten in het voordeel van de praktijktoetsen, maar in het havo en vwo respectievelijk slechts om twee en één punt (waarbij deze laatste twee niet significant afwijken van nul). In de 'lagere' opleidingstypen is het verschil in moeilijkheid tussen de beide toetsvormen dus groter dan in de 'hogere' opleidingstypen.

Met uitzondering van het ivbo doet het random effect van de toetsvorm zich ook voor in de afzonderlijke opleidingstypen (zie Tabel 2). In het vbo, mavo, havo en vwo varieert de regressie van de prestaties op de toetsvorm tussen scholen, terwijl ook de covariantie met het intercept negatief is. Ook binnen de afzonderlijke opleidingstypen is het moeilijkheidsverschil in het voordeel van praktijktoetsen op de ene school groter dan op de andere school. Tegelijkertijd neemt dit verschil af naarmate het gemiddelde prestatieniveau van de opleiding hoger ligt.

Eerder constateerden we dat de schriftelijke toetsen beter discrimineerden tussen scholen dan de praktijktoetsen. In de afzonderlijke opleidingstypen vbo, mavo, havo en vwo zijn het verrassend genoeg juist de praktijktoetsen die een scherper onderscheid tussen scholen en leerlingen maken (alleen in het ivbo is het verschil niet significant). Zo bedraagt de tussen-scholenvariantie in het mavo voor de praktijktoetsen 56.34 versus 31.76 voor de schriftelijke toetsen.

2.6 Vraagtype per opleidingstype

Het fixed effect van het percentage gesloten vragen blijkt in het ene opleidingstype groter dan in het andere (zie Tabel 2). Gegeven een toets van honderd opgaven, met één te verdienen punt per opgave, gaat het vervangen van tien open vragen door tien gesloten vragen in het ivbo gepaard met een toename van de prestaties met 1.6 punt, in het vbo met 1.3 punt en in het mavo met bijna één punt (in het havo en vwo verschilt het regressiegewicht voor het vraagtype niet significant van nul).

Net als in de totale responsgroep hangt de invloed van het percentage gesloten vragen op de prestaties af van de school waaraan de desbetreffende opleiding verbonden is (zie Tabel 2). Zo ligt het regressiegewicht, afhankelijk van de school waaraan de opleiding verbonden is, in 90% van de opleidingen voor ivbo tussen .04 tot .27. In de 5% ivbo-opleidingen met de zwakste regressie nemen de prestaties per tien gesloten vragen toe met hooguit .4 punt en op de 5% ivbo-opleidingen met de sterkste regressie gaat het om tenminste 2.7 punten (gegeven een toets met honderd vragen en één te behalen punt per vraag). In het mavo liggen de regressiegewichten voor 90% van de opleidingen tussen de .00 en .18 en in het vwo tussen -.06 en .12.

Net als in de totale responsgroep zien we in de afzonderlijke opleidingstypen een negatieve samenhang tussen het schoolintercept en het regressiegewicht voor het percentage

gesloten vragen. Ook binnen het ivbo, vbo, mavo, havo en vwo zijn de hellingshoeken minder steil naarmate het gemiddelde prestatieniveau van de school stijgt. Hoe beter de opleiding presteert op de afsluitingstoetsen, hoe minder de samenstelling van de toets naar vraagtype ertoe doet.

3 Discussie

In het onderhavige onderzoek is nagegaan hoe belangrijk de factoren toets, school en leerling zijn voor de prestaties op de afsluitingstoetsen basisvorming (eerste onderzoeksvraag). Geanalyseerd in de totale responsgroep van scholen met opleidingen voor ivbo, vbo, mavo, havo en vwo blijkt de school van groter belang voor de prestaties dan de toets. Van de totale variantie in toetsprestaties bevindt zich 12% tussen toetsen, 24% tussen scholen en 64% tussen leerlingen. De analyse per opleidingstype geeft evenwel een geheel ander beeld te zien. Binnen de opleidingstypen blijkt de school juist van *minder groot* belang dan de toets. Hoe een ivbo-, vbo-, mavo-, havo- of vwo-leerling het doet op de afsluitingstoetsen is meer afhankelijk van welke toets de docent hem of haar voorlegt dan van de school waaraan zijn of haar opleiding verbonden is[3].

Daarnaast is er een interactie met het opleidingstype. Voor (i)vbo-leerlingen zijn de toets en de school van groter belang voor de prestaties dan voor havo- en vwo-leerlingen. Een mogelijke verklaring verwijst naar een grotere differentiatie in het onderwijsaanbod. Mogelijk zijn de 'niveauverschillen' tussen de getoetste onderdelen van de basisvorming in het (i)vbo groter dan in het avo. Wellicht kent het (i)vbo grotere verschillen in de mate waarin de getoetste onderdelen van de basisvorming worden onderwezen dan het avo. En dit hangt mogelijk weer samen met de overladenheid van het curriculum in relatie tot de beschikbare onderwijstijd (Inspectie van het Onderwijs, 1999). Omdat het tempo in het (i)vbo doorgaans wat lager ligt, moeten docenten vaker een keuze uit de vele kerndoelen maken, met als gevolg grotere verschillen in de tijdsbesteding en aandacht voor de getoetste onderdelen van de basisvorming.

De toetsvorm en het vraagtype blijken van belang voor de prestaties (tweede onderzoeksvraag). Gezamenlijk wordt 21% van de verschillen tussen toetsen verklaard. Een hier niet gerapporteerde analyse laat zien dat dit percentage niet in elk opleidingstype gelijk is. In het ivbo, vbo, mavo, havo en vwo gaat het om respectievelijk 34%, 31%, 25%, 10% en 10% van de tussen-toetsenvariantie (vgl. Kuhlemeier, Kleintjes & Van den Bergh, 1999). Voor (i)vbo-leerlingen zijn de toetsvorm en het vraagtype dus van groter belang dan voor havo- en vwo-leerlingen, waarbij het mavo zoals zo vaak een middenpositie inneemt.

3.1 Toetsvorm

Ten aanzien van de interactie van moeilijkheid, toetsvorm en school/opleidingstype wordt onze eerste verwachting bevestigd. Praktijktoetsen blijken gemiddeld beter gemaakt dan schriftelijke toetsen (tweede onderzoeksvraag). Maar voor leerlingen uit de lager presterende scholen en opleidingstypen is het moeilijkheidsverschil ten faveure van de praktijktoetsen groter dan voor leerlingen van hoger presterende scholen en opleidingstypen (derde en vierde onderzoeksvraag). Leerlingen uit de 'lagere' opleidingstypen hebben naar verhouding minder moeite met de praktijktoetsen dan hun leeftijdsgenoten uit de 'hogere' opleidingstypen. Voor dit verschijnsel zijn ten minste twee verklaringen mogelijk.

Een eerste verklaring verwijst naar een reëel verschil in vaardigheid tussen (i)vbo- en avo-leerlingen. Mogelijk is de prestatiekloof tussen beide toetsvormen in het (individueel) beroepsonderwijs kleiner omdat de leerlingen daar wat praktischer ingesteld zijn dan hun leeftijdsgenoten in het algemeen voortgezet onderwijs. Ongetwijfeld zal dit ook samenhangen met een verschil in onderwijsaanbod. Het onderwijzen van praktische vaardigheden kent in het (individueel) beroepsonderwijs immers een langere traditie dan in het algemeen voortgezet onderwijs. Van de andere kant beogen 'performance-based tests', meer dan schriftelijke toetsen, een beroep te doen op complexe vaardigheden en hogere denkprocessen waarbij de leerling leerstof van verschillende vakgebieden

moet integreren (Linn & Baker, 1996). Gezien in dit licht wekt het enige verbazing dat de afgenomen praktijktoetsen gemiddeld beter maakbaar bleken dan de schriftelijke toetsen. Wat een rol kan spelen is dat het meten van complexe vaardigheden van hoger cognitief niveau niet is voorbehouden aan praktijktoetsen. Ook met de schriftelijke afsluitingstoetsen is het mogelijk deze te toetsen. Voorbeelden hiervan in de afsluitingstoetsen basisvorming zijn onder meer 'informatie in verschillende gegevensbestanden opzoeken, selecteren, verzamelen en ordenen', 'rekenvaardigheden toepassen (hoofdrekenen, rekenregels gebruiken, meten en schatten)', 'informatie beoordelen (op betrouwbaarheid, representativiteit en bruikbaarheid), verwerken en benutten', 'op doordachte wijze keuzeproblemen oplossen', 'op basis van argumenten tot een standpunt komen' en 'verschillen in meningen en opvattingen benoemen en hanteren' (Kuhlemeier, 1998). Mogelijk verschillen de ontwikkelde praktijk- en schriftelijke toetsen minder sterk in de intellectuele eisen die zij aan de leerlingen stellen dan de literatuur doet vermoeden.

Een tweede verklaring verwijst naar de wijze waarop de praktijktoetsen beoordeeld zijn. De praktijktoetsen zijn beoordeeld aan de hand van beoordelingsschema's en beoordelingsschalen. Deze zijn gelijk voor alle opleidingstypen, maar laten de docent toch nogal wat speelruimte. Het ligt voor de hand te veronderstellen dat docenten van lager presterende scholen en opleidingstypen hun leerlingen soepeler beoordeelden dan hun collega's van hoger presterende scholen en opleidingstypen (vgl. Heuves & Kuhlemeier, 1998).

In de totale responsgroep, waarin alle opleidingstypen vertegenwoordigd zijn, blijken de schriftelijke toetsen beter te spreiden tussen scholen dan praktijktoetsen. Geanalyseerd in de afzonderlijke opleidingstypen zijn het juist de praktijktoetsen die beter discrimineren tussen scholen. Een plausibele verklaring voor dit 'flip-flap effect' stelt dat de praktijktoetsen niet minder goed discrimineren tussen scholen, maar wel minder goed tussen opleidingstypen. Waarschijnlijk zijn de praktijktoetsen wat minder gevoelig voor verschillen tussen opleidingstypen in onderwijsaanbod dan de schriftelijke toetsen. Mogelijk zal dit verschil verdwijnen als de basisvorming over een aantal jaren integraal is ingevoerd en de thans nog weinig onderwezen praktische vaardigheden meer voet aan de grond hebben gekregen (vgl. Inspectie van het Onderwijs, 1999).

3.2 Vraagtype

Het vraagtype blijkt van belang voor de prestaties (tweede onderzoeksvraag). Hoe meer meerkeuzevragen een toets bevat, hoe hoger de prestaties. Op scholen en opleidingen met hoge gemiddelde prestaties is het percentage gesloten vragen evenwel van minder groot belang dan op scholen en opleidingen met lage gemiddelde prestaties (derde en vierde onderzoeksvraag). Een mogelijke verklaring verwijst naar een differentiële geneigdheid tot raden bij verschillende groepen leerlingen (vgl. Ben-Shakhar & Sinai, 1991; Gafni & Melamed, 1994). Het is bekend dat de raadkans voor een meerkeuzevraag deels afhankelijk is van het vaardigheidsniveau van de leerlingen. Als de leerlingen de leerstof volledig beheersen, hoeft er niet geraden te worden en is de raadkans nul. Hebben zij daarentegen geen enkele kennis van de leerstof, dan is de raadkans het hoogst (bij de gebruikelijke vierkeuze-items: .25) en het voordeel van gesloten vragen boven open vragen het grootst. Dit kan ook verklaren waarom de scorevariantie op school- en leerlingniveau voor toetsen met veel meerkeuzevragen kleiner is dan voor toetsen met veel open vragen (en waarom toetsen met veel meerkeuzevragen vaak minder betrouwbaar zijn dan toetsen met veel open vragen). Raden vermindert de scorevariantie, met een lagere betrouwbaarheid als mogelijk gevolg (Carter & Crone, 1940; Ebel & Frisbie, 1991).

3.3 Kanttekeningen

De gegevens zijn niet speciaal voor de beantwoording van de onderzoeksvragen verzameld. Gebruik is gemaakt van een bestaand bestand met afnamegegevens van de afsluitingstoetsen in de schooljaren 1994/95 en 1995/96. Dit heeft consequenties voor de interpretatie van de onderzoeksuitkomsten.

Ten eerste heeft aan de constructie van de afsluitingstoetsen geen systematisch gekruist ontwerp ten grondslag gelegen met bijvoorbeeld vakken, toetsvormen en vraagtypen als facetten. Wel bevat het toetspakket zowel praktijk- als schriftelijke toetsen en toetsen met gesloten en open vragen in wisselende samenstelling. Bij de interpretatie van het effect van het percentage gesloten vragen moeten we bijvoorbeeld bedenken dat de inhoud van de items niet over vraagtypen constant gehouden is. Gesloten en open vragen zijn met andere woorden wederzijds exclusieve verzamelingen opgaven die behalve qua vraagvorm kunnen verschillen in onder meer onderwerp en vaardigheid. Dit geldt evenzeer voor de praktijktoetsen en de schriftelijke toetsen. Ook die kunnen sterk verschillen qua inhoud en vaardigheid. In een extra analyse is de relatie tussen beide toetsvormen nader onderzocht. Daarbij is een random kruisclassificatiemodel geanalyseerd met de beide toetsvormen als factoren (waarbij op niveau 1 alleen de varianties van beide toetsvormen worden geschat en op niveau 2 de volledige 2*2 covariantiematrix en de tussen-toetsenvariantie). De analyse brengt aan het licht dat de beide toetsvormen op schoolniveau matig gecorreleerd zijn (r = .56). Dit doet vermoeden dat de twee toetsvormen deels andere kennis en vaardigheden meten.

Ten tweede zij opgemerkt dat het aantal en de aard van de afgenomen toetsen niet in elke school of opleidingstype gelijk is. Anders dan bij de centrale schriftelijke eindexamens mochten de docenten een keuze maken uit de meestal twee à vier afsluitingstoetsen die per vak ter beschikking waren gesteld. Ook hadden scholen een grote vrijheid in het bepalen van het moment van afname. Tevens werden niet alle vakken in alle opleidingstypen in dezelfde mate gegeven/gekozen (denk aan de moderne vreemde talen en de kunstvakken). De samenstelling van de respons verschilt derhalve van vak tot vak en van toets tot toets. Daardoor is niet altijd duidelijk in hoeverre verschillen tussen toetsen zijn toe te schrijven aan verschillen in de intrinsieke moeilijkheid van de toetsen, verschillen in onderwijsaanbod dan wel aan verschillen in de samenstelling van de groepen leerlingen aan wie de toetsen zijn voorgelegd (vgl. Kuhlemeier, Kleintjes & Kremers, 1997; Kuhlemeier, Kleintjes & Van den Bergh, 1999).

Noten

1. Het programma MLwiN bleek onder Windows 95 niet meer dan 64 MB RAM intern geheugen te kunnen aanspreken (mondelinge communicatie John Rasbash). Vandaar dat de analyse is uitgevoerd met alleen de gegevens van de homogene opleidingstypen ivbo, vbo, mavo, havo en vwo.
2. In verband met de lengte van het artikel zijn de formules bij de onderscheiden modellen niet in de tekst opgenomen. Deze zijn evenwel in elektronische vorm opvraagbaar bij de tweede auteur.
3. Terzijde zij opgemerkt dat het hier zogeheten bruto schooleffecten betreft. Er is immers niet gecorrigeerd voor verschillen in de samenstelling van de leerlingbevolking naar beginkennis, sociaal milieu en etniciteit (Willms, 1992). De tussenscholenvariantie representeert derhalve het 'bruto' effect van de school, en niet de 'netto' onderwijseffectiviteit (toegevoegde waarde).
4. In model 3 representeert het intercept de gemiddelde prestatie voor een toets met 28% gesloten vragen. Voor andere waarden van het percentage gesloten vragen kan de tussen-scholenvariantie groter of kleiner zijn. De negatieve covariantie tussen het intercept en het regressiegewicht voor het percentage gesloten vragen betekent strikt genomen alleen dat de covariantie (correlatie) rechts van het intercept - vanaf 28% gesloten vragen - negatief is (vgl. Bosker & Snijders, 1990; Van den Bergh & Kuhlemeier, 1997). Op andere schaalpunten, bijvoorbeeld bij 10% of 90% gesloten vragen, is de variantie van het schoolintercept en derhalve ook de covariantie (en de correlatie) met het regressiegewicht voor het vraagtype anders. Een nadere inspectie van Figuur 5 doet evenwel vermoeden dat de covariantie (correlatie) zowel voor toetsen met weinig als met veel gesloten vragen negatief is, zij het dat de covariantie voor toetsen met veel open vragen wat sterker negatief is dan voor toetsen met veel gesloten vragen.

Literatuur

Adams, R., Carson, J. & Cureton, K. (1993). *Item difficulty adjustment study: GRE verbal discretes* (ETS research report no. RR-92-79). Princeton, NJ: Educational Testing Service.

Aitkin, M. & Longford, N. (1986). Statistical modelling in school effectiveness studies (with discussion). *Journal of Royal Statistical Society, A 149*, 1-43.

Alkema, D. & Huson, A. (1971). *Het verbeteren van meerkeuze-vragen aan de hand van item-indices (Rapport nr. 7)*. Leiden: Bureau Onderzoek van Onderwijs.

Bejar, I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement, 7*, 303-310.

Bennett, R.E., Rock, D.A. & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement, 28*, 1, 77-92.

Ben-Shakhar, G. & Sinai, Y. (1991). Gender differences in multiple-choice tests: the role of differential guessing tendencies. *Journal of Educational Measurement, 28*, 1, 23-35.

Bergh, H. van den (1988*). Examens geëxamineerd*. 's-Gravenhage: Instituut voor Onderzoek van het Onderwijs.

Bergh, H. van den & Kuhlemeier, H. (1997). Multiniveau modellen voor de analyse van leerwinst vergeleken. *Tijdschrift voor Onderwijsresearch, 22*, 2, 54-75.

Bosker, R.J. & Snijders, T.A.B. (1990). Statistische aspecten van multiniveau onderzoek. *Tijdschrift voor Onderwijsresearch, 15,* 317-329.

Carter, H.D. & Crone, A.P. (1940). The reliability of new-type or objective tests in a normal classroom situation. *Journal of Applied Psychology, 24*, 353-368.

Ebel, R.L. & Frisbie, D.A. (1991). *Essentials of educational measurement* (5th edition). Englewood Cliffs, NJ: Prentice Hall.

Frary, R.B. (1985). Multiple-choice versus free-response: A simulation study. *Journal of Educational Measurement, 22*, 21-31.

Gafni, N. & Melamed, E. (1994). Differential tendencies to guess as a function of gender and lingual-cultural reference group. *Studies in Educational Evaluation, 20*, 3, 309-319.

Goldstein, H. (1995). *Multilevel statistical models* (2nd Ed.). London: Edward Arnold.

Goldstein, H. & Healy, M.J.R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society, A. 158*, 175-7.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G. & Healy, M. (1998). *A user's guide to MLwiN*. London: University of London, Institute of Education.

Goldstein, H. & Sammons, P. (1997). The influence of secondary and junior schools on sixteen year examination performance: a cross-classified multilevel analysis. *School Effectiveness and School Improvement, 8*, 2, 219-230.

Goldstein, H. & Spiegelhalter, D.J. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society, A. 159*, 385-443.

Groen, H. & Moelands, H. (1989). *Algemene constructieregels voor het aanbrengen van onderscheid tussen C- en D-examens. Een interimverslag*. Cito: Arnhem.

Heuves, T. & Kuhlemeier, J.B. (1998). Discussievaardigheid in de basisvorming: ontwikkeling en beproeving van een meetinstrument. *Tijdschrift voor Taalbeheersing, 20*, 1, 1-19.

Inspectie van het Onderwijs (1999). *Werk aan de basis. Evaluatie van de basisvorming na vijf jaar*. Utrecht: Tonnaer b.v.

Kinney, C.B. & Eurich, A.C. (1938). A summary of investigations. Comparing different types of tests. *School and Society, 36*, 540-544.

Kuhlemeier, J.B. (1998). *Toetsing van algemene vaardigheden in de afsluitingstoetsen basisvorming*. Arnhem: Instituut voor Toetsontwikkeling.

Kuhlemeier, J.B., Kleintjes, F.G.M. & Van den Bergh, H.H. (1999). *Effect van toets, toetsvorm en vraagtype op de moeilijkheid van de afsluitingstoetsen basisvorming* (Publicaties Voortgezet Onderwijs). Arnhem: Insituut voor Toetsontwikkeling.

Kuhlemeier, J.B., Kremers, E.J.J. & Kleintjes, F.G.M. (1997). *De eerste generatie afsluitingstoetsen: gebruik, betrouwbaarheid en maakbaarheid*. Arnhem: Instituut voor Toetsontwikkeling.

Linn, R.L. & Baker, E.L. (1996). Can performance-based student assessments be psychometrically sound? In J. B. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities, 95th yearbook of the National Society for the Study of Education, Part I* (pp. 84-103). Chicago: University of Chicago Press.

Mellenbergh, G.J. (1971). *Studies in studietoetsen*. Amsterdam: Psychologisch Laboratorium.

Rasbash, J. & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross-classified ran-

dom structures using a multilevel model. *Journal of Educational and Behavioral Statistics, 19*, 4, 337-350.

Scheuneman, J.D. & Steinhaus, K.S. (1987). *A theoretical framework for the study of item difficulty and discrimination* (ETS Research Report no. RR-87-44). Princeton, NJ: Educational Testing Service.

Sluijter, C., Kleintjes, F.G.M., Schalk, H.H., Roosmalen, W. van, Hermans, P.H.L., & Bogaerts, C.A.M.J. (1996). *De constructie van beoordelingsschalen bij afsluitingstoetsen voor de basisvorming*. (Onderzoeksrapporten algemeen voortgezet onderwijs). Arnhem: Instituut voor Toetsontwikkeling.

Willms, J.D. (1992). Monitoring school performance. A guide for educators. London: The Falmer Press.

Manuscript aanvaard: 18 februari 2001

Auteurs

Hans Kuhlemeier is werkzaam als onderwijskundig onderzoeker bij de afdeling Beginfase Voortgezet Onderwijs van het Instituut voor Toetsontwikkeling (Cito).

Frans Kleintjes is werkzaam als methodoloog bij de afdeling Psychomerisch Onderzoek en Kenniscentrum van het Instituut voor Toetsontwikkeling (Cito).

Correspondentieadres: H. Kuhlemeier, CITO, Nieuwe Oeverstraat 50, 6801 MG Arnhem, e-mail: Hans.Kuhlemeier@Citogroep.nl

Abstract

This article presents the results of a cross-classified multilevel analysis on the effect of test type and item format on the achievements of 74.988 students from 2267 tracks within 931 junior secondary schools. It shows that the variation between the 47 tests is substantially larger than that between schools. It also shows that performance-based tests (PBT) in general appear to be easier than traditional paper-and-pencil tests (PPT). However, the achievement gap between PBT and PPT is found to be larger at schools with lower average achievement than it is at schools with higher average achievements. It also demonstrates that multiple-choice questions in general are easier than open-ended questions. However, the gap between both item formats is lager at lower achieving schools than it is at higher achieving schools. Finally, interpretations for these results are provided.