Formatieve lessen uit peilingsonderzoek: de toegevoegde waarde van mixture IRT-modellen

D. Van Nijlen en R. Janssen

Samenvatting

Sinds 2002 worden in Vlaanderen peilingen georganiseerd. Zij vormen een cruciaal element in de kwaliteitscontrole voor het Vlaamse onderwijs. Peilingsonderzoek wordt voornamelijk uitgevoerd om summatief een uitspraak te doen over het onderwijsniveau: hoeveel leerlingen bereiken op het einde van een onderwijsniveau de minimumdoelstellingen voor een bepaald domein? Beleidsmakers en het onderwijsveld willen echter ook meer formatieve lessen uit deze peilingsonderzoeken kunnen trekken. Welke specifieke thema's leveren problemen op voor bepaalde studenten? En wie zijn deze studenten? In deze paper wordt geïllustreerd hoe mixture IRT-modellen een belangrijke aanvulling kunnen vormen op de traditionele rapportering over peilingsonderzoek. Zij bieden de mogelijkheid op een inhoudelijk zinvolle manier subgroepen van leerlingen te onderscheiden en hieruit lessen te trekken over waar er zich specifieke problemen voordoen. Dit wordt geïllustreerd op basis van data uit de Vlaamse peiling wiskunde in de eerste graad secundair onderwijs A-stroom in 2009. Uit de analyses blijkt dat twee groepen leerlingen specifieke problemen ondervinden bij werken met veeltermen. Eén groep ondervindt specifiek problemen bij het werken met merkwaardige producten, een andere groep blijkt enkel basisopgaven met betrekking tot eerstegraadsvergelijkingen enigszins onder de knie te hebben. Telkens gaat het om ongeveer één derde van de leerlingen.

1 Inleiding

Sinds 2002 worden in Vlaanderen onderwijspeilingen georganiseerd en dit zowel in het basisonderwijs als in het secundair onderwijs. Met deze peilingen gaat de Vlaamse overheid na in welke mate een bepaalde leerlingengroep de vooropgestelde minimumdoelstellingen (in Vlaanderen uitgedrukt in de vorm van eindtermen) bereiken voor een bepaald domein. Deze peilingen hebben tot doel de kwaliteit van het Vlaamse onderwijs te evalueren en waar mogelijk aanzetten te geven tot reflectie over verbetering van het onderwijs. Peilingen zijn complementair aan andere vormen van externe kwaliteitscontrole zoals doorlichtingen door de onderwijsinspectie en internationale vergelijkende onderwijsonderzoeken. Net zoals internationale onderzoeken zijn peilingen voornamelijk voorzien om uitspraken te doen op systeemniveau, maar in tegenstelling tot internationale onderzoeken zijn peilingen specifiek gericht op het Vlaamse curriculum. Scholen ontvangen bij deelname aan een peiling wel feedback over de prestaties van hun leerlingen, niet op individueel niveau, maar op het niveau van de school. Het gaat in essentie om een low-stakes toetsafname die geen gevolgen heeft voor de school. De ontvangen feedback kan door de school ingezet worden bij de interne kwaliteitscontrole.

Peilingsonderzoek wordt dus in de eerste plaats uitgevoerd om op systeemniveau een uitspraak te doen over de mate waarin een bepaalde leerlingengroep de vooropgestelde minimumdoelstellingen haalt. Men zou kunnen stellen dat het peilingsonderzoek dus voornamelijk een summatief doel nastreeft. Beleidsmakers en het onderwijsveld willen echter uit het peilingsonderzoek ook lessen kunnen trekken die eerder als formatief beschouwd kunnen worden. Zij willen ook informatie verzamelen die relevant is voor het onderwijs- en het leerproces. Meer concreet willen ze ook antwoorden op vragen als 'Met welke specifieke thema's hebben leerlingen het moeilijk?' en 'Wie zijn de leerlingen die het hier moeilijk mee hebben?'.

De huidige paper zal illustreren hoe het gebruik van mixture IRT-modellen een belangrijke toevoeging kan zijn aan de meer traditionele manier van rapporteren over peilingen. Dit zal gebeuren aan de hand van data

uit een peiling wiskunde op het einde van de eerste graad secundair onderwijs in Vlaanderen die plaatsvond in 2009. In een eerste deel zal wat theoretische achtergrond bij de onderzoeksvraag en over het gebruik van mixture IRT-modellen worden gegeven. Daarna wordt de gehanteerde methodologie beschreven. De gebruikte data worden toegelicht in een derde deel. De resultaten van de analyses worden voorgesteld in een volgend deel. De paper wordt afgesloten met een bespreking van de resultaten waar ook ingegaan wordt op het onderwijskundige belang van het onderzoek.

2 Theoretisch kader

Aansluitend op het summatieve doel van peilingsonderzoek wordt bij de analyse van de gegevens veelal verondersteld dat alle leerlingen de toets beantwoorden volgens hetzelfde onderliggend model (DiBello & Stout, 2007). Alle opgaven hebben voor alle leerlingen dezelfde relatieve moeilijkheid, maar leerlingen kunnen natuurlijk wel verschillen in hun vaardigheid. Meer specifiek wordt veelal gebruik gemaakt van modellen uit de itemresponstheorie (IRT) die uitgaan van een continue en unidimensionele onderliggende vaardigheid. De focus in deze analyses komt dan te liggen op verschillen in de mate van beheersing (kwantitatieve verschillen) en niet op verschillen in de aard van problemen dat leerlingen ondervinden (kwalitatieve verschillen).

Dat betekent dan ook dat deze modellen niet zullen volstaan wanneer specifieke items problemen geven voor subgroepen van leerlingen. In dat geval zal de relatieve moeilijkheid van de opgaven namelijk niet voor alle leerlingen hetzelfde zijn. Mixture IRT-modellen bieden de mogelijkheid om niet enkel te focussen op kwantitatieve verschillen in prestaties (verschillen in mate van beheersing), maar eveneens op zoek te gaan naar kwalitatieve verschillen in prestaties (hebben bepaalde leerlingen specifiek moeite met bepaalde opgaven?). Op deze manier bieden zij de mogelijkheid een antwoord te geven op meer formatieve vragen bij grootschalige toetsafnames zoals een peilingsonderzoek.

Het onderscheid tussen kwantitatieve en kwalitatieve verschillen kan gelinkt worden aan het concept van *differential item functioning* (DIF). DIF verwijst naar de aanwezigheid van verschillende meeteigenschappen van items voor verschillende groepen van leerlingen wanneer de overall (latente) vaardigheid van leerlingen in rekening werd gebracht. Concreet betekent dit dat een item DIF vertoont wanneer twee leerlingen met dezelfde vaardigheid uit twee verschillende groepen (bijvoorbeeld meisjes versus jongens) toch niet dezelfde kans hebben om een opgave juist op te lossen. Ruwe verschillen in prestaties wijzen niet per definitie op DIF. Strikt kwantitatieve verschillen in prestaties op een item zullen niet beschouwd worden als DIF aangezien zij puur verschillen in de mate van beheersing weergeven. Kwalitatieve verschillen tussen groepen, specifieke zwaktes of sterktes in het oplossen van bepaalde items, daarentegen kunnen wel beschouwd worden als DIF.

Doorheen de jaren is men anders gaan kijken naar de theoretische betekenis en de praktische waarde van DIF (Zumbo, 2007). Aanvankelijk werd DIF enkel beschouwd als iets dat de meeteigenschappen van een toets verstoorde. Dat betekende dan ook dat de detectie van DIF-items voornamelijk gebeurde vanuit het doel de test 'op te poetsen' en werd ervoor gekozen DIF-items uit een toets te verwijderen. Later gebeurde er echter een verschuiving waarbij men ook op zoek ging naar verklaringen voor het optreden van DIF voor specifieke items en wat er geleerd kan worden over het antwoordproces en leerproces op basis van de items die DIF vertonen. Dat betekende dan ook dat DIF ook interessant werd vanuit een inhoudelijk standpunt.

In een klassieke DIF-analyse wordt gewerkt met manifeste groepen. Cohen en Bolt (2005) stelden echter vast dat er maar een beperkte samenhang is tussen de manifeste indeling die gebruikt wordt om DIF te onderzoeken en de eigenlijke groep van leerlingen die benadeeld of bevoordeeld wordt. Manifeste groepen zijn vaak helemaal niet zo homogeen op vlak van de onderzochte variabele als verondersteld wordt (De Ayala, Kim, Stapleton, & Dayton, 2002; Samuelsen, 2005). Niet alle meisjes gaan zich bijvoorbeeld

bij het oplossen van een opgave gedragen als een typisch meisje, waardoor de gevonden DIF weinig informatief wordt. Het groepslidmaatschap is vaak maar een zwakke proxy voor die aspecten die eigenlijk relevant zijn voor de onderwijspraktijk (Tatsuoka, Linn, Tastuoka, & Yamamoto, 1988). Op die manier zijn de resultaten op basis van manifeste groepen vaak moeilijk inhoudelijk te interpreteren en is de onderwijskundige relevantie vaak beperkt.

Omwille van deze overwegingen is het vaak interessanter op zoek te gaan naar latente, onderliggende groepen van leerlingen en in een tweede stap na te gaan wie deze leerlingen zijn. Het gebruik van latente groepen in de context van DIF heeft de laatste 10 jaar grote aandacht gekregen (Cohen & Bolt, 2005; De Ayala et al., 2002; Samuelsen, 2005; Webb, Cohen, & Schwanenflugel, 2008). Door mixture IRT-modellen (e.g., Rost, 1990; Mislevy & Verhelst, 1990) te gebruiken kunnen een aantal problemen bij het werken met manifeste groepen omzeild worden. Mixture IRT-modellen maken het mogelijk binnen de leerlingen subgroepen op te sporen die kwalitatieve verschillen vertonen en tegelijkertijd de verschillen binnen een groep te kwantificeren.

3 Methode van onderzoek

3.1 Data

De data komen uit de peiling wiskunde eerste graad secundair onderwijs A-stroom in Vlaanderen in 2009. Omdat het bereiken van de eindtermen getoetst wordt op het einde van een onderwijsniveau betekent dit dat leerlingen uit het tweede jaar secundair onderwijs aan deze peiling deelnamen. In Vlaanderen wordt in de eerste graad secundair onderwijs een algemeen geldend curriculum voorzien voor leerlingen uit de A-stroom. Binnen de A-stroom worden vanaf het tweede jaar wel basisopties onderscheiden om het programma meer te laten aansluiten bij de interesses en kwaliteiten van leerlingen. Voor alle basisopties gelden echter wel dezelfde eindtermen. Voor een beperkte groep van leerlingen die de voorziene eindtermen voor het basisonderwijs niet bereiken wordt in de eerste graad van het secundair onderwijs een remediërend programma ingericht, de B-stroom. Deze groep van leerlingen was niet betrokken in deze peiling omdat de eindtermen A-stroom voor hen niet gelden.

In de peiling kwamen de verschillende domeinen binnen wiskunde die opgenomen zijn in de eindtermen aan bod. In de eerste graad secundair onderwijs hebben de eindtermen wiskunde betrekking op de domeinen getallenleer, algebra en meetkunde. De uiteindelijke peilingstoets bestond uit 10 subsets van items die elk betrekking hadden op een topic uit één van deze drie domeinen. De toets werd afgenomen in een geblokt onvolledig design met vier toetsboekjes die elk vier tot zes subsets van items bevatte. Elke student loste één van deze toetsboekjes op. In totaal werd voor het oplossen van de toets 100 minuten voorzien. Binnen elke subset werden de items in dezelfde volgorde gepresenteerd, maar de subsets hadden verschillende posities in de verschillende toetsboekjes. Analyses in de peiling en uitspraken over het al dan niet behalen van de eindtermen gebeurden per subset van items.

In de huidige studie wordt illustratief gefocust op data met betrekking tot twee subsets van items die de eindtermen uit het domein algebra (Tabel 1) dekken. Eén subset bestaat uit 24 zogenaamde kale opgaven met betrekking tot veeltermen, terwijl de andere subset 20 gecontextualiseerde opgaven over probleemoplossen gebruikmakend van algebraïsche uitdrukkingen bevat. Op deze manier vormen deze twee subsets tegelijkertijd een eenheid omdat ze beide betrekking hebben op algebra, maar zijn ze ook een belangrijke aanvulling op elkaar. Bij de toetsafname was het de leerlingen toegestaan een rekenmachine te gebruiken. Leerlingen mochten geen informatieblad met formules gebruiken, zoals soms bij wiskundetoetsen wel gebeurt. De analyses gebeurden op de resultaten van 1567 leerlingen uit 91 scholen. Door middel van achtergrondvragenlijsten die zowel bij leerlingen, ouders als leerkrachten werden afgenomen werd ook nog verdere informatie over de leerlingen en de klaspraktijk verzameld.

De ontwikkeling van de items voor elke subset van items gebeurde op basis van een

Tabel 1

Eindtermen algebra – eerste graad secundair onderwijs

Eindterm	Beschrijving
	Leerlingen
18	gebruiken letters als middel om te veralgemenen en als onbekenden.
19	kunnen twee- en drietermen optellen en vermenigvuldigen en het resultaat vereenvoudigen.
20	kennen de formules voor de volgende merkwaardige produkten: $(a+b)^2$ en $(a+b)(a-b)$; ze kunnen ze verantwoorden en in beide richtingen toepassen.
21	kunnen vergelijkingen van de eerste graad met één onbekende oplossen.
22	kunnen eenvoudige vraagstukken die te herleiden zijn tot een vergelijking van de eerste graad met één onbekende oplossen.
23	ontdekken regelmaat in eenvoudige patronen en schema's en kunnen ze beschrijven met formules.

Rekenen met veeltermen				Verwerkingsniveau
	Aantal items		Aantal items	Reproductief toepassen
ET 19	6	tweeterm + tweeterm	0	
		tweeterm - tweeterm	1	3-19-2
		drieterm + drieterm	1	3-19-4
		drieterm - drieterm	1	3-19-5
		gemengd	1	3-19-7
		tweeterm x tweeterm	1	3-19-17
		tweeterm x drieterm	1	3-19-13
		gemengd	0	
ET 20	6	(a+b)(a-b)	1	3-20-3
		$(a+b)^2$	2	3-20-5 3-20-6
		a^2-b^2	2	3-20-10 3-20-12
		$a^2+2ab+b^2$	1	3-20-15
ET 21	12	x+a=b	3	3-21-4 3-21-12 3-21-20
		ax=b	3	3-21-13 3-21-22 3-21-23
		ax+b=c	2	3-21-6 3-21-14
		gemengd	4	3-21-7 3-21-8 3-21-11 3-21-16
	24		24	

Figuur 1. Toetsmatrijs rekenen met veeltermen

Probleemoplossen gebruikmakend van algebraïsche uitdrukkingen					Verwerkingsniveau	
	Aantal items			Aantal items	Begrips-vorming	Productief toepassen
ET 18	6	letters als onbekende om uitdrukking op te stellen	vorm ax	2	4-18-6	4-18-5
			vorm ax+b	1	4-18-1	
		letters als onbekende om vergelijking op te stellen	vorm ax+b = c	2		4-18-7
						4-18-3
			vorm a(x+b)=c(x+d)	1		4-18-2
ET 22	7	vraagstuk dat leidt tot een vergelijking van de vorm	ax+b=c	1		4-22-11
			ax+b=cx+d	2		4-22-6
						4-22-15
			ax+b(c-x)=e	0		
			(ax+b)+(cx+d)=e	1		4-22-4
			(ax+b)+(cx+d)+(ex+f)=g	3		4-22-3
						4-22-7
						4-22-14
ET 23	7	regelmaat in rij getallen	volgend getal in rij	1		4-23-3
			willekeurig getal in rij	1		4-23-2
			formule opstellen	0		
		regelmaat in figuur	volgende figuur	1		4-23-10
			willekeurige figuur in rij	2		4-23-7
						4-23-11
			formule opstellen	2		4-23-8
						4-23-13
	20			20		

Figuur 2. Toetsmatrijs probleemoplossen gebruikmakend van algebraïsche uitdrukkingen.

toetsmatrijs waar elk item geplaatst werd binnen een (onderdeel van een) eindtermen en binnen een verwerkingsniveau (Bloom, 1956). De onderscheiden verwerkingsniveaus waren feitenkennis, begripsvorming, reproductief toepassen en productief toepassen. De toetsmatrijzen voor de twee opgenomen subsets worden weergegeven in Figuur 1 en Figuur 2. Voor de toets rond rekenen met veeltermen werd binnen elke eindterm nog een meer specifieke inhoudelijke indeling gemaakt die het mogelijk maakte een rijkere inhoudelijke interpretatie aan de resultaten te geven. Alle opgaven met betrekking tot rekenen met veeltermen vallen binnen het verwerkingsniveau reproductief toepassen. Ook binnen de eindtermen voor de subset over probleemoplossen met algebraïsche uitdrukkingen wordt een verdere opdeling gemaakt, voornamelijk naar de vorm van de uitdrukking die gebruikt wordt. Wat betreft het verwerkingsniveau situeren deze items zich voornamelijk op het niveau productief toepassen. Twee items worden geplaatst binnen begripsvorming. Elk item kreeg een uniek nummer toegewezen. Het eerste deel van dit nummer verwijst naar de subset van items waartoe het item behoort. Het middelste element verwijst naar de eindterm waarop het item betrekking heeft. Ten slotte krijgt elk item ook een volgnummer binnen de eindterm.

3.2 Mixture IRT-modellen

Mixture IRT-modellen onderscheiden subgroepen aan de hand van hun antwoordpatroon. Leerlingen worden onderverdeeld in latente klassen en kwantitatieve verschillen in prestatieniveau binnen deze latente klassen worden gemodelleerd aan de hand van klassespecifieke IRT-schalen. In dit geval werd het mixture Rasch model (Rost, 1990) gebruikt.

In het Rasch model wordt de kans om een opgave correct op te lossen (y_{ip}) bepaald door de vaardigheid van een leerling (θ_p) en de moeilijkheid van de opgave (β_i). Als het vaardigheidsniveau van een leerling exact gelijk is aan de moeilijkheid van een item is de kans op een juist antwoord .50. Als de vaardigheid hoger ligt dan de moeilijkheid van een item zal deze kans groter dan .50 zijn. Ligt de vaardigheid lager, dan is deze kans kleiner dan .50. Traditioneel wordt een Rasch model voorgesteld gebruikmakend van een moeilijkheidsparameter, maar het kan volledig equivalent worden voorgesteld met een parameter die de makkelijkheid van een item voorstelt. Deze voorstelling wordt bij de volgende analyses gebruikt.

Het mixture IRT-model breidt het Rasch model (Rasch, 1960) uit met een gewicht voor elke klasse. Dit gewicht geeft de overall kans weer tot een bepaalde klasse te behoren (π_g). Elke leerling krijgt een klassespecifieke vaardigheidsparameter en elke opgave krijgt een klassespecifieke makkelijkheidsparameter toegekend. Formeel krijgt de vaardigheidsparameter en de makkelijkheidsparameter in het mixture Rasch model een extra subscript om aan te geven dat ze kunnen verschillen per latente klasse: θ_{pg} is de vaardigheid van leerling p in klasse g; β_{ig} is een parameter die de makkelijkheid van item i in klasse g weergeeft. Dit komt erop neer dat voor elke latente klasse een specifieke meetschaal opgemaakt wordt.

$$P(y_{ip}=1)=\sum_{g=1}^{G}\pi_g\frac{\exp(\theta_{pg}+\beta_{ig})}{1+\exp(\theta_{pg}+\beta_{ig})} \quad (1)$$

Daarnaast veronderstellen we dat

$$\theta_{pg} \sim N(0,\sigma^2)$$

voor alle latent klassen. Aangezien de klassegewichten moeten sommeren tot één is π_G vastgelegd op

$$1-\sum_{g=1}^{G-1}\pi_g \ .$$

Deze modellen werden geschat met behulp van LatentGOLD (Vermunt & Magdison, 2000). De latente klassen maken het mogelijk kwalitatieve verschillen tussen groepen vast te stellen en tegelijkertijd bieden de IRT-schalen de mogelijkheid voor elke groep de kwantitatieve verschillen in beheersing te modelleren. Voor elke leerling wordt eveneens de kans geschat tot een bepaalde latente klasse te behoren. Eens de latente groepen onderscheiden zijn kan er nagegaan worden welke patroon de itemparameters in de verschillende latente klassen volgen en kan beoordeeld worden waar nu het verschil tussen de groepen juist uit bestaat.

3.3 Koppeling aan achtergrondgegevens

In tweede instantie wordt de opdeling in latente klassen gekoppeld aan beschikbare achtergrondgegevens. Om de hiërarchische structuur van de data (leerlingen binnen klassen binnen scholen) in rekening te brengen wordt hiervoor gebruik gemaakt van multiniveaumodellen (Snijders & Bosker, 2012). Wanneer data een hiërarchische structuur hebben kan niet verondersteld worden dat observaties binnen een groep onafhankelijk van elkaar zijn en dit is een cruciale voorwaarde wanneer klassieke statistische technieken worden toegepast. Leerlingen binnen een klas delen een aantal kenmerken die maken dat de verzamelde observaties niet onafhankelijk zijn (zelfde leerkracht, vaak gelijkaardige familiale achtergrond,...). Wanneer de hiërarchische structuur niet in rekening wordt gebracht leidt dit tot een onderschatting van de standaardfouten van de parameters die de samenhang tussen klasselidmaatschap en de achtergrondkenmerken beschrijven. Bij een onderschatting van de standaardfouten heeft men een verhoogd risico onterecht samenhang te vinden omdat men bij een kleinere standaardfout sneller significantie vindt. Als afhankelijke variabele wordt in deze analyses de voor elke leerling geschatte kans om tot een bepaalde latente klasse te behoren gebruikt. Wanneer meer dan twee latente klassen onderscheiden worden gebeurt deze analyse multivariaat zodat de covariantie tussen de kansen tot een bepaalde latente klasse te behoren mee in rekening gebracht wordt. De kans om tot een bepaalde groep te behoren is namelijk niet onafhankelijk van de kans om tot een andere groep te behoren. Meer specifiek verwacht men een negatieve covariantie omdat een verhoogde

kans tot een bepaalde groep te behoren per definitie samenhangt met een verlaagde kans tot een andere groep te behoren. Wanneer men deze covariantie niet in rekening brengt, vertekent dit de schatting van de parameters en de bijhorende standaardfouten.

4 Resultaten

4.1 Selectie latente klassen

Er werden modellen geschat met één tot en met vijf latente klassen. Daarnaast werd ook een 2PL-model (Birnbaum, 1968) geschat. Dit model biedt de mogelijkheid om na te gaan of eventueel volstaan kan worden met het toelaten van verschillende discriminatiegraden voor items eerder dan het verhogen van het aantal latente klassen. In Tabel 2 worden de resultaten voor deze modellen weergegeven. Voor elk model wordt de loglikelihood samen met het aantal parameters gerapporteerd. Het totaal aantal parameters is gelijk aan het aantal items voor elke latente klas, de variantie van de vaardigheidsparameter (die constant is over de verschillende klassen) en de klassespecifieke gewichten voor de modellen met meer dan één latente klas. Aangezien deze gewichten tot 1 moeten sommeren worden enkel G-1 gewichten geschat. De keuze van het meest geschikte model wordt gebaseerd op het Bayesian Informatie Criterium (BIC; Schwarz, 1978; Nylund, Asparouhov, & Muthén, 2007). Dit informatiecriterium evalueert de geschiktheid van het model door modelcomplexiteit te bestraffen op basis van het aantal parameters en verschaft relatieve informatie over de fit van een model. Het model met de laagste waarde voor het informatiecriterium wordt verkozen.

Op basis van het BIC wordt een oplossing met drie latente klassen verkozen. De drie latente klassen zijn ongeveer even groot: een eerste latente klas bestaat uit 33% van de leerlingen, een tweede uit 35% en de derde uit 32% van de leerlingen.

Tabel 2

Rasch, 2PL en mixture IRT-modellen met een tot vijf latent klassen: loglikelihood (LL), aantal parameters (Npar) en Bayesian Informatie Criterium (BIC)

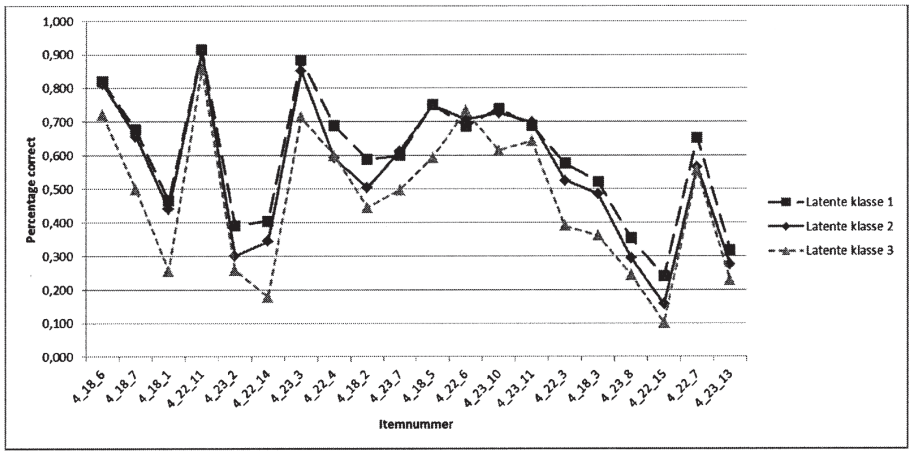
Model	LL	Npar	BIC
Rasch	-38478	45	77287
2PL	-38109	88	76865
Mixed Rasch 2cl	-37822	90	76306
Mixed Rasch 3cl	-37366	135	75724
Mixed Rasch 4cl	-37218	180	75760
Mixed Rasch 5cl	-37097	225	75850

4.2 Interpretatie

Om de opdeling in latente klassen inhoudelijk te interpreteren werden twee benaderingen gebruikt. In de eerste plaats werd voor elke latente klasse nagegaan wat de kans is om een bepaald item juist op te lossen. Aangezien er bij een Rasch-model een één-op-één verband is tussen de proportie juist en de makkelijkheidsparameter uit het model wordt de voorkeur gegeven de resultaten aan de hand van proporties juist weer te geven. Deze zijn namelijk veel eenduidiger te interpreteren voor mensen die niet vertrouwd zijn met IRT-modellen. In tweede instantie zal het lidmaatschap van een bepaalde klasse gekoppeld worden aan bepaalde achtergrondgegevens die vanuit de peilingen beschikbaar zijn. Op deze manier kan verder inzicht verkregen worden in waar nu juist de verschillen tussen de subgroepen liggen (Samuelsen, 2005).

Interpretatie op basis van proporties juist

Voor elke leerling wordt op basis van het antwoordpatroon een kans berekend om tot elk van de drie latente klassen te behoren. Elke leerling werd toegewezen aan een latente klasse op basis van de hoogste van de drie kansen om tot een klasse te behoren. Wanneer de leerlingen zo zijn toegewezen aan een bepaalde latente klasse kan voor elk item per latente klasse berekend worden welke proportie van leerlingen het item juist oplosten. Door deze proporties weer te geven in een figuur kan visueel nagegaan worden of er zich opvallende afwijkingen in de patronen van de moeilijkheid voor de verschillende groepen voordoen. Dit levert ons dus een antwoord op de vraag: zijn er specifieke subgroepen van leerlingen die problemen ondervinden met specifieke opgaven? Door het unieke itemnummer te gebruiken kan makke-

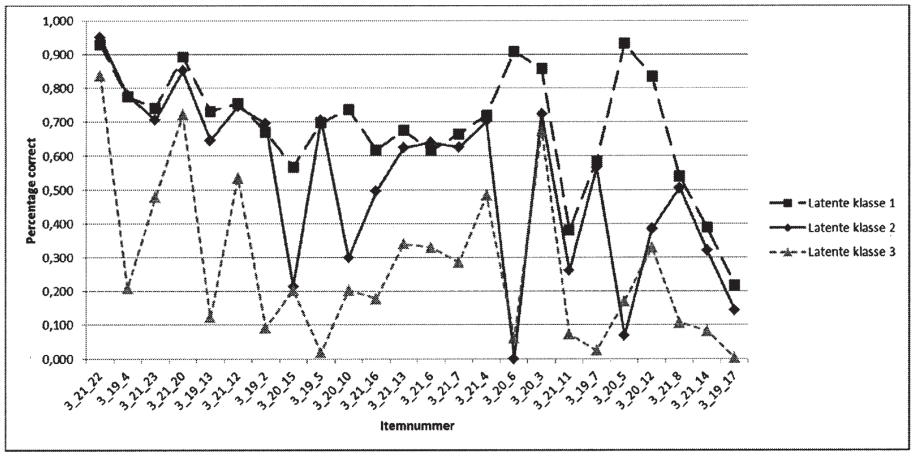


Figuur 3. Grafische illustratie percentage correcte antwoorden per latente klasse – probleemoplossen gebruikmakend van algebraïsche uitdrukkingen

lijk nagegaan worden of het net items met betrekking tot een bepaalde eindterm zijn die een bepaald patroon vertonen.

Globaal genomen werden er geen subgroepen van leerlingen onderscheiden voor de opgaven rond probleemoplossen gebruikmakend van algebraïsche uitdrukkingen (Figuur 3). Er zijn geen opvallende verschillen in prestaties voor specifieke items. Het patroon lijkt algemeen genomen zo te zijn dat de eerste twee latente klassen vrij gelijkaardig presteren met iets betere prestaties voor de eerste groep en dat de derde latente klasse over de hele lijn minder goed presteert. Dit betekent niet dat er geen prestatieverschillen tussen leerlingen waren, maar het patroon in moeilijkheidsgraad voor de opgaven was voor alle leerlingen gelijklopend. Men vindt dus louter kwantitatieve verschillen in de mate van beheersing.

Voor de opgaven rond rekenen met veeltermen kunnen wel duidelijk drie verschillende groepen onderscheiden worden (Figuur 4). De eerste groep presteert algemeen genomen het best op de opgaven uit deze subset en ondervindt geen specifieke problemen met bepaalde typen opgaven, al zijn er natuurlijk ook een aantal opgaven die voor hen vrij lastig zijn. De tweede groep van leerlingen heeft voor de meeste opgaven een gelijkaardig prestatieniveau als de eerste, maar presteren duidelijk minder goed op een aantal opgaven.



Figuur 4. Grafische illustratie percentage correcte antwoorden per latente klasse – rekenen met veeltermen

Wanneer dit resultaat naast de toetsmatrijs gelegd wordt, blijkt het hier te gaan om de opgaven rond merkwaardige producten. De derde groep van leerlingen presteert algemeen genomen duidelijk minder goed en blijkt enkel de basisopgaven rond eerstegraadsvergelijkingen enigszins onder de knie te hebben.

Interpretatie op basis van achtergrondgegevens

In Tabel 3 wordt weergegeven in welke mate een aantal beschikbare achtergrondkenmerken samenhangen met de kans tot een bepaalde groep van leerlingen te behoren. Voor elke leerling wordt een kans berekend om tot de drie latente klassen te behoren. Deze kans wordt in deze analyse als afhankelijke variabele genomen. Dat betekent ook dat de parameters uit deze analyse geïnterpreteerd kunnen worden als kansen. Zo geeft het intercept weer dat de referentieleerling ongeveer 40% kans heeft tot de eerste groep van leerlingen te behoren, een wat lagere kans van 38% om tot de tweede groep te behoren en zo'n 22% kans om tot de derde groep van leerlingen te behoren. De referentieleerling is een leerling die voor alle categorische variabelen in de referentiecategorie zit en voor de continue variabelen steeds gemiddeld scoort. In dit geval is, wanneer we naar de eerste drie variabelen kijken, de referentieleerling dus een jongen die op leeftijd zit en thuis enkel Nederlands spreekt.

Uit de koppeling met de achtergrondgegevens blijkt er een samenhang te zijn tussen het klasselidmaatschap en een aantal leerproblemen. Leerlingen met dyscalculie hebben een duidelijk hogere kans om tot de derde latente klasse te behoren die eigenlijk enkel de basisopgaven met betrekking tot eerstegraadsvergelijkingen onder de knie lijken te hebben. Deze kans ligt bijna 20% hoger dan voor leerlingen zonder dit leerprobleem. De keerzijde hiervan is dat ze een duidelijk lagere kans hebben om tot de eerste latente klasse te behoren, namelijk 23% lager. Leerlingen met dyslexie hebben dan weer een wat hogere kans om tot de tweede latente klasse te behoren. Dit zijn de leerlingen die specifiek problemen ondervinden met merkwaardige producten. Deze verhoogde kans wordt gecompenseerd door een lagere kans om tot de derde latente klasse te behoren.

Ook de studierichting hangt duidelijk samen met klasselidmaatschap. Leerlingen uit technische basisopties hebben een sterk verhoogde kans om te behoren tot de groep van leerlingen die enkel de eenvoudige eerstegraadsvergelijkingen onder de knie hebben. Het gaat hier om een stijging van bijna 30% tegenover een referentieleerling uit moderne wetenschappen. Wanneer we dan in rekening brengen dat een referentieleerling een kans van iets meer dan 20% had om tot deze groep te behoren, zien we dat leerlingen uit technische opties meer dan 50% kans hebben om tot de laatste groep te behoren. Dat betekent enerzijds dat ze een lagere kans hebben om tot de eerste latente klasse te behoren, maar anderzijds ook een verlaagde kans om tot de tweede latente klasse te behoren. Leerlingen uit klassieke talen hebben weinig kans tot de derde groep te behoren en een duidelijk hogere kans om tot de eerste groep te behoren die over de hele lijn vrij goed presteert.

Ten slotte blijken leerlingen uit het gemeenschapsonderwijs een enigszins lagere kans te hebben om tot de eerste latente klasse te behoren. Deze verlaagde kans wordt gelijkmatig gecompenseerd door een lichte (niet-significante) verhoging van de kans om tot de twee andere latente klassen te behoren.

Met de andere variabelen wordt geen samenhang gevonden. Er is geen verschil tussen jongens en meisjes in de kans tot één van de groepen te behoren. We vinden ook geen samenhang met schoolse achterstand. Voor thuistaal blijkt één parameter net significant te zijn, maar wanneer rekening gehouden wordt met meervoudig toetsen blijkt er globaal genomen geen samenhang te zijn met thuistaal. Met SES en het aantal boeken thuis wordt ook geen samenhang gevonden. Ook het schooltype en de score van de school op de onderwijs armoede-indicator hangt niet samen met het klasselidemaatschap.

Vanuit onderwijskundig standpunt kan het interessant zijn om te zien of het klasselidmaatschap ook samenhangt met een aspect uit de klaspraktijk. Vanuit de vragenlijsten was er ook informatie beschikbaar over welk handboek gebruikt werd. Wanneer we geen achtergrondkenmerken in rekening brengen

Tabel 3

Samenhang klasselidmaatschap en achtergrondgegevens

	Klasse 1			Klasse 2			Klasse 3		
Variabele	COEF	SF		COEF	SF		COEF	SF	
Intercept	0.406	0.062		0.376	0.061		0.218	0.050	
Geslacht (meisje)	0.030	0.021		-0.027	0.022		-0.003	0.020	
Schoolse achterstand									
voor	-0.028	0.068		-0.061	0.072		0.089	0.067	
Op leeftijd°									
1 jaar achter	-0.035	0.029		-0.018	0.030		0.053	0.028	
+1 jaar achter	-0.038	0.075		0.103	0.079		-0.064	0.073	
Thuistaal									
Exclusief Nederlands°									
Nederlands met andere taal	0.047	0.031		-0.069	0.033	*	0.022	0.030	
Exclusief andere taal	0.041	0.058		-0.036	0.061		-0.005	0.056	
Aantal boeken thuis									
0-10°									
11-25	-0.013	0.035		0.047	0.037		-0.035	0.035	
26-100	0.028	0.035		-0.014	0.037		-0.014	0.034	
101-200	-0.058	0.038		0.048	0.041		0.010	0.038	
meer dan 200	-0.025	0.042		0.009	0.045		0.015	0.041	
Leerproblemen									
Dyslexie	-0.008	0.036		0.087	0.038	*	-0.079	0.035	*
Dyscalculie	-0.230	0.072	**	0.036	0.077		0.194	0.071	**
AD(H)D	-0.034	0.044		-0.003	0.046		0.037	0.043	
Autismespectrumstoornis	0.148	0.091		-0.089	0.095		-0.059	0.088	
Andere	-0.018	0.036		0.005	0.038		0.013	0.035	
Optiegroep									
Technische opties	-0.180	0.047	***	-0.108	0.047	*	0.288	0.039	***
Klassieke talen	0.180	0.037	***	-0.042	0.037		-0.139	0.032	***
Moderne wetenschappen°									
Overige basisopties	-0.346	0.268		-0.295	0.269		0.641	0.231	**
SES	0.006	0.012		0.010	0.012		-0.016	0.011	
Onderwijs kansarmoede-indicator (schoolniveau)	-0.035	0.047		0.024	0.045		0.011	0.036	
Schooltype									
Autonome middenschool°									
Aso-school	-0.053	0.062		0.021	0.059		0.032	0.045	
Multilaterale school	0.043	0.078		-0.060	0.074		0.017	0.056	
TSO/BSO/KSO-school	0.020	0.068		-0.047	0.065		0.027	0.051	
Officieel gesubsidieerd onderwijs	-0.141	0.061	*	0.061	0.058		0.080	0.045	

° referentiecategorie; *p<.05; **p<.01; ***p<.001

Tabel 4

Samenhang klasselidmaatschap en gebruik handboek zonder achtergrondkenmerken in rekening te brengen

	Klasse 1			Klasse 2			Klasse 3		
Variabele	COEF	SF		COEF	SF		COEF	SF	
Intercept	0.430	0.048		0.290	0.039		0.279	0.043	
Handboek									
1°									
2	-0.193	0.092	*	0.075	0.076		0.118	0.084	
3	-0.220	0.083	**	0.324	0.068	***	-0.104	0.075	
4	-0.167	0.083	*	0.034	0.070		0.133	0.076	
5	-0.299	0.085	***	0.221	0.072	**	0.078	0.078	
6	-0.177	0.104		0.209	0.093	*	-0.031	0.099	
7	0.133	0.085		-0.094	0.069		-0.039	0.076	
8	-0.275	0.084	**	0.008	0.069		0.267	0.076	***
9	-0.290	0.127	*	0.031	0.104		0.259	0.114	*
10	0.102	0.098		-0.041	0.084		-0.060	0.091	
11	-0.223	0.089	*	-0.046	0.078		0.269	0.084	**

° referentiecategorie; *p<.05; **p<.01; ***p<.001

zien we een duidelijke samenhang tussen klasselidmaatschap en de keuze van handboek (Tabel 4). Zo hebben leerlingen die handboek 8 gebruiken een duidelijk lagere kans om tot de eerste groep van leerlingen te behoren, maar een duidelijk hogere kans om tot de derde groep te behoren die enkel de basisopgaven met betrekking tot eerstegraads-

Tabel 5

Samenhang klasselidmaatschap en gebruik handboek rekening houdend met achtergrondkenmerken

	Klasse 1			Klasse 2			Klasse 3		
Variabele	COEF	SF		COEF	SF		COEF	SF	
Intercept	0.458	0.069		0.310	0.065		0.232	0.058	
Handboek									
1°									
2	-0.114	0.084		0.097	0.078		0.017	0.068	
3	-0.235	0.074	**	0.294	0.068	***	-0.060	0.059	
4	-0.021	0.110		-0.022	0.102		0.043	0.090	
5	-0.245	0.100	*	0.154	0.093		0.090	0.081	
6	-0.227	0.096	*	0.205	0.094	*	0.022	0.082	
7	0.104	0.076		-0.120	0.069		0.016	0.060	
8	-0.110	0.084		0.072	0.077		0.039	0.068	
9	-0.145	0.119		0.046	0.107		0.099	0.094	
10	0.094	0.089		-0.027	0.085		-0.068	0.074	
11	-0.135	0.085		0.019	0.085		0.117	0.074	

° referentiecategorie; *p<.05; **p<.01; ***p<.001

vergelijkingen onder de knie hebben. Nu is het echter zo dat specifieke handboeken voor specifieke leerlingengroepen gebruikt worden. Sommige handboeken zullen eerder in technische basisopties gebruikt worden en van die leerlingen weten we dat zij sowieso een grotere kans hebben tot de derde groep van leerlingen te behoren. Daarom is het bij deze analyse cruciaal om eveneens de achtergrondkenmerken in rekening te brengen en dan te evalueren of er nog steeds een samenhang gevonden wordt.

In Tabel 5 worden de resultaten weergegeven wanneer alle bovenvermelde achtergrondkenmerken in rekening worden gebracht. Uit deze analyses blijkt dat een aanzienlijk deel van de samenhang tussen klasselidmaatschap en het handboek verdwijnt. Zo wordt er voor geen enkel handboek nog een samenhang gevonden met de kans tot de derde groep van leerlingen te behoren. Ook zien we geen samenhang meer tussen klasselidmaatschap en het gebruik van handboek 8. Toch blijft er voor een aantal handboeken nog een duidelijke samenhang en telkens vertoont de samenhang een gelijkaardig patroon. De leerlingen die handboek 3, 5 en 6 gebruiken hebben een duidelijk lagere kans om tot de eerste groep leerlingen te behoren die over de hele lijn vrij goed presteren. En voornamelijk voor handboek 3 en 6 zien we dat de keerzijde hiervan is dat deze leerlingen een duidelijk verhoogde kans hebben om tot de groep van leerlingen te behoren die specifiek problemen ondervinden met merkwaardige producten, zonder een verhoogde kans te hebben tot de derde groep te behoren die algemeen genomen vrij zwak presteert.

Aan de leerkrachten werd bij de peiling gevraagd telkens aan te geven of een leerling volgens hen de eindtermen voor wiskunde bereikte. De leerkrachten gaven voor 79% van de leerlingen aan dat zij volgens hen de eindtermen bereikten. We koppelden dit oordeel in een bijkomende multiniveau-analyse ook aan de kans tot een bepaalde latente klasse te behoren. Leerlingen die volgens de leerkrachten de eindtermen niet beheersen hebben 58% kans tot de derde groep van leerlingen te behoren. Daarnaast hebben ze slechts 19% kans om tot de eerste groep te behoren en 23% kans om tot de groep van leerlingen te behoren die specifiek problemen ondervinden met merkwaardige producten. Leerlingen die volgens de leerkrachten wel de eindtermen bereiken hebben een duidelijk lagere kans om tot de laatste groep te behoren die over de hele lijn zwak presteert, namelijk 28%. Over de andere twee groepen worden deze leerlingen gelijkmatig verdeeld, respectievelijk 34 en 38% kans. Uit de samenhang met de oordelen van de leerkrachten over het beheersen van de eindtermen wiskunde blijkt dus dat de groep met specifieke problemen voor merkwaardige producten door leerkrachten zeker niet beschouwd wordt als een zwak presterende groep.

Aangezien deze leerlingen voor andere opgaven wel op het niveau van de eerste groep presteren kan deze analyse een aanzet vormen om na te gaan of het specifieke probleem met merkwaardige producten inderdaad niet toegeschreven kan worden aan de manier waarop dit thema aangebracht wordt in deze handboeken.

5 Conclusie en discussie

Door mixture IRT-modellen te gebruiken bij de analyse van data uit een peiling wiskunde was het mogelijk een opsplitsing te maken in de leerlingen en deze opsplitsing blijkt duidelijk relevant te zijn vanuit onderwijskundig oogpunt. Er blijken twee substantiële groepen van leerlingen te zijn die toch nog problemen ondervinden met specifieke aspecten van het werken met veeltermen. Een derde van de leerlingen ondervindt specifieke problemen bij het werken met merkwaardige producten en een derde van de leerlingen blijkt enkel te kunnen werken met basale eerstegraadsvergelijkingen.

Dit soort resultaten kan een belangrijke aanvulling vormen op de meer traditionele rapportering over peilingsonderzoek waar voornamelijk gekeken wordt hoe de groep in zijn geheel op een bepaald (sub)domein presteert. Ook op vlak van feedback op schoolniveau kan dit mogelijk een interessante aanvulling vormen. Op dit moment krijgt de school in deze feedback onder meer informatie over hoeveel van hun leerlingen de

eindtermen bereiken. Dit soort informatie zou dan eventueel aangevuld kunnen worden door aan te geven hoeveel van hun leerlingen tot een bepaalde groep behoren die specifieke problemen ondervinden. Op die manier kan aan de school een inhoudelijk rijkere feedback worden gegeven.

Daarnaast geeft dit resultaat ook aanleiding tot een reflectie over wat deze resultaten betekenen voor de traditionele manier van rapporteren bij peilingsonderzoek waarbij veelal ervan wordt uitgegaan dat een specifiek domein voor alle leerlingen een gelijkaardig patroon van moeilijkheid kent. In een meer meettechnische context kan men stellen dat voorondersteld wordt dat een bepaald domein unidimensioneel is. Het bestaan van subpopulaties wijst echter op een multidimensioneel karakter van een domein, net zoals DIF wijst op de aanwezigheid van meerdere dimensies. Dit betekent ook dat er een spanningsveld kan ontstaan tussen het centrale idee achter peilingsonderzoek (hoeveel leerlingen bereiken het vooropgestelde minimumniveau voor een bepaald domein?) en de vaststelling dat er subpopulaties bestaan die specifieke problemen ondervinden met bepaalde aspecten van een domein.

In het peilingsonderzoek wordt ook uitgebreide achtergrondinformatie over de leerlingen en hun school verzameld. Door deze informatie te koppelen aan het groepslidmaatschap wordt nog verder inzicht verkregen in voor welke leerlingen specifieke problemen vooral optreden. Als denkoefening werd ingegaan op de samenhang met handboekgebruik. Uit deze analyse blijkt dat het gebruik van mixture IRT-modellen de mogelijkheid biedt inhoudelijke relevante informatie bloot te leggen die het mogelijk maakt formatieve conclusies uit grootschalig peilingsonderzoek te trekken.

Deze paper toont dat ontwikkelingen op methodologisch vlak ook een belangrijke meerwaarde kunnen hebben vanuit vakinhoudelijk oogpunt. Het gebruik van modellen met latente groepen van leerlingen, zoals mixture IRT-modellen, kan ook voor andere vakinhouden een belangrijke aanvulling vormen op een meer klassieke manier van rapporteren. Een ander voorbeeld hiervan kan gevonden worden in het doctoraatsproefschrift van Marian Hickendorff (2011). In haar doctoraat onderzoekt zij onder meer hoe het in rekening brengen van oplossingsstrategieën bij complexe delingen meer inzicht kan verschaffen in de prestatiedaling die voor dit domein werd vastgesteld tussen de peiling van 1997 en 2004 (Hickendorff, Heiser, van Putten, & Verhelst, 2009).

Literatuur

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bloom, B. (1956). *Taxonomy of Educational Objectives, Handbook 1: The Cognitive Domain*. New York: Longman.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42*, 133-148.

De Ayala, R. J., Kim, S. H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing, 3-4*, 243-276.

DiBello, L. V., & Stout, W. (2007). Guest editor's introduction and overview: IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement, 44*, 285-291.

Hickendorff, M., Heiser, W. J., van Putten, C. M., & Verhelst, N. D. (2009). Solution strategies and achievement in Dutch complex arithemtic: Latent variable modeling of change. *Psychometrika, 74*, 331-350.

Hickendorff, M. (2011). *Explanatory latent variabele modeling of mathematical ability in primary school: Crossing the border between psychometrics and psychology*. Leiden, Nederland: Universiteit Leiden.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195-215.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modelling: A Monte Carlo simulation study. *Stuc-*

tural Equation Modeling: A Multidisciplinary Journal, 14, 535-569.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271-282.

Samuelsen, K. (2005). *Examining differential item functioning from a latent class perspective*. Dissertatie. University of Maryland, MD.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461-464.

Snijders, T. A. B., & Bosker, R. J. (2012) *Multilevel analysis: An introduction to basic and advanced multilevel modeling, second edition*. London: Sage.

Tatsuoka, K. K., Linn, R. L., Tatsuoka, M. M., & Yamamoto, K. (1988). Differential item functioning resulting from the use of different solution strategies. *Journal of Educational Measurement, 25*, 301-319.

Vermunt, J. K., & Magdison, J. (2000). *Latent-GOLD.* Belmont, MS: Statistical Innovations.

Webb, M. L., Cohen, A. S., & Schwanenflugel, P. J. (2008). Latent class analysis of differential item functioning on the Peabody Picture Vocabulary Test-III. *Educational and Psychological Measurement*, 68, 335-351.

Zumbo, B. D. (2007). Three generations of DIF analysis: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*, 223-233.

Manuscript aanvaard op: 29 oktober 2012

Auteurs

Daniël Van Nijlen is als doctor-assistent werkzaam aan het Centrum voor Onderwijseffectiviteit en -Evaluatie van de KU Leuven. **Rianne Janssen** is als hoofddocent verbonden aan het Centrum voor Onderwijseffectiviteit en -Evaluatie en de Onderzoeksgroep Kwantitatieve Psychologie en Individuele Verschillen van de KU Leuven.

Correspondentieadres: D. Van Nijlen, Centrum voor Onderwijseffectiviteit en -Evaluatie, KU Leuven, Dekenstraat 2 bus 3773, 3000 Leuven. E-mail: daniel.vannijlen@ppw.kuleuven.be

Abstract

Formative lessons from national assessments: the added value of mixture IRT models

Flanders started conducting national assessments in 2002 as an element of quality control in the Flemish educational system. National assessments primarily serve a summative purpose: how many students achieve the required minimal level in a certain domain? However, policymakers and educational practitioners also want to draw more formative conclusions from these assessments. Are there specific topics that pose problems for specific groups of students? It is illustrated how mixture IRT models can be an addition to the more traditional way of reporting on national assessments. Analyses using data from the 2009 national assessment on mathematics in the first stage of secondary education showed that two subgroups of students experienced specific problems when using polynomials, each consisting of about one third of the students. One group had trouble working with special factoring while another group actually only seemed to master basic items with regard to first degree equations.