Twee jaar WISCAT-pabo

- beschrijving, resultaten en ervaringen -

Gerard Straetmans
Cito Arnhem/Saxion Hogescholen
Theo Eggen
Cito Arnhem/Universiteit Twente

1 inleiding

Zo'n dertig jaar geleden studeerde de eerste auteur van dit artikel aan een pedagogische academie. Hij tekende de volgende herinnering op:

> De opleiding telde toen geen vier maar drie studiejaren met in het laatste jaar ruimte voor een specialisatie. Ik koos voor rekenen, omdat me dat een tamelijk belangrijke vaardigheid leek voor een onderwijzer. Ik herinner me hoe verwonderd ik toen was dat maar zo weinig studenten dezelfde keuze maakten, temeer omdat bij plenaire besprekingen van praktijkervaringen steeds opnieuw bleek dat studenten bij het geven van rekenlessen in de problemen kwamen door gebrek aan eigen rekenvaardigheid.

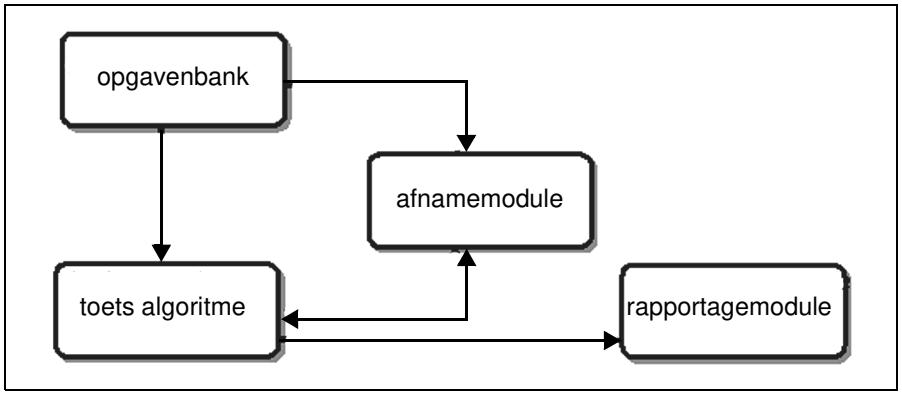
In de decennia daarna werd de problematiek van de tekortschietende rekenvaardigheid ook buiten de opleidingen opgemerkt en werden van overheidswege steeds nieuwe maatregelen opgelegd om die te bestrijden. Sinds begin 2006 weten we meer gefundeerd dat die maatregelen niet echt geholpen hebben, want toen de onderzoeksresultaten werden gepubliceerd bleek dat meer dan de helft van de eerstejaars pabo-studenten onvoldoende rekenvaardig bezat (Straetmans & Eggen, 2005). De landelijke discussie die daarop volgde, aangewakkerd door vaak ongenuanceerde krantenkoppen die suggereerden dat er niet alleen iets mis was met de rekenvaardigheid van pabo-studenten, maar met het hele Nederlandse onderwijs, leidde tot grotere daadkracht bij de besluitvormers. De HBO-raad besloot tot de ontwikkeling en verplichte afname van een gestandaardiseerde rekenvaardigheidstoets en de minister van onderwijs gelastte een onderzoek naar de oorzaken van de tekortschietende rekenvaardigheid bij eerstejaars pabo-studenten teneinde een definitieve oplossing te kunnen bereiken.

Het Cito verwierf de opdracht om een landelijk toetspakket te ontwikkelen waarmee de rekenvaardigheid van eerstejaars pabo-studenten kon worden vastgesteld. In dit artikel krijgt de lezer een indruk van de wijze waarop dit - inmiddels operationele - toetspakket functioneert. Daarbij wordt hoofdzakelijk ingegaan op de toetsinhoudelijke aspecten en niet of nauwelijks op de geautomatiseerde administratieve procedures die nodig zijn om het toetspakket te laten functioneren. Eerst worden globaal de componenten van het toets-

pakket beschreven. Daarna volgt een overzicht van de eerste toetsresultaten bij de doelgroep.

2 hoe zit WISCAT-pabo in elkaar?

Het toetspakket voor rekenen heeft de naam WISCAT-pabo gekregen. WISCAT is een acroniem voor WISkunde/rekenen COmputergestuurd ADaptief TOetspakket. De toevoeging pabo is nodig omdat voor andere doelgroepen ook WISCAT-pakketten ontwikkeld zijn of nog worden.



figuur 1: schematische weergave van WISCAT-pabo

Figuur 1 is een sterk vereenvoudigde schematische weergave van het toetspakket (waarbij, zoals gezegd, de technische infrastructuur buiten beschouwing is gelaten). WISCAT-pabo bestaat uit vier componenten die samen zorgdragen voor de samenstelling, afname, beoordeling en rapportage van toetsen bij kandidaten. Dat werkt globaal als volgt: het toetsalgoritme is een computerprogramma dat regels bevat die precies specificeren hoe tijdens een computergestuurde toetsafname een toets moet worden samengesteld uit een opgavenbank. In de opgavenbank liggen itemteksten, bijbehorend illustratiemateriaal en kwaliteitsgegevens opgeslagen over elke afzonderlijke opgave. De afnamemodule presenteert opgaven één voor één op het scherm, scoort de gegeven antwoorden van de kandidaat als goed of fout en geeft dit resultaat steeds door aan het toetsalgoritme. Als het toetsalgoritme de toets beëindigt, zorgt de rapportagemodule voor de terugkoppeling van resultaten naar kandidaat en docent. Aan drie van deze componenten zal hieronder meer in detail aandacht worden besteed.

de opgavenbank

De opgavenbank bevat een kleine duizend opgaven die tezamen een operationalisatie vor-

men van het begrip rekenvaardigheid. Figuur 2 laat zien welke domeinen met hoeveel items in de opgavenbank vertegenwoordigd zijn.[1]

Domein	Aantal opgaven	Waarvan hoofdrekenen
Basisoperaties, zoals optellen, aftrekken, delen, vermenigvuldigen.	181	141
Operaties met breuken, procenten, verhoudingen en decimale getallen.	431	152
Meten met enkelvoudige en samengestelde grootheden.	175	23
Meetkunde. Interpreteren van plattegronden en ruimtelijke figuren.	113	
Statistische gegevens ordenen, weergeven, samenvatten en interpreteren.	33	
Verbanden beschrijven met (woord)- formules en daarmee rekenen. Grafieken en tabellen aflezen en interpreteren.	51	
Totaal	**966**	**316**

figuur 2: beschrijving van de opgavenbank van WISCAT-pabo.

vraagtypen

Om docenten zo min mogelijk te belasten met extra werk heeft de opdrachtgever besloten tot een computergestuurd toetspakket. Dit legde uiteraard beperkingen op aan de te gebruiken vraagtypen. Bij de constructie van opgaven is gewerkt met twee vraagtypen die zich goed lenen voor geautomatiseerde scoring: de meerkeuzevraag en de kort-antwoordvraag. Bij dit laatste vraagtype gaat het om een opgave waarbij de kandidaat antwoord moet geven door één getal, woord of symbool in te vullen in een antwoordveld.

Het zal duidelijk zijn dat het met dit type vragen niet mogelijk is om zicht te krijgen op de aanpak die kandidaten hanteren bij het oplossen van de rekenopgaven. Dat hoeft ook niet, want het doel van WISCAT-pabo is het vaststellen van een minimaal noodzakelijk geacht rekenvaardigheidsniveau en niet het opsporen van eventuele leerbelemmeringen of misconcepties op rekenkundig gebied.

moeilijkheidsgraad van de opgaven

Alle opgaven zijn vooraf getest bij eerstejaars pabo-studenten om zo bruikbare gegevens te krijgen over belangrijke kwaliteitsaspecten, zoals de moeilijkheidsgraad. De moeilijkheidsgraad van de afzonderlijke opgaven is in hoge mate bepalend voor de meetkwaliteit van de samen te stellen toets. Het vaststellen van de moeilijkheidsgraad van een opgave is echter een kunst op zich. De meest gebruikte maat voor het beschrijven van de moeilijkheidsgraad van een item is de *p*-waarde. Dit is de proportie van een groep personen die het

$$p_i(\theta) = P(X_i = 1 \mid \theta) = \frac{\exp(a_i(\theta - \beta_i))}{1 + \exp(a_i(\theta - \beta_i))};$$

θ = de vaardigheid van de persoon; β_i = de moeilijkheidsgraad van item *i*;
α_i het discriminerend vermogen van item *i* (Verhelst, 1993).

betreffende item correct beantwoord heeft. Een lage *p*-waarde wordt in verband gebracht met een moeilijk item en een hoge *p*-waarde met een gemakkelijk item. Helaas is de *p*-waarde afhankelijk van het vaardigheidsniveau van de groep personen bij wie het betreffende item is afgenomen. Die eigenschap maakt de *p*-waarde ongeschikt als maat voor de moeilijkheidsgraad van items die bedoeld zijn voor herhaald gebruik.

Recente ontwikkelingen binnen de psychometrie hebben het probleem van de zogenoemde steekproefafhankelijke *p*-waarden opgelost door in een model de relatie tussen de moeilijkheidsgraad van het item en de vaardigheid van de persoon expliciet te beschrijven. Binnen de itemresponstheorie (IRT) zijn verschillende modellen geformuleerd waarvan we er hier één beschrijven. In dat model is de moeilijkheidsgraad van een item gedefinieerd als de vaardigheid waarbij de kans op een correct antwoord precies 50 procent is. Een voorbeeld kan dit verduidelijken.
Als we bij hoogspringen de lat precies zo hoog leggen als de sportvrouw of -man kan springen (dat wil zeggen op een hoogte die het gemiddelde is van alle lathoogten waar die persoon overheen gesprongen is), mogen we verwachten dat hij of zij in de helft van alle gevallen erover zal springen en in alle andere gevallen de lat eraf zal 'springen'. Als de moeilijkheidsgraad van het item groter is dan het vaardigheidsniveau van de persoon, wordt de kans op een correct antwoord kleiner dan 50 procent (de lat wordt er vaker afgesprongen dan dat de sporter erover heen springt). Is de moeilijkheidsgraad van het item kleiner dan het vaardigheidsniveau dan wordt de kans op een correct antwoord groter dan 50 procent (de sporter springt vaker over de lat dan dat hij deze eraf springt). Dit proces wordt in een wiskundig model beschreven waarin de kans gespecificeerd wordt op het geven van een goed antwoord door een persoon met een bepaalde vaardigheid. Deze kans is uiteraard afhankelijk van itemkenmerken, zoals de moeilijkheidsgraad en het discriminerend vermogen.

schaalconstructie

In proefafnamen bij meer dan 2500 eerstejaars pabo-studenten zijn, in een proces dat 'kalibreren' heet, uit de afnamegegevens de moeilijkheidsgraad en het discriminerend vermogen van elk item geschat. Met de geschatte parameters kon vervolgens worden gecontroleerd of het gekozen wiskundige model een goede beschrijving en voorspelling is van de proefafnamegegevens. Items die zich niet 'gedroegen' volgens het model werden verwijderd. De resterende items hadden een moeilijkheidsgraad die geldig is voor elke toekomstige kandidaat uit de doelgroep. Deze items konden geordend worden naar moeilijkheidsgraad om zo een schaal te vormen voor het meten van rekenvaardigheid. Het schaalbegrip houdt in dat een student, die een bepaald item correct beantwoordt, met een grotere kans

ook correct zal antwoorden op items met lagere schaalwaarden. Echter, naarmate items met hogere schaalwaarden worden aangeboden, zullen de kansen op een correct antwoord steeds verder afnemen.

items en studenten op dezelfde schaal

Het grote voordeel van de aldus geconstrueerde rekenvaardigheidsschaal is niet alleen dat het nu mogelijk is om een gewenst beheersingsniveau te definiëren dat onafhankelijk is van de gebruikte toets (de cesuurscore op de schaal geldt voor elke uit de geschaalde opgavenbank samen te stellen toets), maar ook dat de moeilijkheidsgraad van toetsopgaven en de rekenvaardigheid van personen op dezelfde schaal te positioneren zijn. Op die laatste eigenschap komen we terug bij de bespreking van het toetsalgoritme. Eerst gaan we in op het gewenste beheersingsniveau.

hoe goed moeten eerstejaars pabo-studenten kunnen rekenen?

Het construeren van een rekenvaardigheidsschaal is alleen zinnig als op die schaal ook een punt aanwijsbaar is dat bereikt moet zijn om een positieve beslissing over voortzetting van de studie te kunnen nemen. Bij het ontwikkelen van zo'n kwantitatieve beheersingsstandaard is uitgegaan van de in kwalitatieve termen omschreven standaard zoals die door een cesuurcommissie van pabo-rekendocenten is vastgesteld: Eerstejaars pabo-studenten moeten aan het eind van het eerste inschrijvingsjaar even goed kunnen rekenen als een goede leerling uit groep 8 van het basisonderwijs. Met een goede leerling bedoelde de cesuurcommissie een leerling wiens rekenprestaties tot de beste 20 procent van groep 8 leerlingen behoren. Door een deel van de opgavenbank ook te laten maken door een representatieve steekproef van leerlingen uit groep 8 kon de vaardigheidsverdeling van deze groep worden afgebeeld op de voor de pabo-studenten geconstrueerde schaal. Binnen die vaardigheidsverdeling is het punt op de schaal gezocht waaronder de prestaties van 80 procent van de groep 8 leerlingen vallen. Dat punt (schaalwaarde 103 op een schaal met een bereik van 0 tot 200) is de cesuur die door WISCAT-pabo gebruikt wordt voor het nemen van zak-/slaagbeslissingen.

het toetsalgoritme[2]

Doordat de vaardigheid van personen, de moeilijkheidsgraad van opgaven en de beheersingsstandaard op één en dezelfde schaal zijn af te beelden, wordt de weg vrijgemaakt voor adaptief toetsen. Bij een adaptieve toets wordt de moeilijkheidsgraad van de toets zorgvuldig afgestemd op de vaardigheid van de kandidaat. Dat is praktisch gezien alleen uitvoerbaar bij computergestuurde toetsafnamen waar toetssamenstelling en toetsafname ongeveer op hetzelfde moment kunnen plaatsvinden. Globaal werkt het adaptieve algoritme ongeveer als volgt:

1 Er wordt een opgave gepresenteerd op het beeldscherm.
2 De student geeft antwoord.
3 Het programma scoort het antwoord als 'goed' of 'fout' en schat op grond van alle tot dan beschikbare itemscores wat de vaardigheid is van de student.

4 Het programma beslist of de toets al beëindigd mag worden of dat er een nieuwe opgave moet worden aangeboden.
5 Als het laatste het geval is, wordt een opgave uit de opgavenbank geselecteerd die zo goed mogelijk past (onder andere qua moeilijkheidsgraad) bij de geschatte vaardigheid van de betreffende student. Meer nauwkeurig geformuleerd komt het erop neer dat die opgave wordt gekozen die op dat moment van de toetsafname de meeste informatie kan geven over de vaardigheid van de kandidaat. Of nog anders gezegd: een opgave die de vaardigheid van de kandidaat kan schatten met de kleinste meetfout.
6 Zodra de laatste opgave van de toets beantwoord is, krijgt de student de toetsuitslag op het scherm te zien.

Door deze speciale manier van toetssamenstelling krijgt elke kandidaat een toets te maken die nauwkeurig op zijn of haar vaardigheid is afgestemd. Het belangrijkste voordeel daarvan is dat daarmee toetsen verkregen worden die nauwkeuriger meten dan traditioneel samengestelde toetsen. Dit voordeel kan uiteraard ook 'benut' worden door een bepaalde meetnauwkeurigheid te bereiken met een kortere toets dan gebruikelijk. Verder is het zo dat elke student op zijn eigen niveau wordt uitgedaagd waardoor toetsen nooit te moeilijk of te makkelijk zijn. Dit heeft als bijkomend voordeel dat elke kandidaat een andere toets maakt waardoor de noodzaak vervalt om alle kandidaten op hetzelfde moment te toetsen. En ten slotte moet het voordeel van de vergelijkbaarheid van toetsprestaties genoemd worden. Doordat toetsscores worden omgezet in vaardigheidsschattingen op de geconstrueerde vaardigheidsschaal kunnen prestaties op toetsen direct met elkaar vergeleken worden. Dat is erg handig als men de voortgang wil beoordelen van kandidaten in het betreffende domein.

Adaptief toetsen kent ook nadelen. Zo houdt een adaptieve procedure geen rekening met de inhoudelijke opbouw van een toets. Dat is voor de pabo- rekentoets een ongewenste eigenschap want er dient een profielscore gegenereerd te worden met het oog op snelle, doelgerichte remediëring in geval van onvoldoende beheersing.
Om dit toch te kunnen bereiken zijn aan het adaptieve algoritme restricties opgelegd voor wat betreft toetslengte en inhoudelijke itemselectie. De toetslengte is vast en bedraagt vijftig items. Deze moeten zodanig gekozen worden dat voldoende items worden aangeboden uit alle deeldomeinen waarvoor een profielscore moet worden opgesteld:
– hoofdrekenen;
– basisvaardigheden;
– breuken, procenten, verhoudingen en decimale getallen;
– meten en meetkunde.

De opgelegde restricties hebben evenwel tot gevolg dat het adaptieve algoritme niet altijd het item zal selecteren dat vanuit psychometrisch oogpunt bezien de beste keuze zou zijn.

Er zijn nog andere mechanismen werkzaam die het adaptieve algoritme verhinderen om steeds het best passende item te kiezen. Die mechanismen hebben tot doel om over- en onderbenutting van de opgavenbank zoveel mogelijk tegen te gaan. Van overbenutting wordt gesproken als er items zijn die heel vaak geselecteerd worden in een toets. Onderbe-

nutting verwijst naar de situatie dat er items zijn die bijna nooit geselecteerd worden. Overbenutting is ongewenst, omdat dit de geheimhouding van het toetsmateriaal ernstig bedreigt, vooral in opleidingssituaties waar cursisten na elkaar in plaats van tegelijk getoetst worden. Onderbenutting is ongewenst, omdat slechts beperkt gebruik wordt gemaakt van de voorhanden zijnde items. Beide situaties versterken elkaar en zullen leiden tot een versnelde veroudering van de opgavenbank.
Om overbenutting tegen te gaan wordt in WISCAT-pabo een procedure toegepast die moet garanderen dat een item uit de itembank niet vaker dan in 30 procent van de af te nemen toetsen wordt opgenomen (Sympson & Hetter, 1985). Sympson & Hetter-achtige methoden voor afnamecontrole zijn effectief in het tegengaan van overbenutting van bepaalde items in een opgavenbank, maar helpen niet (voldoende) in het tegengaan van onderbenutting van de opgavenbank. Revuelta en Ponsoda (1998) hebben de zogenaamde progressieve methode voor afnamecontrole voorgesteld om onderbenutting tegen te gaan. Bij deze methode vindt itemselectie plaats op basis van een mix van aselecte trekking en maximale informatie (zie punt 5 in de beschrijving van het toetsalgoritme) bij de lopende vaardigheid. Deze methode heeft uiteraard als nadeel dat puur toeval een (kleine) invloed heeft op de itemselectie. Eggen (2001) heeft deze methode gegeneraliseerd naar de situatie waarin de oorspronkelijke progressieve methode slechts werkzaam is in een (nader te definiëren) eerste deel van de toets, waarna wordt overgegaan op de exclusieve selectie op basis van maximale informatie.

de rapportagemodule

Zodra de laatste toetsopgave beantwoord is, krijgt de kandidaat de (voorlopige) uitslag te zien op het beeldscherm. In het eerste operationele jaar werden de resultaten gepresenteerd in de vorm van een grafisch scoreprofiel. Dat is een grafisch weergegeven overzicht van de behaalde resultaten op de hele toets en op de vier onderdelen. Helaas bleken de grafisch weergegeven schalen onvoldoende bestendig tegen de verschillende beeldschermresoluties die scholen hanteren waardoor sommige studenten tegenstrijdige informatie kregen over de uitslag.

Historisch overzicht per student **Resultaten voor Katinka de Jonge[3], nr 2006453**			
Landelijk vastgestelde norm: 103			
poging	1	2	3
datum afname	06-09-05	12-01-06	06-06-06
Toetsresultaat	**88 (O)**	**97 (O)**	**105 (V)**
Hoofdrekenen (15)	96	103	109
Niet-hoofdrekenen (35)	91	95	102
Basisvaardigheden (20)	95	100	111
Breuken, proc., enz. (15)	84	99	102

Meten/meetkunde (15)	69 *	72 *	99

figuur 3: individueel resultatenoverzicht (*O*) = onvoldoende; *V* = voldoende;
* blijft significant achter bij toetsresultaat)

Daarom werd besloten het grafisch scoreprofiel te vervangen door een eenvoudige zak- of slaagmededeling en een expliciete vermelding van het onderdeel of de onderdelen waarvan het resultaat significant achterblijft bij het resultaat op de hele toets.
Voor docenten of andere supervisoren produceert de rapportagemodule meer cijfermatig georiënteerde resultaten die naar keuze per individu of per groep opgemaakt kunnen worden (fig.3).

3 de kwaliteit van WISCAT-pabo

de meetnauwkeurigheid

In een adaptieve toets worden de items opgenomen die de meeste informatie geven over de vaardigheid van de getoetste persoon. De informatie in een item en ook in alle items die de toets vormen, hangt direct samen met de nauwkeurigheid waarmee de vaardigheid van een persoon gemeten kan worden. Hoe meer informatie een toets kan genereren over de vaardigheid van een persoon des te groter de nauwkeurigheid waarmee gemeten wordt.
De nauwkeurigheid van WISCAT-pabo is op twee manieren onderzocht:
– op basis van gesimuleerde afnamen;
– op basis van operationele afnamen.

gesimuleerde afnamen

Bij een simulatiestudie worden tijdens een toetsafname voor fictieve kandidaten (waarvan de vaardigheid gekozen wordt door de onderzoeker) op de volgende wijze de antwoorden gegenereerd:
– De ware vaardigheid (het vaardigheidsniveau dat de onderzoeker kiest en waar de simulatie mee start) van de fictieve kandidaat wordt gekozen.
– Het eerste aan te bieden item wordt aselect gekozen uit de opgavenbank.
– Met de dan beschikbare parameters wordt het gebruikte IRT-model geëvalueerd, uitmondend in een kans op correcte beantwoording van het betreffende item.
– Daarna wordt een aselect getal g getrokken uit het interval (0,1). Voor kandidaat v met vaardigheid θ_v en item i wordt het van toepassing zijnde IRT-model geëvalueerd. Als $p_i(\theta_v) \geq g$ wordt het item gescoord als 'correct beantwoord' in het andere geval als 'fout beantwoord'.
– De lopende vaardigheidsschatting wordt vervolgens gebruikt voor de selectie van het volgende item.

Op basis van deze simulatiestudie is een gemiddelde standaardfout voor de hele toets

gevonden van 8,78 (uitgedrukt op een getransformeerde schaal). De standaardfout van een vaardigheidschatting vindt een zinvolle toepassing in de constructie van betrouwbaarheidsintervallen voor de ware vaardigheid van een persoon: $(\theta_k - \gamma.se\ (\theta_k), \theta_k + \gamma.se\ (\theta_k))$. Daarbij is γ een te specificeren constante die afhangt van de gewenste nauwkeurigheid. Als bijvoorbeeld $\gamma = 1{,}6449$, dan kan hiermee een 90 procent betrouwbaarheidsinterval voor de ware vaardigheid worden bepaald.

operationele afnamen

Op grond van de afnamegegevens van kandidaten gedurende het eerste operationele jaar is de betrouwbaarheid geschat van WISCAT-pabo met een in de literatuur voor adaptieve toetsen gebruikelijke methode (Thissen, 2000).
In het studiejaar 2006-2007 werd WISCAT-pabo 17699 keer afgenomen: 17610 afnamen bij pabo studenten en 89 afnamen bij kandidaten in een vooropleiding, die van plan waren zich het volgende studiejaar in te schrijven bij een pabo. De geschatte betrouwbaarheid (geoperationaliseerd als de proportie ware variantie in de geobserveerde testvariantie) in deze groepen bedroeg 0,90 als alleen de pabo studenten werden meegenomen en 0,91 als de schatting gebaseerd werd op alle 17699 kandidaten.

de beslissingsnauwkeurigheid

Wat is de kwaliteit van beslissingen die over studenten genomen worden? Bij het nemen van beslissingen over studenten op basis van toetsresultaten kunnen twee fouten gemaakt worden:

- de vaardigheid van de student wordt ten onrechte beoordeeld als voldoende (het resultaat op de toets voldoet aan de norm maar in werkelijkheid schiet de vaardigheid tekort);
- de vaardigheid van de student wordt ten onrechte beoordeeld als onvoldoende (het resultaat op de toets voldoet niet aan de norm maar in werkelijkheid is de vaardigheid toereikend).

Omdat het vaststellen van de ware vaardigheid praktisch gezien niet haalbaar is, is de kwaliteit van genoemde beslissingen moeilijk te achterhalen. Simulatiestudies kunnen een uitweg bieden. Op basis van het gekozen itemresponsmodel kan een toetsafname gesimuleerd worden waarbij zowel het ware vaardigheidsniveau (het vaardigheidsniveau dat de onderzoeker kiest en waar de simulatie mee start) als het geschatte vaardigheidsniveau (de schatting van de vaardigheid nadat de laatste toetsopgave 'beantwoord' is) bekend zijn en met elkaar vergeleken kunnen worden en het resultaat van die vergelijking te classificeren in een van de volgende categorieën (fig.4).
Als dit een groot aantal keren wordt herhaald voor 'kandidaten' van uiteenlopende vaardigheid, dan wordt een goed beeld verkregen van de kwaliteit van beslissingen die theoretisch haalbaar is.

	geschatte vaardigheid

ware vaardigheid		onvoldoende	voldoende
	onvoldoende	Correcte zakbeslissing	Ten onrechte genomen slaagbeslissing
	voldoende	Ten onrechte genomen zakbeslissing	Correcte slaagbeslissing

figuur 4: correcte en incorrecte zak-/slaagbeslissingen.

De uitgevoerde simulatiestudies met vijfduizend fictieve kandidaten hebben geleid tot het opstellen van een tabel voor beslissingsnauwkeurigheid (fig.5). In deze figuur staan percentages correcte en incorrecte beslissingen gebaseerd op toetsscores die onder het gebruikte itemresponsmodel gegenereerd zijn conform de hierboven beschreven simulatieprocedure. De percentages geven een indicatie voor de omvang van de beslissingsfouten die in reële toepassingssituaties te verwachten zijn(fig.5).

		geschatte vaardigheid	
		onvoldoende	voldoende
ware vaardigheid	onvoldoende	42,4	5,9
	voldoende	5,7	46,0

figuur 5: percentage correcte en incorrecte beslissingen op basis van gesimuleerde toetsafnamen volgens het WISCAT-pabo toetsalgoritme.

De tabel laat zien dat in ruim 88 procent van alle gevallen een correcte beslissing wordt genomen. Dit mag geïnterpreteerd worden als een hoge beslissingsnauwkeurigheid.

validiteit

Een belangrijk kwaliteitskenmerk is de validiteit van een toets. Strikt genomen is de term ‘valide toets’ verkeerd; toetsen zelf kunnen niet valide zijn. Validiteit heeft betrekking op de beslissingen die op grond van de toetsuitslag genomen worden. Dit betekent in principe dat een toets in situatie *A* valide kan zijn (lees: tot valide beslissingen leidt) en in situatie *B* niet. Daarom moet voor elke nieuwe toepassing de validiteit opnieuw onderzocht worden. Er zijn verschillende bronnen waaruit geput kan worden om een bijdrage te leveren aan het bewijs van validiteit (Messick, 1989).

inhoud-gerelateerd bewijs

Voor de operationalisatie van het construct ‘rekenvaardigheid’ zijn items ontwikkeld op basis van eindtermen die oorspronkelijk zijn opgesteld voor het vak rekenen-wiskunde, zoals dat verzorgd wordt in de lagere niveaus van de volwasseneneducatie. Dit eindtermendocument is op haar beurt grotendeels ontleend aan de kerndoelen voor rekenen in het basisonderwijs (Ministerie van Onderwijs, 1993). Beide documenten zijn opgesteld door rekenspecialisten uit de betrokken onderwijstypen en zijn het resultaat van vele screenings-

en verbeteringsronden. Het is daarom aannemelijk dat deze documenten een deugdelijke beschrijving geven van wat in de betreffende onderwijstypen onder rekenvaardigheid begrepen wordt. Een commissie van pabo-rekendocenten heeft bevestigd dat het betreffende eindtermendocument, na een lichte aanpassing, ook een goede beschrijving gaf van de gewenste rekenvaardigheid van eerstejaars pabo-studenten. Op grond van dit aangepaste eindtermendocument zijn items ontwikkeld die tezamen het zogeheten 'doeldomein' (Kane, 2006) vormen, ten aanzien waarvan we de verwachte score van een kandidaat willen weten. De verwachte score wordt geschat op grond van de prestatie op een toets, die een representatieve steekproef vormt uit het doeldomein.

construct-gerelateerd bewijs

Een belangrijke bron voor validering is onderzoek waaruit geconcludeerd kan worden dat de toets het construct meet waarin de toetsgebruiker geïnteresseerd is. Een construct is een hypothetische kwaliteit van een persoon (bijvoorbeeld rekenvaardigheid) die aanwezig verondersteld wordt om het gedrag van die persoon te kunnen verklaren. Het onderzoek dat in het kader van de proefafnamen werd uitgevoerd om de kwaliteit van de opgaven te beproeven, kan als bewijsbron in bedoelde zin fungeren. Dergelijk onderzoek beoogt een schaal te construeren voor het veronderstelde construct. De schaal wordt opgebouwd uit alle items die zich volgens een vooraf gespecificeerd itemresponsmodel 'gedragen'. Van die items kan gezegd worden dat ze allemaal hetzelfde construct meten. Omdat een belangrijke eigenschap van de itemresponstheorie is dat de vaardigheid van een persoon geschat kan worden met elke willekeurige deelverzameling van items (= toets) uit de geschaalde opgavenbank, kan een toetsgebruiker erop vertrouwen dat elke toets die uit de opgavenbank wordt samengesteld het construct meet, mits de kandidaat deel uitmaakt van de populatie waarvoor de schaal geconstrueerd is. Omdat de toetsen die door WISCAT-pabo worden afgenomen, zijn samengesteld uit een geschaalde opgavenbank mag daaruit geconcludeerd worden dat ze het construct meten dat aan de schaal ten grondslag ligt. Nader onderzoek zal uit moeten wijzen of dit construct als rekenvaardigheid getypeerd mag worden of dat er misschien (ook) een andere vaardigheid in het geding is. Vanuit de theorie over het gemeten construct kunnen bijvoorbeeld hypotheses geformuleerd worden over relaties met scores afkomstig van andere instrumenten (convergente en discriminerende validering) (Messick, 1989). Dergelijk onderzoek werd mogelijk gemaakt door het normeringsonderzoek dat werd uitgevoerd voor de ontwikkeling van een cesuurscore voor WISCAT-pabo.

		Taken							
Taken		1	2	3	4	5	6A	6B	7
1	Begrijpend lezen								
2	Woordenschat	0,56							
3	Spelling	0,53	0,41						
4	Informatiebewerking	0,56	0,51	0,54					

5	Herkennen persoonsvorm	0,32	0,17	0,44	0,38				
6A	Rekenen Eindtoets Basisond.	0,47	0,43	0,54	0,64	0,31			
6B	Rekenen pabo	0,42	0,39	0,54	0,54	0,33	**0,67**		
7	Schrijven	0,60	0,46	0,58	0,64	0,50	0,56	0,53	
	Aantal proefpersonen	323	323	323	323	323	323	323	323
	Aantal items	30	30	30	30	30	21	9	30

figuur 6: intercorrelatie tussen taken in Proeftoets 1

		Taken						
Taken		1	2	3	4A	4B	5	6
1	Aardrijkskunde							
2	Natuuronderwijs	0,43						
3	Geschiedenis	0,61	0,56					
4A	Rekenen Eindtoets Basisond.	0,46	0,43	0,47				
4B	Rekenen pabo	0,45	0,43	0,46	**0,65**			
5	Natuuronderwijs	0,47	0,60	0,58	0,40	0,38		
6	Aardrijkskunde	0,65	0,49	0,60	0,47	0,45	0,55	
	Aantal proefpersonen	458	458	458	458	458	458	458
	Aantal items	30	30	30	19	11	30	30

figuur 7: intercorrelaties tussen de taken in Proeftoets 10

In dat verband konden intercorrelaties berekend worden tussen de verschillende taken, waaronder ook taken bestaande uit WISCAT-pabo items, die proefpersonen gemaakt hadden in het kader van een proeftoetsing voor de Eindtoets Basisonderwijs 2002.

Figuur 6 en 7 geven een overzicht van alle berekende intercorrelaties tussen de taken (subtoetsen) van respectievelijk Proeftoets 1 en Proeftoets 10, zoals gemaakt door leerlingen in groep 8 in het kader van de proeftoetsing Eindtoets Basisonderwijs 2002. Ten behoeve van onderhavig onderzoek naar de constructvaliditeit van WISCAT-pabo zijn taak 6 (uit Proeftoets 1) en taak 4 (uit Proeftoets 10) hier opgesplitst in subtaak *A*, bevattende een set rekenitems voor de Eindtoets Basisonderwijs en subtaak *B*, bevattende een set rekenitems voor WISCAT-pabo.
Dat beide rekentaken hoog met elkaar correleren ($r_{6A,6B} = 0{,}67$ en $r_{4A,4B} = 0{,}65$), terwijl de correlaties met de andere taken beduidend lager zijn, is een sterke aanwijzing voor de constructvaliditeit van WISCAT-pabo.

Nog een aanwijzing voor de constructvaliditeit van WISCAT-pabo heeft betrekking op het - tijdens de proefafname van de WISCAT-pabo opgavenbank - gevonden verschil in gemiddelde toetsscore tussen mannen en vrouwen. Dit grote verschil (vrouw: gemiddelde geschatte vaardigheid = 96,88; man: gemiddelde geschatte vaardigheid = 118,70) kon niet verklaard worden door de verschillen in vooropleiding. Het is bekend uit de literatuur (Van der Velden, 1996) dat rekenprestaties van mannen en vrouwen vaak significant van elkaar verschillen, in die zin dat mannen hoger scoren dan vrouwen, terwijl die verschillen zich niet voordoen bij een schoolvak als bijvoorbeeld taal. Ook dit is een aanwijzing dat met de opgavenbank en de toetsen die eruit worden samengesteld vooral rekenvaardigheid wordt gemeten.
Een andere construct-gerelateerde bewijsbron voor de validiteit van de toetsscores die WISCAT-pabo oplevert, is gebaseerd op de proefafnamen van de opgaven uit de opgavenbank. De antwoorden die proefpersonen uit diverse doelgroepen tijdens de verschillende proefafnamen gegeven hebben, werden behalve voor het schatten van de moeilijkheidsgraden van de items ook gebruikt om een beeld te krijgen van de gemiddelde vaardigheid van studenten afkomstig uit verschillende vooropleidingen (fig.8).

Vooropleiding	**Gemiddelde vaardigheid**	**Standaarddeviatie**
mbo	84,7	30,6
havo	104,7	30,7
vwo	129,2	37,7
Totaal	100,0	30,0

figuur 8: gemiddelde en standaarddeviatie vaardigheid pabo-eerstejaars naar vooropleiding (proefafname resultaten)

In figuur 8 zijn de gemiddelden en standaarddeviaties van de geschatte vaardigheden van deze groepen uitgedrukt op de schaal voor rekenen-wiskunde. De gevonden verschillen tussen de gemiddelde vaardigheidsschattingen liggen in de verwachte richting: mbo'ers scoren gemiddeld lager dan havisten die op hun beurt gemiddeld weer lager scoren dan de vwo'ers. De ruim zeventienduizend operationele afnamen gedurende het eerste invoeringsjaar bevestigen de verschillen in rekenvaardigheid tussen de vooropleidingen (fig.9).

Vooropleiding	**Gemiddelde deviatie**	**Standaarddeviatie**
mbo	94	27,7
havo	109,8	26,4
vwo	137,3	32,6
Totaal	105	30,2

figuur 9: gemiddelde en standaarddeviatie vaardigheid pabo-eerstejaars naar vooropleiding (operationele resultaten).

De duidelijke en statistisch significante verschillen tussen de gemiddelde vaardigheidsschattingen, in de intuïtief veronderstelde richting, zijn een steuntje in de rug voor de veronderstelling dat met deze opgaven verschillen in rekenvaardigheid tussen personen gemeten kunnen worden.

4 de operationele resultaten

Tijdens de studiejaren 2006-2007 en 2007-2008 zijn meer dan 30.000 toetsen afgenomen. Hiermee is een schat aan informatie verworven over de behaalde toetsprestaties en over de ervaringen van de gebruikers. Achtereenvolgens zullen we in deze en de volgende paragraaf aandacht besteden aan deze onderwerpen. In totaal werd WISCAT-pabo in 2006-2007 17.610 keer en in 2007-2008 15.124 keer afgenomen. De gemiddelde scores per type vooropleiding staan in figuur 10.
De verschillen tussen de gemiddelden van de studenten met verschillende vooropleiding zijn in de verwachte richting en zijn statistisch significant. Er is een fors effect van de vooropleiding op de WISCAT-pabo score. Ook tekent zich duidelijk een toename in de gemiddelde score af bij alle vooropleidingen ten opzichte van 2006-2007.

Vooropleiding	Gemiddelde 2007-2008 (sd)	Gemiddelde 2006-2007 (sd)
mbo	98,9 (27,6) (N=6481)	94,0 (27,7) (N=7817)
havo	112,7 (26,0) (N=6628)	109,8 (26,4) (N=1872)
vwo	142,8 (34,0) (N=992)	137,3 (32,6) (N=1453)
Onbekend	117,0 (36,6) (N=1023)	102,4 (32,96) (N=168)
Totaal	109,1 (30,3) (N=15124)	105,0 (30,2) (N=17610)

figuur 10: WISCAT-pabo score en vooropleiding.

Als we kijken naar de gemiddelde scores op de deelvaardigheden in het studiejaar 2007-2008 (fig.11), dan zien we dat de verschillen tussen de vooropleidingen nagenoeg gelijk zijn.

Voor-opleiding		WISCAT Score	Score hoofd-rekenen	Score niet-hoofd-rekenen	Score basisvaar-digheden	Score breuken e.d.	Score meten
mbo	Gemiddelde	98,9	102,4	97,6	98,7	98,1	100,5
	Std. deviatie	27,6	39,5	26,3	29,6	30,2	32,8
havo	Gemiddelde	112,7	117,4	110,9	114,3	112,4	111,8
	Std. deviatie	26,0	38,15	24,9	27,9	29,2	31,6
vwo	Gemiddelde	142,8	149,5	139,4	140,5	142,9	143,3
	Std. deviatie	34,0	41,1	33,5	35,3	35,5	37,3
Onbe-kend	Gemiddelde	117,0	122,3	114,9	115,8	117,9	117,1
	Std. deviatie	36,6	45,5	35,6	37,0	39,3	40,0
Totaal	Gemiddelde	109,1	113,4	107,3	109,4	108,7	109,4
	Std. deviatie	30,3	41,3	29,1	31,9	33,0	34,8

figuur 11: gemiddelde totaal- en deelscores per vooropleiding in studiejaar 2007-2008

Binnen een vooropleidingsgroep zijn er nauwelijks verschillen tussen de deelvaardigheden basisvaardigheden, breuken en meten. Het is wel zo dat vooral havisten en vwo'ers relatief iets hoger scoren op hoofdrekenen ten opzichte van niet-hoofdrekenen. Ten opzichte van 2006-2007 is in deze algemene trends niets veranderd.
De geconstateerde algemene vooruitgang is ook bij alle deelvaardigheden aanwezig met een enkele uitzondering: de vwo'ers zijn ten opzichte van 2006-2007 veel vooruit gegaan

met de score op de basisvaardigheden en bijna niet met de gemiddelde scores op breuken.

In totaal maakten in het schooljaar 2007-2008 10.072 verschillende kandidaten een of meerdere keren de toets. Dat zijn er 906 minder dan in het schooljaar 2006-2007. De verhoudingen van de aantallen per vooropleiding zijn in beide jaren ongeveer gelijk (fig.12).

Vooropleiding	**Percentage 2007-2008**	**Percentage 2006-2007**
mbo	41,6	39,8
havo	48,8	48,4
vwo	9,6	11,7

figuur 12: percentage 'unieke' kandidaten per vooropleiding.

Het percentage studenten dat uiteindelijk slaagde voor de toets is gestegen van 75,6 procent in studiejaar 2006-2007 naar 79,1 procent in studiejaar 2007-2008 (fig.13). De verschillen tussen de vooropleidingen zijn groot. Bijna alle vwo'ers, ruim 85 procent van de havisten en 65 procent van de mbo'ers heeft in het schooljaar 2007-2008 uiteindelijk voldaan aan de vastgestelde norm. Voor elke vooropleidinggroep liggen deze percentages, respectievelijk 5 procent, 3 procent en 2,5 procent voor mbo, havo en vwo, hoger dan in het schooljaar 2006-2007.

vooropleiding		**Geslaagd in 2007-08**		**Geslaagd in 2006-07**	
		nee	ja	nee	ja
mbo	aantal	1323	2527	1709	2614
	%	34,4	65,6	39,5	60,5
havo	aantal	611	3898	865	4390
	%	13,6	86,4	16,5	83,5
vwo	aantal	24	862	65	1207
	%	2,7	97,3	5,1	94,9
onbekend	aantal	147	680	40	88
	%	17,8	82,2	31,2	68,8
Totaal	aantal	2105	7967	2679	8299
	%	20,9	79,1	24,4	75,6

figuur 13: geslaagd en vooropleiding

5 enkele ervaringen van gebruikers

De ervaringen van gebruikers, studenten en docenten, zijn tot op heden niet systematisch

bevraagd. Toch is het mogelijk, doordat studenten en docenten de ontwikkelaars van het toetspakket soms op eigen initiatief benaderden, een indruk te krijgen van wat de belangrijkste 'pijnpunten' zijn van de inzet van WISCAT-pabo. De meest saillante stellen we hieronder aan de orde.

moeilijk te begrijpen scoring

Bij conventionele toetsen wordt de toetsscore (meestal uitgedrukt als het aantal correct beantwoorde opgaven) via een of andere eenvoudige transformatie omgezet in een schoolcijfer of zak-/slaagbeslissing. Voor studenten zijn dergelijke relaties tussen toetsscore en beslissing inzichtelijk en daardoor acceptabel. Bij adaptieve toetsing is die relatie aanzienlijk complexer en bijgevolg minder inzichtelijk en acceptabel voor studenten. Ook voor de docenten is het meestal onduidelijk hoe zak-/slaagbeslissingen gerelateerd zijn aan de toetsprestaties. Het is begrijpelijk dat studenten daar opmerkingen over maken en om opheldering vragen. Alhoewel de gebruikte scoreprocedures op brede schaal geaccepteerd worden door psychometrici en goed gedocumenteerd zijn in de psychometrische literatuur, is het bijzonder lastig ze voor leken op een begrijpelijke wijze uit te leggen.

adaptiviteit zou stress oproepen

Voor studenten verschilt het maken van een adaptieve toets op bepaalde punten wezenlijk van het maken van een conventionele toets.

- Een opmerkelijk verschil bij het maken van een adaptieve toets ten opzichte van een conventionele (beeldscherm of papier) is dat er bij de eerstgenoemde geen toetsboekje is waar je fysiek of digitaal in kunt bladeren. De kandidaat moet de items waaruit de toets bestaat maken in de volgorde waarin ze door het toetsalgoritme gepresenteerd worden. Voor veel personen is dat onprettig, omdat ze eraan gewend zijn door een toets te kunnen bladeren voor een eerste oriëntatie op de opgaven. De dwingende noodzaak om elk gepresenteerd item onmiddellijk te moeten beantwoorden maakt hen zenuwachtig. Helaas is hier geen oplossing voor, omdat bij een adaptieve toets de samenstelling en de afname min of meer gelijktijdige processen zijn. Een goede voorbereiding op de aard van adaptieve toetsing kan helpen om minder snel gestrest te raken.
- Het feit dat de toets zich aanpast aan de vaardigheid van de student door de moeilijkheidsgraad van de volgende opgave te verhogen of juist te verlagen wordt door sommige studenten ook als een vervelende eigenschap beschouwd. Met name als er een opgave wordt gepresenteerd die opvallend eenvoudiger is dan de vorige, schijnt dit voor die studenten het signaal te zijn dat men 'op weg is' naar een onvoldoende resultaat. De uitwerking van de maatregel om onderbenutting tegen te gaan (zie pag.144, 'het toetsalgoritme') kan dit effect nog versterken omdat een puur toevallige selectie tot onverwacht makkelijke (ook onverwacht moeilijke, maar die vormen in dit kader een minder groot probleem) items kan leiden. De spanning die hiermee gepaard gaat zou de toetsprestatie in negatieve zin beïnvloeden. Op dit moment is het niet duidelijk in hoeverre deze klachten een systematisch probleem vormen en of ze werkelijk veroor-

zaakt worden door het adaptieve karakter van de toets. Wij vermoeden van niet. Dat vermoeden wordt ingegeven door de resultaten van een onderzoek dat we enige tijd geleden hebben uitgevoerd bij cursisten uit de basiseducatie. In dat onderzoek lieten we negentig cursisten elk twee rekentoetsen maken volgens het *design* in figuur 14 (Straetmans & Eggen, 1998).

Groep	Eerste toets	Tweede toets	Aantal personen
1	PBT	CBT	14
2	CBT	PBT	15
3	PBT	CAT	16
4	CAT	PBT	15
5	CBT	CAT	15
6	CAT	CBT	15

figuur 14: *design* voor onderzoek naar effect van afnameprocedure op toetsprestaties

Elke toets was samengesteld uit dezelfde opgavenbank en de toetsscores werden gerapporteerd op dezelfde onderliggende rekenvaardigheidsschaal. Het ging om een conventionele *paper-based* toets (PBT), een daaraan identieke *computer-based* toets (CBT) en een computergestuurde adaptieve toets (CAT). De inhoud van laatstgenoemde kon uiteraard niet volledig van tevoren worden vastgelegd en was dus niet identiek aan die van de andere twee toetsen. De eigenschappen van het gebruikte psychometrische model maakten het echter mogelijk om de geschatte vaardigheid op grond van een toetsprestatie op een van de hierboven genoemde toetsen direct te vergelijken met de geschatte vaardigheid op grond van een toetsprestatie op een van beide andere toetsen. De analyses lieten zien dat de gemiddelde vaardigheidsschattingen van PBT, CBT en CAT niet significant van elkaar verschilden. Wel bleken de vaardigheidsschattingen van de CAT significant nauwkeuriger dan die van beide andere toetsvormen.

eenzijdige gerichtheid op resultaat.

Sommige studenten merken op dat WISCAT-pabo zich te eenzijdig richt op de resultaten van rekenprocessen terwijl het voor (toekomstige) docenten juist zo belangrijk is om gericht te zijn op de processen zelf. Het klopt dat WISCAT-pabo geen oog heeft voor de rekenprocessen en dat alleen het resultaat telt. Daar zijn twee redenen voor. In de eerste plaats is het bij een computergestuurde toets vooralsnog niet mogelijk om het rekenproces van de kandidaat te analyseren en te beoordelen. In de tweede plaats is dat, gezien het doel van de toets, niet nodig. Het gaat niet om uitvoerige diagnostiek (opsporen en benoemen van misconcepties en leerbelemmeringen) maar om efficiënte onderscheiding van de studenten die al dan niet beschikken over de minimaal noodzakelijke voorkennis. Daarmee wil niet gezegd zijn dat het in kaart brengen en beoordelen van de aanpak niet belangrijk is; alleen moet

daar een andere toets voor worden ingezet.

6 tot slot

In dit artikel hebben we laten zien dat ingewikkelde meetproblematiek op onderwijskundig gebied het hoofd kan worden geboden door twee krachtige technologieën met elkaar te verbinden: moderne testtheorie en computergestuurde toetsing. De moderne testtheorie heeft de ontwikkeling van schalen mogelijk gemaakt waarop zowel de vaardigheid van een persoon als de moeilijkheidsgraad van toetsopgaven kan worden uitgedrukt en waarmee in principe de weg is vrij gemaakt voor toetsing-op-maat bij groepsgewijs afgenomen toetsen. Door de ontwikkeling van de personal computer kon dit ook in de praktijk worden gebracht, omdat de grote verwerkingssnelheid van een computer het mogelijk maakt de samenstelling van een toets, de toetsafname en de scoring als min of meer gelijktijdige processen af te handelen.
CAT is inmiddels de kinderschoenen ontgroeid en wereldwijd neemt de belangstelling toe om deze krachtige technologie voor concrete toetsdoeleinden in te zetten. De voordelen zijn vaak zeer aansprekend; zo ook in het geval van de toetsing van rekenvaardigheid bij eerstejaars pabo-studenten. De specifieke meetproblematiek die hierbij aan de orde was, kon dankzij de inzet van CAT het hoofd worden geboden. De eerste ervaringen zijn positief, maar lijken ons ook te leren dat acceptatie van dergelijke ingewikkelde toetsconcepten door de betrokkenen niet vanzelfsprekend is. Het is noodzakelijk om voldoende tijd en middelen vrij te maken teneinde alle betrokkenen uitvoerig in te kunnen lichten over de werkwijze en (eigen)aardigheden van CAT.

noten

1 De opgavenbank is, na twee operationele jaren, in september 2008 deels ververst. De tweehonderd meest gebruikte opgaven zijn vervangen door circa driehonderd nieuwe opgaven.
2 Lezers die meer in detail willen weten wat adaptief toetsen is en hoe het werkt, worden verwezen naar: http://edres.org/scripts/cat/catdemo.htm. Straetmans & Eggen (1998) en Eggen (2006)).
3 Gefingeerde naam.

literatuur

Eggen, T.J.H.M. (2001). Overexposure and underexposure of items in computerized adaptive testing. *Measurement and Research Department Reports 2001-1.* Arnhem: Cito. (De aanpassing van de Revuelta & Ponsoda procedure voor onderbenutting.)

Eggen, T.J.H.M. (2006). Adaptief toetsen in examens. *Examens*, 01-2006, 21-24.

Kane, M.T. (2006). Content-related validity evidence in test development. In: S.M. Downing & T.M. Haladyna (eds.). *Handbook of Test Development.* Mahwah (NJ): Lawrence Erlbaum Associates, Publishers.

Messick, S. (1989). Validity. In: R.L. Linn (ed.). *Educational Measurement (3rd ed.)*. New York: Macmillan, 13-104.

Ministerie van Onderwijs (1993). *Besluit Kerndoelen Basisonderwijs.* 's-Gravenhage: Sdu Uitgeve-

rij.

Revuelta, J. & V. Ponsoda (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 38*, 311-327.

Straetmans, G.J.J.M. & T.J.H.M. Eggen (1998). Computerized Adaptive Testing: What it is and how it works. *Educational Technology, 38*(1), 45-52.

Straetmans, G.J.J.M. & T.J.H.M. Eggen (1998). Comparison of Test Administration Procedures for Placement Decisions in a Mathematics Course. *Educational Research and Evaluation, 4*(3), 259-275.

Straetmans, G.J.J.M. & T.J.H.M. Eggen (2005). Afrekenen op rekenen: Over de rekenvaardigheid van pabo-studenten en de toetsing daarvan. *Tijdschrift voor hoger onderwijs, 23*(3), 123-139.

Sympson, J.B. & R.D. Hetter (1985). *Controlling item-exposure rates in computerized adaptive testing*. Paper presented at the annual conference of the Military Testing Association, San Diego.

Thissen, D. (2000). Reliability and measurement precision. In: H. Wainer (ed.). *Computerized Adaptive Testing. A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates, 159-184.

Velden, L.F.J. van der (1996). *Context, visie, aanpak en effectiviteit. De bestrijding van achterstanden van Nederlandse leerlingen in het basisonderwijs*. Groningen: Rijksuniversiteit (proefschrift).

Verhelst, N.D. (1993). Itemresponstheorie. In: T.J.H.M. Eggen & P.F. Sanders (red.). *Psychometrie in de Praktijk*. Arnhem: Cito, 83-178.