*Gerard Straetmans*
*Theo Eggen*

# Afrekenen op rekenen: over de rekenvaardigheid van pabo-studenten en de toetsing daarvan

*Dr. Gerard J.J.M. Straetmans (gerard.straetmans@citogroep.nl) en dr. ir. T.J.H.M. Eggen zijn als toetsdeskundigen werkzaam bij Cito in Arnhem; Gerard Straetmans is tevens lector Assessment bij Saxion Hogescholen.*

***Wie anderen wil helpen bij de verwerving van kennis en vaardigheden moet zelf voldoende niveau hebben op de betreffende vakgebieden. Bij de lerarenopleidingen voor het basisonderwijs (pabo's) wordt echter al heel lang geconstateerd dat veel instromende studenten 'dat' niveau niet halen. De Minister van Onderwijs wil nu een verplichte toetsing en een daarop gebaseerd bindend studieadvies tijdens het eerste studiejaar van de pabo. Maar wat is eigenlijk voldoende niveau en hoe kun je betrouwbaar en valide meten of studenten daaraan voldoen? In dit artikel worden deze vragen beantwoord voor het vakgebied rekenen/wiskunde.***

## Introductie

De tekortschietende rekenvaardigheid van pabo-studenten is sinds enkele decennia de inzet van hevige discussies en zwalkend overheidsbeleid. Al in 1987 beval de onderwijsinspectie de minister van Onderwijs, Cultuur en Wetenschappen aan om strenger te selecteren, zowel aan de poort als tijdens de propedeutische fase (Ministerie van Onderwijs en Wetenschappen, 1987). In het jaar daarop honoreerde de minister deze aanbeveling met wat toen de 'wiskunde-maatregel' werd genoemd. Voortaan moesten instromers in het pabo-onderwijs wiskunde A in hun vakkenpakket hebben. In 1990 werd de wiskunde-maatregel, om onduidelijke redenen, alweer ingetrokken. Waarschijnlijk zal het snel toenemende tekort aan leraren hier debet aan zijn geweest; de pabo's konden het zich gewoon niet meer permitteren om kieskeurig te zijn. Met de invoering van het Studiehuis in 1998 werd wiskunde voor havisten alsnog een verplicht eindexamenvak. De rekenvaardigheid van de eerstejaars pabo-student bleef ondertussen onverminderd zwak. Voor de minister kennelijk voldoende bewijs om begin 2005 te constateren dat wiskunde in het examenpakket geen garantie kan bieden voor een voldoende rekenvaardigheid en te pleiten voor afschaffing ervan als verplicht vak in het profiel Cultuur en Maatschappij van de havo. Om de rekenvaardigheid van pabo-studenten te waarborgen, acht de minister andere maatregelen effectiever, zoals aparte, voorbereidende rekenprogramma's in de vooropleiding en/of verplichte toetsing van de rekenvaardigheid in het eerste studiejaar en daarop gebaseerde bindende studieadviezen (OCW,

2005). Hoewel de mogelijkheid van bindende studieadviezen al langer openstaat voor pabo's wordt daar tot op heden weinig gebruik van gemaakt. Een verplichting in dit kader zal vrijwel zeker leiden tot een discussie over de kwaliteit van de beslissingen die op grond van toetsuitslagen genomen worden.

Dit artikel wil een bijdrage leveren aan de discussie over de tekortschietende kennis en vaardigheden van eerstejaars pabo-studenten. Onze positie daarin is dat het zoeken naar oplossingen, in de zin van remediërende maatregelen, pas zinvol wordt als helder is wat onder voldoende kennis en vaardigheid moet worden verstaan en hoe kan worden vastgesteld wie daaraan voldoet. Vanuit het oogpunt van financiële haalbaarheid en transparantie lijkt het niet verstandig om dit aan individuele pabo's over te laten. Standaardisatie, van zowel het gewenste niveau als van het toetsinstrumentarium, lijkt de beste garantie voor het nemen van transparante, betrouwbare en valide beslissingen over studenten. In dit artikel wordt voor het vakgebied rekenen/wiskunde een toetsprocedure beschreven die tegemoetkomt aan genoemde kwaliteitscriteria. Deze toetsprocedure maakt gebruik van een omvangrijke opgavenbank waarvan het functioneren van de opgaven onderzocht is bij vele honderden eerstejaars pabo-studenten. Een deel van de opgaven is ook bij leerlingen uit groep acht van het basisonderwijs afgenomen. De daarmee verkregen gegevens worden behalve voor het beschrijven van de kwaliteit van de opgaven ook gebruikt voor de operationalisatie van 'voldoende rekenvaardigheid' en voor het typeren van de rekenvaardigheid van de populatie eerstejaars pabo-studenten.

## Wat is voldoende rekenvaardigheid?

Het is merkwaardig om enerzijds te moeten constateren dat er binnen de pabo-gemeenschap grote overeenstemming bestaat over het tekort aan rekenvaardigheid van eerstejaars studenten en anderzijds dat de belangrijkste actoren binnen die gemeenschap geen eenduidig antwoord weten te geven op de vraag wat het gewenste rekenvaardigheidsniveau zou moeten inhouden. Docenten rekenen en wiskunde op pabo's hebben het doorgaans over een vaardigheid die voldoende ontwikkeld moet zijn om de rekenopgaven die in het basisonderwijs aan de orde komen, op allerlei manieren te kunnen uitleggen aan kinderen. Sommigen formuleren het wat concreter als de rekenvaardigheid conform die van goede leerlingen uit groep 8. Maar meestal komt het erop neer dat het gewenste rekenvaardigheidsniveau niet kwalitatief omschreven wordt maar slechts kwantitatief wordt uitgedrukt als het percentage correct te beantwoorden opgaven op de gebruikte rekentoets. Op zichzelf is dat niet verkeerd. Maar omdat er geen landelijke toets is, zijn er vermoedelijk net zoveel gewenste rekenvaardigheidsniveaus als toetsen die gebruikt worden om de rekenvaardigheid van eerstejaars studenten in kaart te brengen.

Het is een situatie die enigszins vergelijkbaar is met de situatie op het gebied van maten en gewichten in Nederland vóór de Franse tijd. Elke streek en dikwijls ook nog elke grote plaats daarbinnen gebruikte toen zijn eigen maten en gewichten. De Amsterdamse duim was 2,6 cm, die van Den Bosch 2,9 cm en die van Gelderland 2,27 cm. Zonder kennis te nemen van de afmeting van de gehanteerde duim kon je niet weten hoeveel je van iets

gekocht had. Om de chaos die dit met zich meebracht te verminderen, introduceerden de Fransen het metrieke stelsel met onder andere de meter als eenheid voor lengte.

Een soortgelijke standaardisatie als bij de invoering van het metrieke stelsel wordt ook nagestreefd door de psychometrie: de wetenschap die zich bezighoudt met het ontwikkelen van onderwijskundige en psychologische meetinstrumenten. De psychometrie is inmiddels zo ver ontwikkeld dat voor allerlei kennisgebieden schalen ontwikkeld kunnen worden waarmee de kennis en vaardigheid waarover een persoon beschikt, gemeten kan worden. Met behulp van dergelijke schalen is het niet alleen mogelijk om een gewenst vaardigheidsniveau te definiëren dat onafhankelijk is van de gebruikte toets, maar ook om de moeilijkheidsgraad van toetsopgaven en de rekenvaardigheid van personen te positioneren op dezelfde schaal. Een eigenschap die adaptieve toetsafnames mogelijk maakt. Daar komen we later in dit artikel op terug. Eerst volgt een toelichting op het schaalbegrip en de ontwikkeling van een schaal voor rekenvaardigheid.

## Een schaal voor rekenvaardigheid

Resultaten van metingen worden uitgedrukt op een schaal. Dat kan een heel eenvoudige (nominale) schaal zijn die de te meten objecten slechts benoemd in onderscheidende zin: 'man' of 'vrouw' bijvoorbeeld. Iets ingewikkelder zijn de zogeheten ordinale schalen die de te meten objecten kunnen ordenen op grond van de te meten eigenschap, zoals bijvoorbeeld bij 'soldaat', 'onderofficier', 'officier', 'hoofdofficier'. In beide genoemde schaaltypen kunnen de waarden van de te meten eigenschap ook met getallen worden aangeduid. Maar echt rekenen met deze getallen is niet goed mogelijk omdat er geen vaste afstanden tussen de opeenvolgende waarden zitten. Interessant wordt het pas vanaf de intervalschalen. Bij dergelijke schalen is er een meeteenheid die de schaal specificeert waarmee de te meten eigenschap numeriek gerepresenteerd wordt. Bij een lengteschaal, bijvoorbeeld, is die meeteenheid vaak de centimeter.

Vanwege de betere mogelijkheden die dit biedt om prestaties van leerlingen onderling of met een of andere standaard te vergelijken, wordt er in het onderwijs ook gestreefd naar meten op intervalniveau. In dit artikel bespreken we de ontwikkeling van een intervalschaal voor het meten van de rekenvaardigheid van pabo-studenten.

Een schaal voor rekenvaardigheid zou volgens bovenstaande redenering gespecificeerd worden door een reeks items die (aspecten) van rekenvaardigheid meten. Hoe meer items uit die reeks correct beantwoord zijn door een persoon, des te hoger zijn positie op de schaal en, zo wordt geredeneerd, des te groter zijn rekenvaardigheid. Problematisch in dit verband is dat in de toetsscore (de sommatie van itemscores) de vaardigheid van de persoon en de moeilijkheidsgraad van de gemaakte items op onontwarbare wijze vervlochten zijn. Heeft iemand veel items correct of fout beantwoord omdat hij een hoge respectievelijk lage rekenvaardigheid heeft of omdat de opgaven zo gemakkelijk respectievelijk moeilijk waren? Dit probleem is alleen te verhelpen door de positie op de schaal niet alleen te laten bepalen door het aantal correct beantwoorde opgaven maar ook door de moeilijkheidsgraad van die opgaven.

*Over moeilijkheidsgraden*

Het vaststellen van de moeilijkheidsgraad van een opgave is 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 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. Tabel 1 illustreert dit verschijnsel. De proefpersonen uit de tweede proefafname waren kennelijk vaardiger dan de proefpersonen uit de eerste afname: gemiddeld maakte de eerstgenoemde groep ruim 8 van de 10 items goed terwijl laatstgenoemde groep bijna 6 van de 10 items correct beantwoordde. Dit verschil in vaardigheid uit zich ook in de p-waarden; die zijn in de rechterkolom aanzienlijk hoger dan in de linkerkolom terwijl het toch om dezelfde items gaat.

Tabel 1 Fluctuerende p-waarden.

| item | proefafname I<br>aantal proefpersonen: 433<br>gemiddelde score: 5,9 | proefafname II<br>aantal proefpersonen: 546<br>gemiddelde score: 8,1 |
|---|---|---|
| 1 | 0,51 | 0,76 |
| 2 | 0,63 | 0,85 |
| 3 | 0,47 | 0,69 |
| 4 | 0,51 | 0,80 |
| 5 | 0,62 | 0,85 |
| 6 | 0,43 | 0,68 |
| 7 | 0,61 | 0,80 |
| 8 | 0,71 | 0,89 |
| 9 | 0,72 | 0,89 |
| 10 | 0,72 | 0,91 |

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. Er zijn verschillende modellen waarvan er hier één wordt beschreven. In dat model is de moeilijkheidsgraad van een item gedefinieerd als de vaardigheid waarbij de kans op een correct antwoord precies 50% is. Een voorbeeld kan dit verduidelijken. Als bij hoogspringen de lat exact zo hoog wordt gelegd als de sportman of -vrouw kan springen, mag worden verwacht 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% (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% (de sporter springt vaker over de lat dan dat hij deze eraf springt). Het functioneren van dit model kan wiskundig beschreven worden met de volgende vergelijking, die de kans specificeert op een correct antwoord (de score op item *i*, $X^i = 1$) als een persoon met een bepaalde vaardigheid een item van een bepaalde moeilijkheidsgraad maakt:

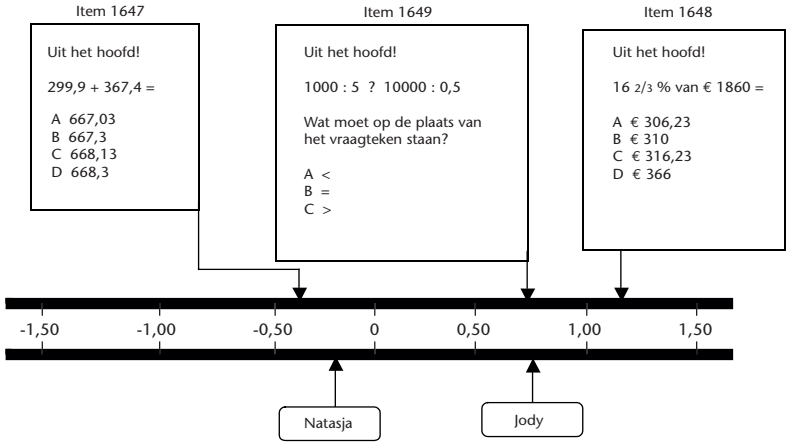
$$P = (X_i = 1/\ \theta,\ \delta_i) = \frac{e^{(\theta-\delta_i)}}{1 + e^{(\theta-\delta_i)}}$$

Daarbij geeft $\theta$ de vaardigheid aan van de persoon en $\delta_i$ de moeilijkheidsgraad van item i. De constante e (2,718) is een schalingsfactor (Verhelst, 1993).

In proefafnames bij de doelgroep worden, in een proces dat 'kalibreren' heet, uit de afnamegegevens de moeilijkheidsgraden van de items geschat en vervolgens, op basis van de dan bekende itemparameters, de vaardigheden van de proefpersonen. Met de dan beschikbare item- en persoonsparameters kan worden gecontroleerd of het gekozen wiskundig model een goede beschrijving en voorspelling geeft van de proefafnamegegevens. Items die zich niet 'gedragen' volgens het model worden verwijderd. De resterende items hebben moeilijkheidsgraden die geldig zijn voor elke toekomstige kandidaat uit de doelgroep. Deze items kunnen geordend worden naar moeilijkheidsgraad om zo een schaal te vormen. Het schaalbegrip houdt in dat een persoon die een bepaald item correct beantwoordt een grotere kans heft ook correct te 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 personen op één schaal*

Het voordeel van de gebruikte psychometrie is dat de geschatte item- en persoonsparameters op dezelfde schaal kunnen worden afgebeeld. De getallen die op deze schaal worden gebruikt, hebben geen absolute betekenis. Vergelijkbaar met bijvoorbeeld het meten van temperatuur zijn het nulpunt en de meeteenheid vrij te kiezen. Gebruikelijk (Verhelst, 1993) is om te kiezen voor een schaal waarbij het gemiddelde van de itemmoeilijkheden 0 is. De positie die iemand op de schaal inneemt, kan eenvoudig geïnterpreteerd worden in termen van kansen op een correcte beantwoording van de items op dezelfde schaal. Figuur 1 illustreert dit aan de hand van een voorbeeld.




Figuur 1 Items en persoonsparameters op één en dezelfde schaal.

Student Jody heeft een rekentoets gemaakt die is samengesteld uit de gekalibreerde itembank. Op grond van de gegeven antwoorden is zijn vaardigheid op de rekenvaardigheidsschaal geschat op 0,70. Op de rekenvaardigheidsschaal zijn ook drie hoofdrekenitems afgebeeld (de lezer dient zich te realiseren dat in werkelijkheid vele honderden items een positie innemen op de rekenvaardigheidsschaal). Hun positie zegt iets over de relatieve moeilijkheidsgraad. Item 1649 heeft een geschatte moeilijkheidsgraad van 0,696. De kans dat Jody dit item correct zal beantwoorden is volgens het bovenstaande wiskundig model ongeveer 50%. Op alle items met een lagere schaalwaarde dan item 1649 heeft Jody meer dan 50% kans om een correct antwoord te geven. Voor item 1647 bijvoorbeeld, met een moeilijkheidsgraad van –0,294, is die kans zelfs 95%. Zou Jody items krijgen aangeboden met een hogere schaalwaarde dan die van item 1649, dan dalen zijn kansen op het geven van correcte antwoorden onder de 50%. De kans bijvoorbeeld dat Jody item 1648, met een moeilijkheidsgraad van 1,086, correct weet te beantwoorden, is maar 24%. Natasja's rekenvaardigheid blijft ver achter bij die van Jody; op grond van een door haar gemaakte toets wordt die geschat op –0,13. Haar kansen op een correcte beantwoording van de drie genoemde items zijn dan ook veel kleiner, namelijk 8%, 62% en 2,5% voor respectievelijk item 1649, 1647 en 1648.

## De rekenvaardigheid van eerstejaars pabo-studenten

De hierboven beschreven technieken zijn gebruikt om een itembank van ongeveer 900 items te kalibreren op grond van een grootschalige proefafname waaraan vele honderden eerstejaars studenten, afkomstig van vijftien pabo's verspreid over het hele land hebben deelgenomen. Tabel 2 geeft globaal de inhoudelijke opbouw van de itembank weer.

Tabel 2 De itembank rekenen/wiskunde voor pabo's.

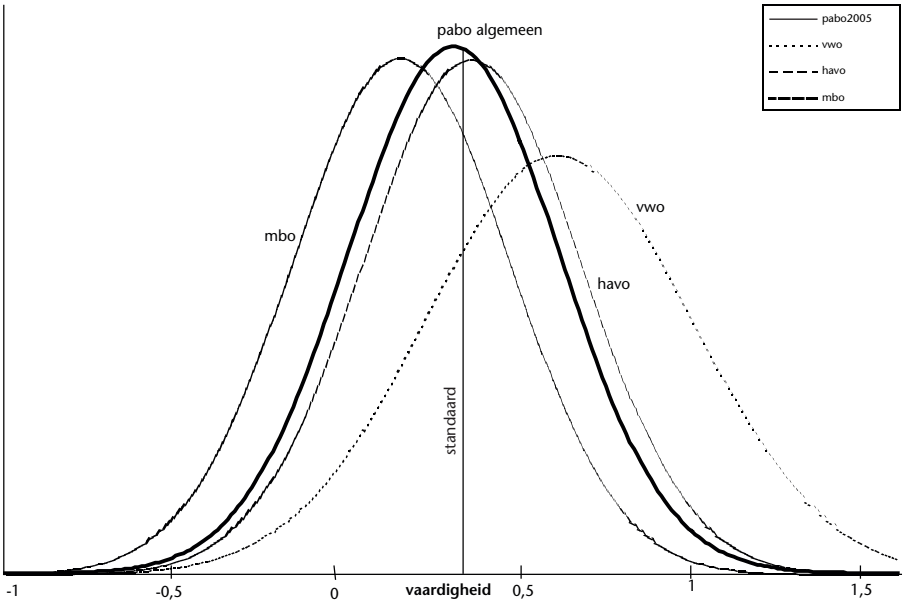
| domein | subdomein | aantal items | waarvan 'hoofdrekenen' |
|---|---|---|---|
| Rekenen | Basisvaardigheden | 203 | 141 |
| | Breuken, procenten, verhoudingen en decimale getallen | 349 | 131 |
| | Meten | 142 | 14 |
| Meetkunde | | 105 | |
| Informatieverwerking, kans en statistiek | | 33 | |
| (Woord)algebra, verbanden, grafieken en functies | | 55 | |
| | | **887** | **286** |

*De propedeutische eis gekwantificeerd*

De circa 900 items vormen samen een schaal voor rekenvaardigheid. De toetsresultaten die de eerstejaars pabo-studenten bij de proefafnames behaald hebben, zijn gebruikt om hun vaardigheden te schatten op deze schaal. Figuur 2 laat de verdeling zien van de

eerstejaars pabo-studenten (qua vaardigheid) op de geconstrueerde schaal voor rekenvaardigheid. Interessanter echter is dat op dezelfde rekenvaardigheidsschaal ook een standaard is afgebeeld. De standaard is een schaalwaarde die het gewenste rekenvaardigheidsniveau representeert van een pabo-student aan het eind van het eerste studiejaar. Dat niveau is in kwalitatieve termen door docenten rekenen/wiskunde van een aantal pabo's in het zuiden van het land als volgt omschreven: Aan het einde van de propedeutische fase dient de student een rekenvaardigheid te hebben die vergelijkbaar is met die van een goede leerling uit groep 8. De volgende procedure is gevolgd om deze kwalitatieve omschrijving om te zetten in een schaalwaarde op de rekenvaardigheidsschaal:

1 Uit de gekalibreerde itembank is een representatieve verzameling van items getrokken.
2 Deze items zijn onder toetscondities gemaakt door enkele honderden leerlingen uit groep 8 van het basisonderwijs.
3 De vaardigheidsverdeling van de groep-8-leerlingen is geschat op de rekenvaardigheidsschaal.
4 Op de rekenvaardigheidsschaal is vervolgens het punt gezocht waaronder 80% van de groep-8-leerlingen zit (de percentielscore 80 is de operationalisatie van de 'goede' leerling uit groep 8 die de rekendocenten van de pabo's op het oog hebben).
5 Dit punt op de rekenvaardigheidsschaal (met schaalwaarde 0,33) wordt de standaard waaraan toekomstige toetsresultaten van pabo-studenten zullen worden afgemeten.




Figuur 2 Vaardigheidsverdeling van eerstejaars pabo-studenten (algemeen en naar vooropleiding) voor rekenen/wiskunde.

*Hoe goed rekenen eerstejaars pabo-studenten?*
Wat uit figuur 2 in één oogopslag valt op te maken, is dat meer dan de helft van de eerstejaars pabo-studenten een geschatte vaardigheid heeft die niet voldoet aan de propedeutische eis en dus zal zakken voor een toets waarin die norm wordt toegepast. Verder wordt duidelijk dat er grote verschillen zijn in rekenvaardigheid tussen eerstejaars studenten. Dit wordt met name zichtbaar als bij het opmaken van vaardigheidsverdelingen rekening wordt gehouden met de aard van de vooropleiding. Tabel 3 presenteert zowel de gemiddelde vaardigheidschattingen en standaardafwijkingen van de diverse deelgroepen als de verwachte afwijspercentages die daaruit zijn af te leiden.

Tabel 3 Beschrijving van de vaardigheidsverdeling voor rekenen/wiskunde van eerstejaars pabo-studenten.

| | **Eerstejaars pabo-studenten** | | | | | |
|---|---|---|---|---|---|---|
| | **algemeen** | **mannen** | **vrouwen** | **mbo** | **havo** | **vwo** |
| gemiddelde vaardigheid | 0,307 | 0,499 | 0,275 | 0,150 | 0,355 | 0,607 |
| standaardafwijking | 0,343 | 0,327 | 0,331 | 0,314 | 0,315 | 0,387 |
| afwijspercentage | 53% | 30% | 57% | 72% | 47% | 24% |

De cijfers in tabel 3 geven aan dat de geluiden over de tekortschietende rekenvaardigheid van eerstejaars pabo-studenten niet uit de lucht gegrepen zijn. Waar 53% niet voldoet aan een norm, mag gerust gesproken worden van een probleem. Zoals verwacht heeft de genoten vooropleiding een grote invloed op de gemeten rekenvaardigheid. Van studenten met een vwo-diploma voldoet naar schatting een kwart niet aan de norm. Bij havisten is dit iets minder dan de helft en bij mbo'ers bijna driekwart. Het gevonden verband tussen vooropleiding en het hebben van een rekenvaardigheid die voldoet aan de propedeutische eis is zeer significant (chi-kwadraat = 75,18; df=2; $p < .01$). Opmerkelijk is ook het zeer grote verschil in rekenvaardigheid tussen mannen en vrouwen. Dit verschil kon niet verklaard worden door verschillen in vooropleiding (chi-kwadraat = 4,47; df = 2; n.s.).

Behalve met cijfers is de rekenvaardigheid van eerstejaars pabo-studenten ook in beeld te brengen aan de hand van enkele karakteristieke opgaven. In principe kunnen dergelijke opgaven voor elk vaardigheidsniveau gepresenteerd worden, maar hier wordt dit alleen gedaan voor de eerstejaars pabo-student met een gemiddelde vaardigheid. In figuur 3 zijn daartoe drie items opgenomen, aangeduid met de letters A, B en C, waarvan de kans op een correcte beantwoording door een student met een gemiddeld vaardigheidsniveau respectievelijk 0,20, 0,50 en 0,80 bedraagt.

*Meten van rekenvaardigheid: wensen vanuit de praktijk*
In de vorige paragrafen hebben we het begrip rekenvaardigheidsschaal aan de orde gesteld en daarnaast de markering op die schaal van het gewenste vaardigheidsniveau van eerstejaars pabo-studenten besproken. In het vervolg van dit artikel gaat onze aandacht uit naar de toetsing op grond waarvan de rekenvaardigheid van studenten op de schaal is af te beelden.

item A Kans op correcte beantwoording door de gemiddelde eerstejaars pabostudent: 0,20

Een auto van € 22000 wordt 20% goedkoper. De nieuwe prijs wordt daarna nog eens met 10% verlaagd.

Wat is het percentage van de totale prijsverlaging?

_____ %

item B Kans op correcte beantwoording door de gemiddelde eerstejaars pabostudent: 0,50

De tuin van Rinus is 8,40 meter lang en 5,60 meter breed.

Rinus maakt een plattegrond van zijn tuin met een schaal van 1 : 20.

Wat worden lengte en breedte van de tuin op de plattegrond?

Lengte: _____ centimeter

Breedte: _____ centimeter

item C Kans op correcte beantwoording door de gemiddelde eerstejaars pabostudent: 0,80

Uit het hoofd!

0,2 x 1,5 =

A 0,03
B 0,3
C 3
D 30

Figuur 3 Items met een hoge (item A), gemiddelde (item B) en lage (item C) moeilijkheidsgraad voor de eerstejaars pabo-student met een gemiddelde vaardigheid voor rekenen/wiskunde.

In gesprekken met docenten rekenen/wiskunde van diverse pabo's zijn de volgende wensen geformuleerd ten aanzien van de toetsing:

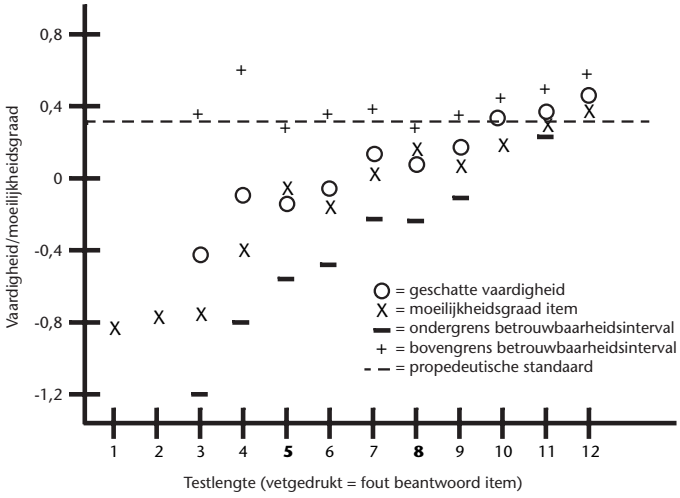
- Het toetsinstrumentarium moet tegemoet kunnen komen aan de grote verschillen in rekenvaardigheid bij eerstejaars pabo-studenten.
- Toetsresultaten moeten eenvoudig te interpreteren zijn in termen van beheersing en aanwijzingen geven voor eventuele remediëring.
- Het toetsinstrumentarium moet inzicht geven in de ontwikkeling van de rekenvaardigheid bij individuele studenten.
- Omdat de toetsuitslagen worden gebruikt voor het nemen van ingrijpende beslissingen over studenten is het zaak dat aannemelijk gemaakt kan worden dat deze beslissingen terecht zijn.
- Het gebruik van het toetsinstrumentarium moet zo min mogelijk inspanning vragen van de kant van de docent.

Een zwaar, maar haalbaar wensenpakket. Aan de opgesomde wensen kan tegemoet worden gekomen door moderne schalingstechnieken, zoals hierboven besproken, te combineren met moderne methodes om toetsen af te nemen. Daarbij wordt gedoeld op speciale toepassingen van *computer-based testing*, die in de literatuur bekend staan als computergestuurde adaptieve toetsing (CAT).

## Computergestuurde Adaptieve Toetsing

Anders dan bij andere vormen van computergestuurd toetsen gaat het bij CAT niet alleen om de afname op een beeldscherm van een opgeslagen toets en het automatisch verwerken van de responsen, maar (vooral) om de geautomatiseerde samenstelling van een toets uit een itembank. Er zijn globaal drie manieren waarop software toetsen kan samenstellen uit een itembank. Volgens de eerste manier worden items puur toevallig uit de itembank getrokken. Behalve op de toetslengte heeft men geen enkele controle op de toets die door de software wordt samengesteld. Bij de tweede manier probeert de software zich te houden aan een toetsmatrijs die voorschrijft hoeveel items over welke leerstofonderwerpen in de toets moeten worden opgenomen. De derde manier, tenslotte, probeert die items te kiezen die de meetfout van de toets zo klein mogelijk maken. Dit wordt bereikt door items te kiezen die qua moeilijkheidsgraad zo goed mogelijk zijn afgestemd op de vaardigheid van de persoon die getoetst wordt. Deze derde manier wordt als uitgangspunt gehanteerd bij adaptief toetsen. Figuur 4 geeft het principe van adaptief toetsen grafisch weer. Op de horizontale as wordt de toetslengte weergegeven. Op de verticale as worden zowel de geschatte vaardigheden van de persoon (aangegeven met cirkeltjes) als de moeilijkheidsgraden van de items (aangegeven met kruisjes) afgebeeld. De gestippelde lijn die parallel loopt aan de horizontale as geeft de standaard weer ofwel de vaardigheid waarover kandidaten moeten beschikken voor een positieve uitslag op de toets. In dit specifieke geval is het algoritme dat de samenstelling van de toets regelt zo ingesteld, dat het adaptieve proces pas begint bij de selectie van het vierde item. Op die manier kan men geforceerd eenvoudige items aanbieden bij de start van de toets, in een poging om eventuele toetsangst te reduceren. Nadat de kandidaat antwoord heeft gegeven op het derde item wordt voor de eerste keer de vaardigheid geschat. Uiteraard kan deze schatting na slechts drie beantwoorde items niet erg nauwkeurig zijn. Het programma schat behalve de vaardigheid ook de gemaakte meetfout en gebruikt die om een betrouwbaarheidsinterval om de geschatte vaardigheid te leggen. De min- en plus-tekens staan voor respectievelijk de onder- en bovengrens van dit betrouwbaarheidsinterval. Het betrouwbaarheidsinterval geeft de (zelf te kiezen) waarschijnlijkheid aan dat de ware vaardigheid van de kandidaat ligt tussen de onder- en de bovengrens. De ware vaardigheid is de vaardigheid van de persoon op de rekenvaardigheidsschaal. Uit figuur 4 valt eenvoudig op te maken dat de nauwkeurigheid van de vaardigheidsschattingen snel groter wordt als het aantal beantwoorde items toeneemt. In dit specifieke geval is er gekozen voor een dynamische stopregel. Dat wil zeggen dat het toetsalgoritme de toetsafname beëindigt zodra het betrouwbaarheidsinterval rondom de meest recente vaardigheidsschatting in zijn geheel onder of boven de standaard ligt. Hier doet zich die situatie voor na beantwoording van het

twaalfde item: er is (in dit specifieke geval) 90% kans dat de ware vaardigheid van de kandidaat boven de gebruikte aftestgrens ligt. Er kan daarom met grote zekerheid worden geconcludeerd dat de kandidaat de betreffende vaardigheid beheerst. Het adaptieve karakter van de toetsafname is zichtbaar in de positioneringen van kruisjes en cirkeltjes op de schaal. Het kruisje in een bepaalde kolom heeft doorgaans een positie op de schaal die dicht in de buurt ligt van het cirkeltje in de links daarvan gelegen kolom.




Figuur 4 Verloop van een adaptieve toetsafname bij een fictieve student.

Deze speciale manier om toetsen samen te stellen leidt tot aansprekende voordelen. Het belangrijkste voordeel is de grotere efficiëntie van het toetsproces. Omdat de moeilijkheidsgraad van elk aangeboden item nauwkeurig is afgestemd op de vaardigheid van de kandidaat, wordt de verhouding tussen informatie en meetfout gunstiger en kan met minder items dezelfde meetnauwkeurigheid worden bereikt als met een conventionele toets. In de literatuur wordt vaak gesproken over reducties in toetslengte van 40% of meer. (Wainer, Dorans, Eignor & Flaugher, 2000). Een ander voordeel is dat studenten niet geconfronteerd worden met toetsen die veel te moeilijk of veel te gemakkelijk zijn. Dat is extra belangrijk in situaties waar het niveau van de te meten vaardigheid gekenmerkt wordt door een grote spreiding in de doelgroep. Conventioneel geconstrueerde toetsen zijn in zo'n situatie voor grote aantallen kandidaten te moeilijk of te makkelijk. Dat leidt niet alleen tot gevoelens van frustratie of verveling maar ook tot onzuivere metingen aangezien een veel te moeilijk of veel te gemakkelijk item geen informatie toevoegt aan wat daarover al bekend was vóór de beantwoording van dat item. Nog een voordeel dat tot de verbeelding spreekt, is dat elke kandidaat een andere toets maakt. Het risico dat studenten informatie over toetsinhouden aan elkaar doorgeven wordt daarmee aanzienlijk gereduceerd en dat maakt de weg vrij voor een meer flexibele toetsplanning.

## WISCAT-pabo

De hiervoor besproken principes van adaptief toetsen zijn uitgewerkt in het toetspakket WISCAT-pabo dat binnenkort voor de pabo's beschikbaar komt. Het pakket wordt ontwikkeld door Cito in opdracht van en in samenwerking met zes pabo's van Fontys Hogescholen en Hogeschool Zuyd. WISCAT is een acroniem voor WISkunde/rekenen Computergestuurde Adaptieve Toetsing. Dit pakket kan op afroep een toets-op-maat samenstellen voor een student, de items presenteren op een beeldscherm, de antwoorden op de open of gesloten vragen nakijken en scoren, en een rapportage verzorgen voor de student en voor diens docent.

*Het toetsalgoritme*

In de vorige paragraaf is uitgelegd hoe optimaal meten met een adaptieve toets in zijn werk gaat. Bij WISCAT-pabo is het van belang dat het toetsalgoritme dat zorgdraagt voor de samenstelling van een adaptieve toets niet alleen let op de hoeveelheid informatie die het volgende aan te bieden item kan geven over de vaardigheid van de kandidaat. Omdat er een scoreprofiel zal worden gegeven, moeten bepaalde leerstofonderdelen met voldoende aantallen items in de toets vertegenwoordigd zijn. Dit betekent dat de vakinhoudelijke samenstelling van een toets onder controle moet worden gehouden, wat gebeurt door restricties op te leggen aan het toetsalgoritme. Het toetsalgoritme moet bij de selectie van items niet uitsluitend op de informatiewaarde van het item letten maar ook op de realisatie van de toetsmatrijs. De specificaties van die toetsmatrijs zijn:

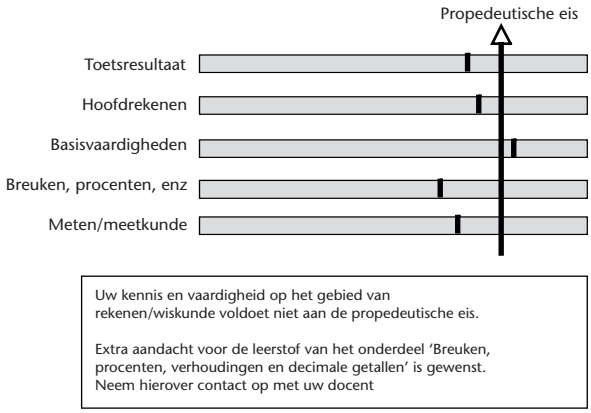
– de toetslengte bedraagt 50 items;
– de eerste 15 items zijn uit het domein hoofdrekenen en moeten binnen 15 minuten gemaakt worden. Hiervan komen er 7 uit het domein basisvaardigheden, 7 uit het domein verhoudingen, breuken, procenten en decimale getallen en 1 uit het domein meten en meetkunde;
– de overige 35 items komen volgens onderstaande verdeelsleutel uit de domeinen:
  - basisvaardigheden 8 items
  - verhoudingen, breuken, procenten en decimale getallen 8 items
  - meten en meetkunde 14 items
  - informatieverwerking, kans en statistiek, (woord)algebra,
  - verbanden, grafieken en functies 5 items

De items uit deze domeinen komen in willekeurige volgorde aan bod.

*Rapportage voor de student*

Zodra de laatste toetsopgave beantwoord is, krijgt de student de uitslag te zien op het beeldscherm. De resultaten worden gepresenteerd in de vorm van een scoreprofiel. Dat is een grafisch weergegeven overzicht van de behaalde resultaten op de hele toets en op de vier domeinen (zie figuur 5). De deelscore van elk domein is steeds gebaseerd op 15 items. Daartoe worden de scores op de items uit deel 1 (hoofdrekenen) en deel 2 (niet hoofdrekenen) die uit hetzelfde domein komen, samengenomen.

Een student kan eenvoudig aflezen of zijn kennis en vaardigheid voldoet aan de beheersingsstandaard (de propedeutische eis). Het resultaat op de hele toets is daarvoor beslis-




Figuur 5 Scorerapportage voor de student.

send. Tevens kan de student aflezen of de resultaten op de domeinen min of meer met elkaar 'in de pas lopen' en of er misschien sprake is van een serieuze achterstand ten opzichte van het resultaat op de hele toets. In het laatste geval verschijnt hierover een schriftelijke mededeling direct onder het scoreprofiel.

*Rapportage voor de docent*

De behaalde resultaten van studenten worden opgeslagen in een database waaruit een docent op verschillende manieren overzichten kan samenstellen. Ter illustratie van de mogelijkheden laat tabel 4 een voorbeeldrapportage zien van alle behaalde toetsresultaten van (fictieve) student R. Gimbrère. Op 28 september 2004 heeft deze student een toets afgelegd met onvoldoende resultaat. De *overall* schatting van 0,05 is fors lager dan de propedeutische eis van 0,33. Het vet weergegeven resultaat op domein Breuken, procenten, enzovoort geeft aan dat het resultaat op dit domein zodanig achterblijft bij de overall schatting dat extra aandacht voor dit leerstofonderdeel geboden is. Het evalueren van het achterblijven van deelschattingen ten opzichte van de overall schatting vindt plaats met een statistische toetsing en heeft zijn rationale in een praktische implicatie van het gebruikte testmodel, namelijk dat het voor de schatting van de vaardigheid niet uitmaakt welke set van items uit een geschaalde itembank daarvoor gebruikt is. Voor de schatting van de vaardigheid maakt het dus niet uit of daarvoor bijvoorbeeld hoofdrekenitems zijn gebruikt of items uit het domein Meten/meetkunde. Als dat wél zo is, is dat misschien een aanwijzing dat er sprake is van ernstige hiaten of misconcepties ten aanzien van het domein met de negatief afwijkende vaardigheidsschatting. In de maanden oktober en november 2004 heeft student Gimbrère veel geoefend en gerichte instructie ontvangen voor het domein Breuken, procenten, verhoudingen en decimale getallen. Op 1 december 2004 heeft hij opnieuw een toets afgelegd. Het resultaat op de hele toets is aanzienlijk verbeterd, maar onvoldoende om te voldoen aan de propedeutische eis. De schattingen voor alle deeltoetsen zijn ook hoger en de schatting voor de deeltoets Breuken, procenten, enzovoort blijft niet meer significant achter ten opzichte van de overall schatting. Al met al een gunstige ontwikkeling die vermoedelijk zal leiden tot het voldoen aan de propedeutische eis vóór het einde van het studiejaar.

Tabel 4 Voorbeeld van een overzichtsrapportage per student.

Resultaten voor: R. Gimbrère (fictieve student)

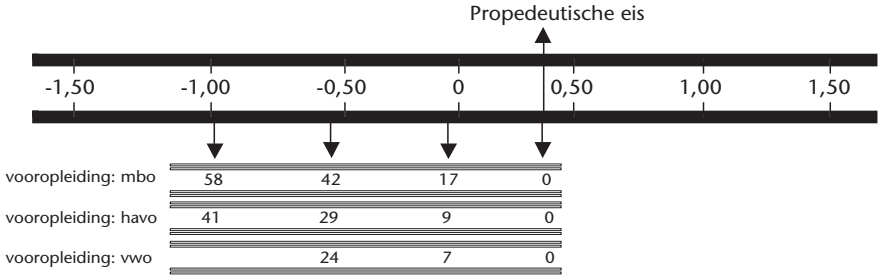
| Datum toetsafname | 28-09-04 | 01-12-04 |
|---|---|---|
| Propedeutische eis | 0,33 | 0,33 |
| | | |
| Toetsresultaat | 0,05 | 0,21 |
| Beslissing | onvold. | onvold. |
| domein: hoofdrekenen | 0,16 | 0,25 |
| domein: basisvaardigheden | 0,14 | 0,23 |
| domein: breuken, procenten, enz. | **-0,52** | 0,16 |
| domein: meten/meetkunde | 0,12 | 0,20 |

*Evaluatie van het wensenpakket*

Tot slot wordt nog eens naar de wensen gekeken die docenten rekenen/wiskunde van pabo's hebben geformuleerd ten aanzien van een toetsinstrumentarium. In hoeverre kan een toetspakket als WISCAT-pabo daaraan voldoen?

Het toetsinstrumentarium zou in de eerste plaats tegemoet moeten kunnen komen aan de grote verschillen in kennis en vaardigheid van instromende pabo-studenten op het gebied van rekenen/wiskunde. Het is een wijdverbreid gebruik om zich bij het samenstellen van een toets te laten leiden door het gemiddelde niveau van kennis en vaardigheid in de groep. Echter, als de spreiding van kennis en vaardigheid erg groot is, zoals in het geval van eerstejaars pabo-studenten, zal een op die manier samengestelde toets voor veel kandidaten te moeilijk en voor anderen juist te gemakkelijk uitvallen. Figuur 2 maakt dit duidelijk. De vaardigheidsverdeling van de studenten met een mbo-vooropleiding ligt veel meer naar links op de schaal dan die van studenten met een vwo-vooropleiding. Een toets die items bevat die optimaal zijn afgestemd op de gemiddelde vaardigheid van eerstejaars pabo-studenten zou voor mbo'ers over het algemeen te moeilijk en voor vwo'ers juist te gemakkelijk zijn. WISCAT-pabo stelt, zoals hiervoor uitgelegd, geheel automatisch voor elke individuele student een toets-op-maat samen en voorkomt daarmee dat studenten ontmoedigd raken door een te moeilijke toets of zich onvoldoende uitgedaagd voelen door een veel te eenvoudige toets.

Eenvoudige interpretatie van toetsscores was een tweede wens. WISCAT-pabo komt hieraan tegemoet door het aantal correct beantwoorde opgaven in een uit de itembank samengestelde toets te 'vertalen' naar een positie op de rekenvaardigheidsschaal. Omdat op deze schaal ook de propedeutische eis is gemarkeerd, is de betekenis van het behaalde toetsresultaat gemakkelijk af te lezen. Het op de rekenvaardigheidsschaal afgebeelde toetsresultaat kan eenduidig worden geïnterpreteerd in termen van 'voldaan' of 'niet voldaan' aan de propedeutische eis, maar ook in termen van de afstand tot het te bereiken doel. Voor dit laatste is informatie nodig over de tijd die studenten er gemiddeld over doen om bepaalde schaalafstanden te overbruggen. Omdat afnamegegevens steeds centraal worden opgeslagen, kunnen dergelijke gegevens na enige tijd verkregen worden en ter beschikking worden gesteld aan de gebruikers (zie figuur 6).




Figuur 6 Gemiddelde doorlooptijd in weken (fictieve data) voor de verwerving van een rekenvaardigheid conform de propedeutische eis.

De bijzondere eigenschappen van het gebruikte testmodel leiden ertoe dat de resultaten op alle uit de geschaalde itembank samen te stellen toetsen direct met elkaar vergeleken kunnen worden. De prestaties van studenten die verschillende toetsen hebben gemaakt, zijn zonder meer vergelijkbaar. Hetzelfde geldt voor de resultaten op twee of meer verschillende toetsen die één bepaalde student in de loop van het studiejaar gemaakt heeft. Gekoppeld aan een rapportage op leerstofonderdelen (domeinen) biedt dit mogelijkheden om de ontwikkeling van de rekenvaardigheid te volgen (derde wens) en tijdig gerichte maatregelen te nemen bij dreigende stagnatie.

De vierde wens had te maken met de rechtvaardiging van de zak-/slaagbeslissing. In hoeverre slaagt WISCAT-pabo erin om gerechtvaardigde beslissingen te nemen over zakken of slagen? Bij het nemen van beslissingen over studenten op basis van toetsresultaten kunnen twee fouten worden gemaakt:

1 De eerste fout doet zich voor als de gemeten vaardigheid van de student voldoet aan de propedeutische eis terwijl de ware vaardigheid niet voldoet aan de propedeutische eis. De student is dan ten onrechte geslaagd.
2 De tweede fout houdt in dat de gemeten vaardigheid van de student niet voldoet aan de propedeutische eis terwijl de ware vaardigheid wel daaraan voldoet. In dat geval wordt de student ten onrechte afgewezen.

Omdat de ware vaardigheid niet kan worden gekend, is de kwaliteit van zak-/slaagbeslissingen moeilijk te achterhalen. Simulatiestudies kunnen echter een uitweg bieden. Op basis van het gekozen testmodel kan een toetsafname worden gesimuleerd 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 om het resultaat van die vergelijking te classificeren in één van de cellen van tabel 5.

Tabel 5 Classificatie van vaardigheidsniveau ten behoeve van zak-/slaagbeslissingen.

| | | **Geschatte vaardigheid** | |
|---|---|---|---|
| | | < prop. eis | ≥ prop. eis |
| **Ware** | < prop. eis | correcte zakbeslissing | ten onrechte genomen slaagbeslissing |
| **vaardigheid** | ≥ prop. eis | ten onrechte genomen zakbeslissing | correcte slaagbeslissing |

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.

De uitgevoerde simulatiestudies hebben geleid tot het opstellen van een tabel voor beslissingsnauwkeurigheid (tabel 6) die geldig is voor alle toetsen die met WISCAT-pabo worden afgenomen. In de tabel staan percentages correcte en incorrecte beslissingen gebaseerd op toetsscores die onder het gebruikte testmodel gegenereerd zijn.

Tabel 6 Percentage correcte en incorrecte beslissingen op basis van WISCAT-pabo-resultaten.

| | | **Geschatte vaardigheid** | |
|---|---|---|---|
| | | onvoldoende | voldoende |
| **Ware vaardigheid** | onvoldoende | 50,6 | 4,6 |
| | voldoende | 4,2 | 40,6 |

De tabel laat zien dat in ruim 90% van alle gevallen een correcte beslissing genomen wordt en dat de twee soorten verkeerde beslissingen ongeveer even vaak zullen voorkomen. Zoals gezegd geven deze resultaten de theoretische beslissingsnauwkeurigheid weer. In werkelijkheid zullen personen zich niet altijd conform het gebruikte testmodel gedragen. De percentages zijn daarom niet meer dan een indicatie voor de omvang van de beslissingsfouten die in reële toepassingssituaties te verwachten zijn.

De laatst geformuleerde wens ging over het gebruiksgemak. Goede toetsen construeren kost veel tijd. Deze afnemen, de resultaten bepalen en bespreken en naar aanleiding daarvan een passend remediëringstraject opzetten en begeleiden, zijn evenzeer arbeidsintensieve activiteiten. Bij een computergestuurd adaptief toetspakket hoeft een docent zich niet meer te bekommeren om toetsconstructie en om administratieve activiteiten rondom toetsafnames. Daarmee blijft meer tijd over voor taken die niet of nauwelijks te automatiseren zijn, zoals het opsporen en wegnemen van misconcepties of leerbelemmeringen.

## Tot slot

De discussie over de kwaliteit van de lerarenopleiding primair onderwijs is in volle hevigheid losgebarsten nu recente visitaties duidelijk hebben gemaakt dat er sprake is van 'overladenheid zonder diepgang' en het 'dringend geboden is het algemene niveau en de geletterdheid van studenten te verhogen' (HBO-raad, 2003). Voor velen kwamen die conclusies niet als een verrassing; in de landelijke pers werden de al of niet vermeende tekortkomingen van pabo-studenten al langer breed uitgemeten. Zonder de juistheid ervan in twijfel te willen trekken, is het zaak om te constateren dat deze conclusies over het algemeen getrokken zijn zonder verwijzing naar een helder gedefinieerde prestatiestandaard. Dit artikel beschrijft een mogelijke bijdrage aan de kwaliteitsverbetering van de pabo's door de ontwikkeling van een gestandaardiseerd toetspakket om de reken-

vaardigheid van eerstejaars studenten in beeld te brengen, te volgen en te evalueren op grond van een nauwkeurig gedefinieerde prestatiestandaard. De in dit artikel gekozen en beargumenteerde prestatiestandaard heeft bijvoorbeeld als consequentie dat, naar verwachting, bijna 50% van de instromers met een havo-diploma hieraan niet zal voldoen. Het op deze wijze inzichtelijk maken van de gevolgen van een gekozen prestatiestandaard kan een bijdrage leveren aan de discussie over het gewenste niveau van rekenvaardigheid van instromende pabo-studenten.

Het mag duidelijk zijn dat deze aanpak ook gevolgd kan worden voor andere vakken of vakonderdelen waarop de eigen vaardigheid van studenten vaak achterblijft bij wat gewenst wordt, zoals bijvoorbeeld spelling en wereldoriëntatie.

## Referenties

HBO-raad (2003) *Moed tot meesterschap.* Den Haag: HBO-raad.

Ministerie van Onderwijs en Wetenschappen (1987). *Het onderwijs in het vakgebied 'Rekenen en Wiskunde' op de PABO. Rapport aan de Minister van Onderwijs en Wetenschappen opgesteld door de Inspectie Hoger Onderwijs, geleding Opleidingen Onderwijsgevenden.* Inspectierapport 11. 's-Gravenhage: MOW.

OCW (2005) *Meer kwaliteit en differentiatie: de lerarenopleidingen aan zet. Beleidsagenda lerarenopleidingen 2005-2008.* http://www.minocw.nl/brief2k/2005/doc/27721b.pdf

Verhelst, N.D. (1993) Itemresponstheorie. In: T.J.H.M. Eggen & P.F. Sanders (Red.), *Psychometrie in de Praktijk.* Arnhem: Cito Instituut voor Toetsontwikkeling.

Wainer, H., Dorans, N.J., Eignor, D. & Flaugher, R. (2000) *Computerized Adaptive Testing: a primer* (2nd ed.). Mahwah: Lawrence Erlbaum.