De tijd is rijp om in het onderwijs gebruik te maken van de voordelen die adaptieve toetsing biedt.

Adaptieve toetsing: van gimmick naar kroonjuweel

Gerard Straetmans
Dr. G.J.J.M. Straetmans is gepensioneerd lector assessment. Ruim veertig jaar heeft hij zich in allerlei werkverbanden ingezet voor de kwaliteitsverbetering van toetsing en examinering in het onderwijs.
E-mail: gerkit@fo.nl

Cor Sluijter
Dr. C. Sluijter is onder meer docent bij de Master Toetsdeskundige van Fontys, vice-president van de Association for Educational Assessment Europe en oud-voorzitter van de NVE. E-mail: c.sluijter@fontys.nl

De tijd lijkt rijp om een begin te maken met grootschalige implementatie van adaptieve toetsing in het onderwijs. De toegenomen behoefte aan objectieve informatie over leervorderingen en de groeiende wens om het onderwijs verder te flexibiliseren zorgen ervoor dat er steeds meer getoetst wordt. Het construeren, afnemen en nakijken van de toetsen en het bespreken van de toetsresultaten met de kandidaten is daardoor steeds meer tijd gaan kosten. Niet meer dan logisch dat er wordt gezocht naar een manier om de toetsing efficiënter te laten verlopen. Adaptieve toetsing voorziet in die behoefte. Er zijn inmiddels enkele systemen operationeel. Bij lang niet iedereen in de toetsgemeenschap is bekend wat adaptieve toetsing precies inhoudt en hoe het werkt. Deze bijdrage gaat in op het concept adaptieve toetsing en maakt duidelijk hoe het werkt en welke voor- en nadelen er zijn.

Inleiding

In 1980 liep de eerste auteur van deze bijdrage stage bij het Cito in Arnhem. Het stagedoel behelsde het meewerken aan de ontwikkeling van een opgavenbank voor het vak Onderwijskundige Basiskennis van een tweedegraads lerarenopleiding en het uitvoeren van onderzoek ten dienste van die ontwikkeling. Een van de onderzoeksactiviteiten was het beproeven van de kwaliteit van de ontwikkelde opgaven bij een steekproef uit de doelgroep. Daarbij werd niet alleen gebruik gemaakt van technieken afkomstig uit de klassieke testtheorie (KTT), maar ook uit wat destijds nog latente-trektheorie heette. Een op dat moment nieuwe ontwikkeling binnen de psychometrie die inmiddels bekend staat als itemresponstheorie (IRT).

Naast nieuwe inzichten over het meten van menselijke mentale eigenschappen, bood de nieuwe theorie diverse mogelijkheden voor verbetering van de toetspraktijk (Van der Linden, 1983). Op grond van deze theorie werd het onder andere mogelijk 'toetsen-op-maat' of 'adaptieve toetsen', zoals ze later werden genoemd, te ontwikkelen en af te nemen. Sommige universitaire vakgroepen hadden voor onderzoeksdoeleinden een kleine opgavenbank beschikbaar waaruit met behulp van een algoritme dergelijke adaptieve toetsen konden

worden samengesteld en afgenomen. Vol trots demonstreerden de ontwikkelaars de mogelijkheden op congressen en studiedagen, maar aangezien de toehoorders op hun werk vaak (nog) niet beschikten over opgavenbanken, laat staan over de benodigde computerfaciliteiten, dachten ze niet dat de toepasbaarheid voor de onderwijspraktijk groot zou zijn. Een leuke gimmick, dat wel.

Nu, ruim 40 jaar later, is de tijd rijp om in het onderwijs gebruik te maken van de voordelen die adaptieve toetsing biedt. Door de toegenomen behoefte aan objectieve informatie over leervorderingen en de wens om het onderwijs steeds meer te flexibiliseren wordt er meer getoetst dan vroeger. Het construeren, afnemen en nakijken van de toetsen en het bespreken van de toetsresultaten met de kandidaten is daardoor steeds meer tijd gaan kosten. Niet meer dan logisch dat er wordt gezocht naar een manier om het toetsproces efficiënter te laten verlopen. Digitale toetsing voorziet tot op zekere hoogte al in die behoefte, maar adaptieve toetsing heeft als bijzondere toepassing van digitale toetsing nog meer te bieden, zoals verderop in dit artikel duidelijk zal worden. Er zijn inmiddels enkele systemen operationeel: in verschillende onderwijssectoren, in verschillende domeinen en voor het nemen van diverse onderwijskundige beslissingen, zoals plaatsing, vaststelling van de voortgang, diagnostisering, afsluiting/certificering en onderwijsevaluatie. Toch weten maar weinigen precies wat adaptieve toetsing inhoudt en hoe het werkt. In deze bijdrage zullen we daarom het concept adaptieve toetsing en de werking ervan uitleggen. Ook staan we stil bij de vaardigheidsschaal, de sine qua non van adaptieve toetsing, en bij de voor- en nadelen die deze toetsmethodiek met zich meebrengt. Verder bespreken we twee adaptieve toetsen die in het Nederlandse onderwijs op grote schaal worden toegepast voor het nemen van verschillende onderwijskundige beslissingen. En we eindigen met een pleidooi voor nog meer grootschalige toepassingen.

Adaptieve toetsing is ‘le raison d’être’ van de itemresponstheorie.

Adaptieve toetsing: wat is het en hoe werkt het?

Een oud principe

Het principe van adaptieve toetsing is beslist niet nieuw. Bij mondelinge examinering werd en wordt het, vaak onbewust, toegepast. Zo zal een verstandige examinator, wanneer blijkt dat een gestelde vraag te moeilijk of te makkelijk is geweest, de volgende te stellen vraag respectievelijk makkelijker of moeilijker laten zijn. De reden is dat het stellen van te moeilijke of te makkelijke vragen onvoldoende informatie oplevert om een goed beeld te krijgen van het vaardigheidsniveau van de kandidaat (Wainer et al., 2000). Het volgende voorbeeld, afkomstig uit Straetmans & Eggen (2007a, p.19), maakt duidelijk dat afstemming van de moeilijkheidsgraad van een taak op de respons van een kandidaat op een voorgaande taak, bevorderlijk is voor de hoeveelheid

informatie die daarmee wordt verkregen over het vaardigheidsniveau van de kandidaat. “Stel u voor dat u een gymdocent bent en dat u wilt vaststellen hoe hoog de brugklasleerlingen in uw groep kunnen springen. Het ligt voor de hand dat u eerst een grove inschatting maakt van de capaciteit van de betreffende leerlingen op dit gebied. U gebruikt daarvoor de vuistregel dat langere personen over het algemeen hoger kunnen springen dan korte personen. Op grond daarvan komt u tot de conclusie dat het vermoedelijk zinloos is om de betreffende leerlingen te laten springen over lathoogtes lager dan 60 cm en hoger dan 160 cm. Immers, de uitslag van de sprong over dergelijke latposities is zeer voorspelbaar en zal dus niet of nauwelijks bijdragen aan uw kennis over de hoogspringcapaciteit van de betreffende leerling. U kiest ervoor om in het midden van het interval 60-160 cm te beginnen, namelijk op 110 cm. U neemt de leerling tijdens diens poging over de lat te springen nauwkeurig waar en signaleert dat die er ruimschoots overheen springt. U heeft uit deze eerste sprong veel informatie gekregen, namelijk dat de leerling waarschijnlijk nog een flink stuk hoger kan springen dan 110 cm. U besluit daarom de lat op 130 cm te leggen. De sprong over deze lathoogte mislukt maar net. U concludeert hieruit dat de capaciteit van de leerling dichter bij 130 cm zal liggen dan bij 110 cm en legt de lat vervolgens op 125 cm. Als de leerling hier overheen springt, concludeert u dat de hoogspringcapaciteit van deze leerling ergens ligt tussen de 125 en 130 cm. Met deze schatting bent u tevreden en u beëindigt het assessment voor deze leerling.” Vergelijk deze procedure met een standaardassessment hoogspringen waarbij de gymdocent leerlingen over lathoogtes vanaf 60 cm laat springen en bij iedere geslaagde poging de lat 5 cm hoger legt. Bij de adaptieve variant zullen er in de regel aanzienlijk minder sprongen nodig zijn om vast te stellen hoe hoog iedere leerling kan springen. Dat komt omdat in de adaptieve procedure de lathoogte voor de volgende sprong telkens wordt afgestemd op de resultaten die een leerling behaalde bij eerdere sprongen (lathoogtes). Het resultaat van een sprong over een lathoogte die dicht in de buurt ligt van de hoogspringcapaciteit levert meer informatie op over die hoogspringcapaciteit dan een sprong over een lathoogte ver onder of boven de hoogspringcapaciteit. De adaptieve procedure kan daardoor bij de meeste leerlingen veel sneller tot een conclusie leiden over hun hoogspringcapaciteit en is daardoor veel efficiënter.

Noodzaak van een gemeenschappelijke schaal

Voor de onderlinge afstemming van moeilijkheidsgraad en vaardigheid is het nodig dat deze gemeten eigenschappen van respectievelijk opgaven/taken en personen op een en dezelfde schaal liggen en dus te beschrijven zijn met dezelfde meeteenheid. Bij het zojuist gegeven voorbeeld werden lathoogte – moeilijkheidsgraad − en hoogspringcapaciteit – vaardigheid − beide uitgedrukt in centimeters. Maar bij toetsing van kennis zijn fysieke metingen niet aan de orde. De klassieke testtheorie gebruikt de ruwe score − het aantal behaalde punten op een toets − als een maat voor het prestatieniveau van een persoon. En de p-waarde − de gemiddelde score van een groep personen op een opgave gedeeld door de maximaal haalbare score van de betreffende opgave − als een maat voor de moeilijkheidsgraad van die opgave. Anders dan bij het voorbeeld van het hoogspringen is het bij toetsing van kennis lastig om op basis van de op een zeker moment behaalde score te bepalen welke p-waarde een volgende opgave idealiter zou moeten hebben. Niet alleen omdat het andere meeteenheden betreft, maar ook omdat de p-waarde van een opgave, zelfs als die berekend zou zijn op grond van de antwoorden van een voor de doelgroep perfect representatieve groep personen, niet te gebruiken is als indicator van de moeilijkheidsgraad voor *individuele* personen. Zo is de kans dat een uiterst vaardige kandidaat een correct antwoord geeft op een opgave met

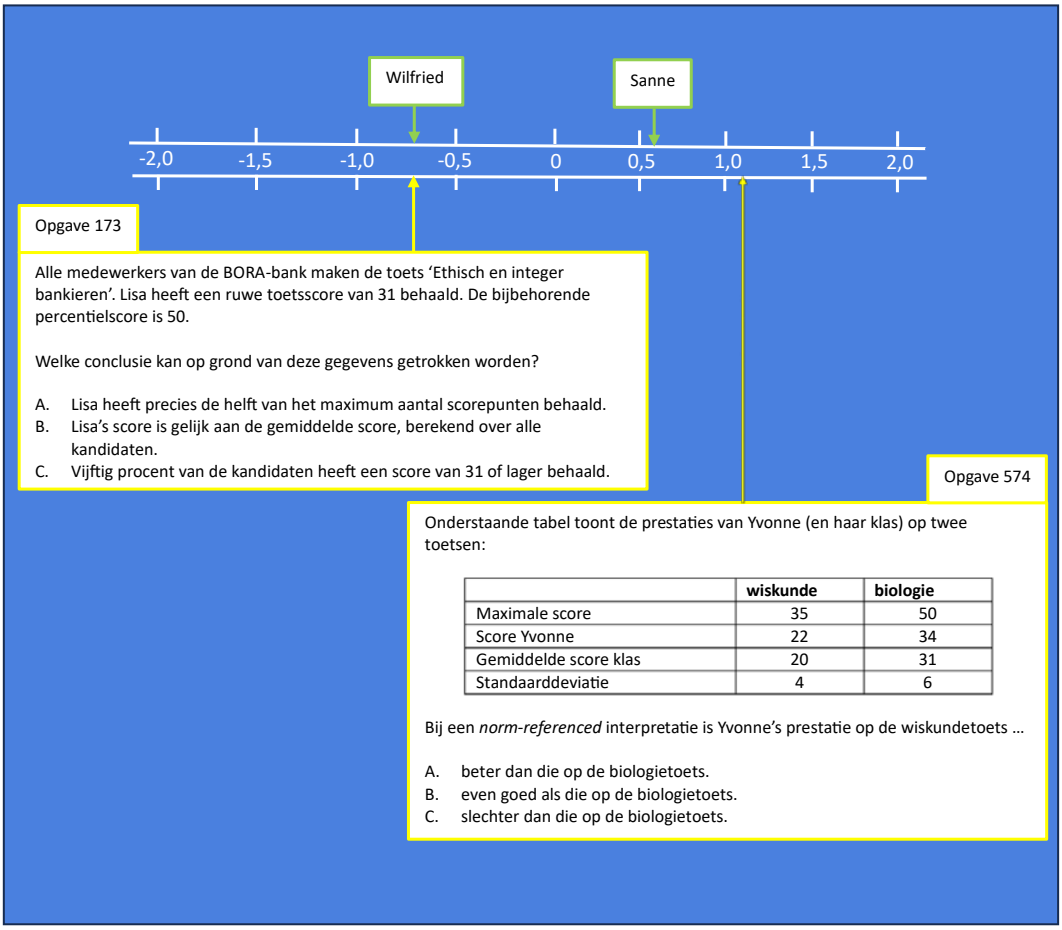
een p-waarde van 0,75, groter dan 0,75. En voor personen met een geringe vaardigheid zal de kans op een correct antwoord kleiner zijn dan 0,75 (Suen, 1990, p.84).

IRT-vaardigheidsschaal

Genoemde problemen zijn op te lossen door gebruik te maken van principes en procedures uit de IRT (Verhelst, 1993). Anders dan de KTT, die zich alleen richt op de observeerbare toetsscore van een kandidaat, gaat de IRT uit van het bestaan van een niet direct waarneembare – latente – psychologische eigenschap die ten grondslag ligt aan de geobserveerde toetsprestatie van een kandidaat (Odendahl, 2011, p.118). Voorbeelden van dergelijke psychologische eigenschappen in het onderwijs zijn: rekenvaardigheid, spellingsvaardigheid, luistervaardigheid, digitale vaardigheid, enzovoorts Zo'n latente eigenschap wordt gerepresenteerd door een numerieke schaal waarop zowel de vaardigheid van personen als de moeilijkheid van opgaven af te beelden zijn. Zie het tekstkader voor een summiere uitleg over de statistische uitwerking hiervan.

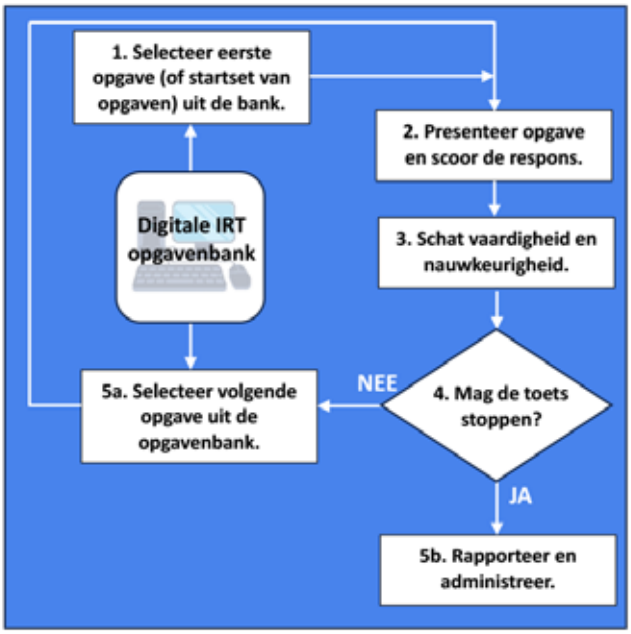
Figuur 1 illustreert het principe van een dergelijke numerieke gemeenschappelijke schaal voor personen (qua vaardigheid) en opgaven (qua moeilijkheidsgraad). De kandidaten Wilfried en Sanne hebben een toets gemaakt die uit de opgavenbank 'Onderwijskundig meten' is samengesteld en hun op basis van IRT-procedures geschatte vaardigheden zijn respectievelijk -0,7 en 0,6. De geschatte moeilijkheden van de opgaven in de bank liggen ook op deze schaal. Zo

Een IRT-vaardigheidsschaal is het resultaat van een zogenoemd kalibratie-onderzoek waarin een representatieve steekproef van personen uit de doelgroep van de toets opgaven (die een operationalisatie zijn van de te toetsen eigenschap) beantwoordt. Voorafgaand aan dit onderzoek heeft de onderzoeker een model gekozen dat een goede beschrijving geeft van de relatie tussen de moeilijkheidsgraad van een opgave en de vaardigheid van een persoon. In de IRT zijn diverse mathematische modellen ontwikkeld die de kans op een correct antwoord geven als een functie van de vaardigheid van de persoon en een of meer eigenschappen, zoals bijvoorbeeld de moeilijkheidsgraad, van opgaven. Bekende modellen zijn onder andere het Rasch (1960) model, het Birnbaum 2- en 3-parameter model (Birnbaum, 1968) en in Nederland het Verhelst-Eggen model (Adèr, Mellenbergh & Hand, 2008), dat ook wel bekend staat als het One Parameter Logistic Model (Glas & Verhelst, 1993). De hier genoemde bronnen stellen hoge eisen aan de wiskundige kennis van lezers. Een toegankelijker geschreven werk over de basisprincipes van IRT vindt men bij Baker (1985).
Statistische analyse van de in het kalibratie-onderzoek verzamelde gegevens maakt vervolgens duidelijk hoe het gesteld is met de modelpassing van iedere opgave. Voor alle opgaven waarvan het gekozen model een goede beschrijving en voorspelling geeft van de antwoorden van de personen, wordt de relatieve moeilijkheidsgraad (aangeduid als ß) geschat; wederom door middel van een statistische procedure. De opgaven met een onvoldoende modelpassing worden niet opgenomen in de resulterende opgavenbank. Anders dan de p-waarde is ß een indicator van de moeilijkheidsgraad van een opgave voor elke persoon in de doelgroep, ongeacht zijn of haar vaardigheid. Ordening van de opgaven op hun ß-waarde levert een numerieke schaal op, waarop ook de positie van een persoon te schatten is op grond van zijn of haar score op een reeks opgaven afkomstig uit de gekalibreerde opgavenbank. Het schaalbegrip houdt in dit verband in dat een kandidaat die een bepaalde opgave correct beantwoordt een steeds grotere kans zal hebben op een correct antwoord, naarmate opgaven lagere schaalwaarden hebben. Evenzo zal diezelfde kandidaat een steeds kleinere kans hebben om correct te antwoorden naarmate de schaalwaarde van opgaven toeneemt.



Figuur 1. Opgaven en studenten op een (fictieve) IRT-schaal 'Onderwijskundig meten'.

heeft bijvoorbeeld opgave 173 dezelfde positie op de schaal als Wilfried, namelijk -0,7. Volgens het door de ontwikkelaars van de opgavenbank gekozen IRT-model betekent dit dat Wilfried 50 procent kans heeft om deze opgave correct te beantwoorden. Maar de kans dat Wilfried opgave 574 correct zal beantwoorden is veel kleiner dan 50 procent, want deze opgave heeft een moeilijkheidsgraad van 1,15. De moeilijkheidsgraden van de opgaven in de bank zijn, zoals uitgelegd in het tekstkader, geldig voor alle personen uit de doelgroep van de IRT-geschaalde opgavenbank. De moeilijkheidsgraden zijn relatief bepaald ten opzichte van de andere opgaven in de bank, maar zijn onafhankelijk van de vaardigheid van de personen in de doelgroep. En hetzelfde geldt voor de vaardigheidsschattingen van personen. Die zijn onafhankelijk van de uit de opgavenbank samengestelde toetsen. Wilfried krijgt, uitgezonderd een verschil als gevolg van meetfout, dezelfde geschatte vaardigheid na het maken van een relatief gemakkelijke als na het maken van een relatief moeilijke toets. De nauwkeurigheid van die schattingen kan wel verschillen, afhankelijk van de informatie die de opgaven in beide toetsen bijdragen. 'Informatie' moet hier in statistische zin worden opgevat als de bijdrage die een opgave kan leveren aan de nauwkeurigheid van de vaardigheidsschatting (Fisher, 1922). Naarmate de moeilijkheidsgraad van een opgave op de schaal dichter bij de geschatte



Figuur 2. Interacties van CAT-componenten tijdens een toetsafname.

vaardigheid van een persoon ligt, zal die opgave meer informatie leveren. De moeilijkheidsgraad van bijvoorbeeld opgave 173 is voor Wilfried en Sanne gelijk, alleen de informatie die deze opgave in statistische zin levert over hun vaardigheid verschilt. Die is voor de schatting van Wilfrieds vaardigheid veel groter dan voor de schatting van de vaardigheid van Sanne.

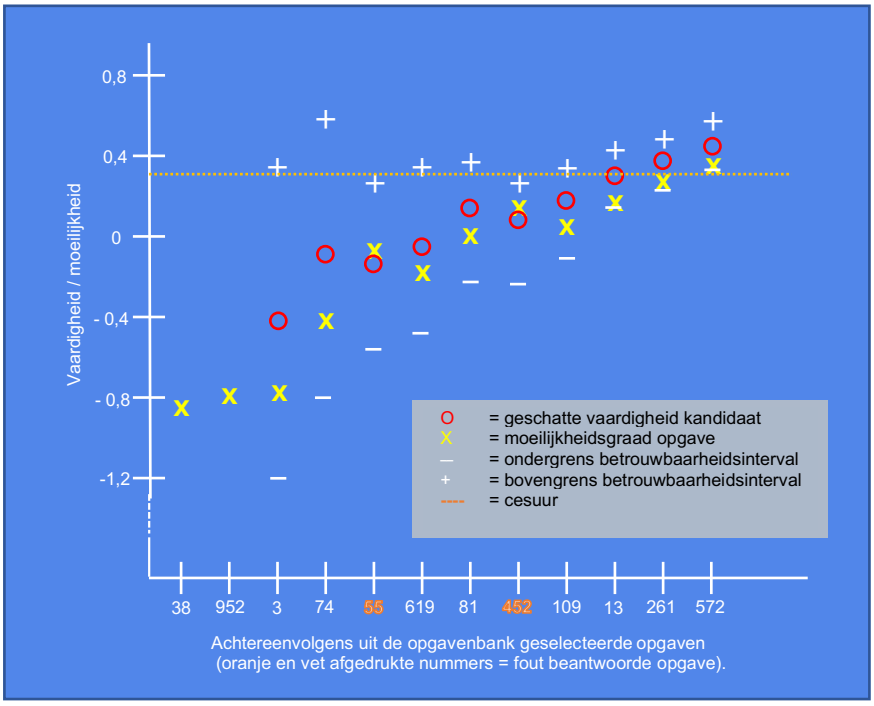
Inzet van een computer

Wanneer een kandidaat begint aan een adaptieve toets is nog onbekend uit welke opgaven die toets zal bestaan, want afname en samenstelling zijn bij adaptieve toetsing gelijktijdige processen. Na elk nieuw antwoord van de kandidaat volgt op grond van alle tot dan gegeven antwoorden een schatting van de positie van de kandidaat op de vaardigheidsschaal. En op basis daarvan selecteert het algoritme een volgende opgave uit de bank. Daarvoor zijn verschillende methoden voorhanden (Van der Linden & Pashley, 2010). De meest gebruikelijke is het selecteren van een opgave met een moeilijkheidsgraad die zo dicht mogelijk ligt bij de geschatte positie van de kandidaat op de vaardigheidsschaal. De daarvoor benodigde berekeningen vergen de inzet van een computer. Vandaar dat een adaptieve toets vaak wordt aangeduid met het acroniem CAT, wat staat voor Computerized Adaptive Test, of in het Nederlands Computergestuurde Adaptieve Toets.

Hoe werkt het?

Een CAT is een softwarepakket dat uit de volgende componenten bestaat: a) een digitale, IRT-gekalibreerde opgavenbank; (b) een afnamemodule die opgaven één voor één op een beeldscherm presenteert en de gegeven antwoorden van een kandidaat scoort als goed of fout, (c) een algoritme dat de start, voortgang en beëindiging van een toets regelt en (d) een rapportagemodule die het toetsresultaat kenbaar maakt aan de kandidaat en alle afnamegegevens administreert. Figuur 2 geeft schematisch weer hoe deze componenten samenwerken tijdens een adaptieve toetsafname.

In dit artikel beperken we ons tot een bespreking van één component: het algoritme. Figuur 3 geeft grafisch weer hoe het algoritme het



Figuur 3. Verloop van een adaptieve toetsafname bij een fictieve kandidaat.

toetsafnameproces reguleert (Straetmans & Eggen, 1998). Op de horizontale as staan de opgaven die achtereenvolgens geselecteerd en beantwoord zijn. Op de verticale as worden zowel de geschatte vaardigheden van de persoon (aangegeven met cirkeltjes) als de moeilijkheidsgraden van de geselecteerde opgaven (aangegeven met kruisjes) afgebeeld. De gestippelde lijn die parallel loopt aan de horizontale as is de cesuur, die aangeeft hoe vaardig kandidaten moeten zijn voor een positieve uitslag op de toets.

Bij de start kan het toetsalgoritme nog geen opgave selecteren die het beste past bij de vaardigheid van de kandidaat, omdat er dan nog onvoldoende bekend is over die vaardigheid. Aselect trekken van de eerste opgave is een veel gekozen oplossing voor dit probleem. Maar er zijn ook andere mogelijkheden, zoals bijvoorbeeld het selecteren van een opgave met een gemiddelde moeilijkheid. In het voorbeeld van Figuur 3 zijn de eerste drie opgaven aselect getrokken uit een deelverzameling van relatief eenvoudige opgaven. Op die manier kan men geforceerd eenvoudige opgaven aanbieden bij de start van de toets, bijvoorbeeld om eventuele toetsangst te reduceren.

Na beantwoording van de derde opgave volgt de eerste vaardigheidsschatting. Na slechts drie beantwoorde opgaven is die natuurlijk nog niet erg nauwkeurig, omdat er statistisch gezien nog niet veel informatie is verzameld. Het programma schat behalve de vaardigheid ook de gemaakte meetfout en gebruikt die om een betrouwbaarheidsinterval af te bakenen rondom de geschatte vaardigheid. Het min- en plusteken staan voor respectievelijk de onder- en bovengrens van dit betrouwbaarheidsinterval. Het betrouwbaarheidsinterval geeft met een in te stellen zekerheid, bijvoorbeeld 90 procent, aan dat de ware vaardigheid van de kandidaat tussen de aangegeven onder- en bovengrens zal liggen. Duidelijk te zien is dat de nauwkeurigheid van de vaardigheidsschattin-

gen snel toeneemt naarmate er meer opgaven beantwoord zijn. Vanaf de vierde opgave is het toetsproces adaptief; na elk antwoord volgt een nieuwe vaardigheidsschatting op basis van alle tot dan gegeven antwoorden. Alle nog niet in de toets gebruikte opgaven in de bank worden geordend naar de hoeveelheid statistische informatie die ze kunnen geven over de meest recente vaardigheidsschatting. De opgave met de grootste informatiewaarde wordt geselecteerd als volgende te presenteren opgave, tenzij andere selectieregels dat verhinderen. Zo zijn er regels die ervoor zorgen dat een adaptieve toets inhoudelijk op elk moment zoveel mogelijk een afspiegeling is van de te toetsen leerdoelen, content-balancing genoemd (Kingsbury & Zara,1989). En ook zijn er regels die verhinderen dat bepaalde opgaven te vaak (over-exposure) of bijna nooit (under-exposure) geselecteerd worden (Mills & Stocking, 1996). Over-exposure vergroot het risico dat opgaven bekend worden onder potentiële kandidaten en under-exposure is zonde van het geld en de tijd benodigd voor het ontwikkelen en beproeven van de opgaven.
Er zijn diverse manieren om een adaptieve toetsafname te beëindigen. De meest eenvoudige manier is stoppen na het aanbieden van een bepaald aantal opgaven. In het voorbeeld van figuur 3 is er echter gekozen voor een dynamische stopregel. Dat wil zeggen dat de adaptieve toets stopt zodra het betrouwbaarheidsinterval om de meest recente vaardigheidsschatting in zijn geheel onder of boven de cesuur ligt. Bij deze stopregel varieert het aantal opgaven dat een kandidaat voorgelegd krijgt. In het voorbeeld is dat al na beantwoording van de twaalfde opgave (met nummer 572), omdat dan met 90 procent zekerheid geconcludeerd kan worden dat de ware vaardigheid van deze kandidaat boven de cesuur ligt en de kandidaat de betreffende leerdoelen dus voldoende beheerst. Het adaptieve karakter van de toetsafname is zichtbaar in de positioneringen van kruisjes en cirkeltjes op de schaal. Het kruisje (moeilijkheidsgraad opgave) in een bepaalde kolom heeft doorgaans een positie op de schaal die dicht in de buurt ligt van het cirkeltje (vaardigheidsschatting) in de links daarvan gelegen kolom.

Belangrijkste voordelen CAT

Ten opzichte van conventionele toetsing heeft CAT verschillende voordelen te bieden voor toetsontwikkelaars, toetsgebruikers en kandidaten. De belangrijkste worden hieronder kort besproken.

- *Grotere efficiëntie*. Het meest aansprekende voordeel van CAT is de grotere efficiëntie vergeleken met meer conventionele toetsen. Terwijl bij conventionele toetsen de moeilijkheidsgraad van de opgaven vaak wordt afgestemd op de gemiddelde kandidaat, richt een CAT zich bij het bepalen van de moeilijkheidsgraad van de opgaven op het niveau van de individuele kandidaat. Hierdoor kan een CAT de gewenste meetnauwkeurigheid bereiken met ongeveer de helft van het aantal opgaven dat nodig is in een conventionele toets (Vispoel et al., 1994; Wainer et al., 2000). Bovendien geldt die gewenste meetnauwkeurigheid voor alle kandidaten. Terwijl bij een conventionele – lineaire − toets de geschatte betrouwbaarheid minder geldig is voor kandidaten met relatief lage en hoge scores. De efficiëntie van CAT is een belangrijke overweging in situaties waar tijd beperkt is. Wanneer echter sprake is van zwaarwegende beslissingen over kandidaten, is het verstandiger de grotere efficiëntie te benutten door de meetnauwkeurigheid te vergroten en reductie van de toetslengte achterwege te laten.
- *Ondersteunt flexibilisering van het onderwijs.* In traditioneel onderwijs werken alle onderwijsdeelnemers op hetzelfde moment aan dezelfde leerdoelen en krijgen ze op hetzelfde moment dezelfde toets voorgelegd, waaruit moet blijken wie de leerdoelen in voldoende mate beheerst. Dit dient zowel de objectiviteit als de efficiëntie (slechts één toetsversie nodig)

van de toetsing. Nu steeds meer onderwijsinstellingen de wens hebben om het onderwijs te flexibiliseren en deelnemers toe te staan in hun eigen tempo door het curriculum te gaan, moet deze strategie verlaten worden. CAT maakt dit mogelijk. In principe kan een deelnemer op elk willekeurig moment een toets maken. Immers, doordat elke deelnemer in principe een andere toets maakt, wordt het risico van het aan elkaar doorgeven van toetsopgaven sterk gereduceerd, terwijl de behaalde toetsresultaten onderling toch vergelijkbaar zijn.

- *Computergestuurd.* Adaptieve toetsing is computergestuurd en dat brengt voordelen die een toets op papier niet kan bieden. Zo zijn er administratieve voordelen die bijvoorbeeld inhouden dat er geen toetsboekjes geproduceerd hoeven te worden en dat het nakijken, scoren en administreren van de resultaten geautomatiseerd is. Verder zijn er toetstechnische voordelen te behalen doordat de computer (a) geluid en bewegende beelden kan toevoegen aan een toetsopgave, (b) kan interacteren met de kandidaat en (c) responsen van uiteenlopende aard kan accepteren. Dit vergroot de mogelijkheden voor het toetsen van hogere cognitieve vaardigheden.

Door de toenemende beschikbaarheid en betere toepasbaarheid van algoritmes zal adaptieve toetsing steeds eenvoudiger te implementeren zijn.

Belangrijkste nadelen CAT

CAT biedt aantrekkelijke voordelen boven conventionele toetsing, maar nadelen zijn er ook.

- *Hoge ontwikkel- en onderhoudskosten.* Een vereiste voor CAT is een IRT-gekalibreerde opgavenbank. Kalibratie houdt in dat het functioneren van alle opgaven in de bank op voorhand onderzocht moet worden bij een representatieve steekproef van proefpersonen uit de doelgroep. Afhankelijk van het gekozen IRT-model zijn hiervoor 150 tot 500 proefpersonen per opgave nodig. De kennis en vaardigheden die voor het uitvoeren van dergelijk onderzoek nodig zijn ontbreken bij de meeste docenten, zodat dure expertise moet worden ingekocht. Dat laatste geldt ook voor het onderhoud van de opgavenbank. Onderzoek naar itemdrift (Is de moeilijkheid van opgaven over tijd veranderd, bijvoorbeeld doordat die opgaven bekend zijn geworden bij potentiële kandidaten?), modelpassing (Is de modelpassing in de loop van de operationele fase veranderd en zo ja, in positieve of negatieve zin?) en inhoudelijke passing (Zijn de leerdoelen, de selectieprocedure van het algoritme en de samenstelling van de gegenereerde toetsen congruent?) is eveneens werk voor psychometrici en toetsdeskundigen.
- *Moeilijk te doorgronden scoring.* Bij con-

ventionele toetsing wordt de toetsscore via een of andere eenvoudige transformatie uitgedrukt in een schoolcijfer of zak-/slaagbeslissing (Sluijter & Straetmans, 2024). Voor onderwijsdeelnemers is de relatie tussen de toetsscore en de uiteindelijke onderwijskundige beslissing meestal inzichtelijk en mede daardoor acceptabel. Bij adaptieve toetsing is die relatie aanzienlijk complexer en daardoor minder inzichtelijk en vaak ook minder acceptabel. Als twee onderwijsdeelnemers (een vaardige, hierna aangeduid als A en een weinig vaardige, hierna aangeduid als B) een CAT maken (A zal dan waarschijnlijk een toets maken die gemiddeld moeilijker is dan B) en ze beiden 30 van de 50 opgaven correct hebben beantwoord, dan zal de geschatte vaardigheid van A waarschijnlijk aanzienlijk groter zijn dan die van B. Immers, de geschatte vaardigheid hangt grotendeels af van de moeilijkheidsgraad van de aangeboden opgaven in de toets. Ook voor de docenten, die meestal niet bekend zijn met de principes van de IRT, is het vaak onduidelijk hoe zak-/slaagbeslissingen gerelateerd zijn aan de toetsprestaties. Zowel toetsgebruikers als kandidaten dienen door middel van een instructie te worden voorbereid op de voor hen ongebruikelijke scoring.

- *Beperkte vrijheid kandidaat.* Bij conventionele toetsen, ongeacht of ze op papier of op een beeldscherm worden gemaakt, bepaalt de kandidaat zelf de volgorde waarin de opgaven worden beantwoord. Het voorlopig overslaan van een opgave, door het toetsboekje bladeren of eerder gegeven antwoorden veranderen, zijn hier in principe allemaal mogelijk. Maar niet bij een CAT, waar de software bepaalt in welke volgorde de opgaven worden beantwoord. Voor bepaalde kandidaten, vooral onzekere personen, kan dat nadelig uitwerken op hun toetsprestatie. Diverse onderzoekers hebben in de afgelopen jaren geëxperimenteerd met aanpassingen van het algoritme die kandidaten meer vrijheid geven op dit gebied. Vergeleken met een normale CAT werden ongeveer dezelfde vaardigheidsschattingen en meetnauwkeurigheden gevonden, maar waren daar wel meer opgaven voor nodig (Lunz & Bergstrom, 1994; Rocklin, 1994; Eggen, 2004).

CATs in Nederland

Dat de genoemde voordelen van CAT niet alleen in theorie bestaan, moge blijken uit de beschrijving van en ervaringen met twee op grote schaal ingevoerde CATs in het Nederlandse hoger onderwijs.

WISCAT-pabo

Deze CAT is in 2006 op verplichte basis ingevoerd voor eerstejaars studenten van de pabo. Er was hiertoe besloten na ophef in de landelijke pers over een onderzoek waarin werd aangetoond, dat meer dan de helft van genoemde studenten niet in staat werd geacht om gedurende de opleiding een niveau van rekenvaardigheid te verwerven dat nodig is om leerlingen in het basisonderwijs te leren rekenen (Straetmans & Eggen, 2005). De toets had tot doel om aan het eind van het eerste studiejaar een bindend studie-advies af te kunnen geven over het al dan niet mogen vervolgen van de opleiding. Eerstejaars pabostudenten werden vóór 2006 ook al getoetst op hun rekenvaardigheid, maar daarbij was geen sprake van een gestandaardiseerde rekentoets die landelijk werd ingezet en ook niet van een aan het toetsresultaat gekoppeld bindend studie-advies. De meeste pabo's gebruikten een zelf ontwikkelde rekentoets en een zelf vastgestelde cesuur, meestal uitgedrukt als een te behalen percentage correct beantwoorde opgaven. Dit betekende dat er net zoveel gewenste rekenvaardigheidsniveaus waren als door individuele pabo's gebruikte toetsen. Het Cito verwierf de opdracht om een gestandaardiseerde, landelijk in te zetten rekentoets te ontwikkelen inclusief een cesuur. De samenwerkende pabo's kozen in onderling

overleg voor een cesuur die overeenkwam met de rekenvaardigheid van het tachtigste percentiel leerlingen aan het einde van groep acht. Dat hield dus in dat pabo-studenten vóór het eind van hun eerste leerjaar aan moesten tonen beter te kunnen rekenen dan 79 procent van alle groep 8 leerlingen.

De toetsontwikkelaars besloten een IRT-geschaalde opgavenbank voor rekenvaardigheid te ontwikkelen met een bijbehorende CAT, inclusief een rapportage- en administratiemodule. Deze keuze werd gemaakt om twee problemen op te lossen die zich zouden voordoen bij het inzetten van een klassieke toets (Straetmans & Eggen, 2007a):

- *Tijdstip van afname*. De opdrachtgever had aangegeven dat het niet mogelijk was om de toets op een voor alle pabo's gelijk tijdstip af te nemen. Dit leverde een probleem op ten aanzien van de geheimhouding van het toetsmateriaal. Het aan elkaar doorgeven van opgaven is een voor de hand liggende en nauwelijks te bestrijden handelwijze van kandidaten als die niet gelijktijdig worden getoetst. Doorgaans probeert men de nadelige gevolgen hiervan voor de betrouwbaarheid van de toetsscores tegen te gaan door met verschillende toetsversies te werken, maar dat zou in dit specifieke geval tot wel erg veel verschillende toetsversies hebben geleid, wat de ontwikkelings- en onderhoudskosten te hoog zou doen oplopen.
- *Heterogene doelgroep.* De populatie van eerstejaars pabo-studenten was zeer heterogeen qua rekenvaardigheid, met name als gevolg van de verschillen in genoten vooropleiding. De meeste studenten hadden een havodiploma (ongeveer 45 procent) of een mbo-diploma (ongeveer 35 procent). Ongeveer 12 procent had een vwo-diploma. Dan was er nog een restcategorie (ongeveer 8 procent) van, onder andere, zij-instromers. Onderzoek liet zien dat de rekenvaardigheden van deze subgroepen significant van elkaar verschilden (Straetmans & Eggen, 2005, p.130). Voor de toetsconstructie was dat een lastig gegeven. Om nauwkeurig te kunnen meten moet een toets qua moeilijkheidsgraad passen bij de vaardigheid van de kandidaten. Maar als die vaardigheden zeer uiteenlopen heeft de toetsconstructeur geen goed ijkpunt voor het bepalen van de moeilijkheidsgraad van de te construeren en in een toets op te nemen opgaven.

Meer gedetailleerde informatie over het ontwerp en de werking van WISCAT-pabo kan gevonden worden bij Straetmans & Eggen (2007b).

De implementatie van WISCAT-pabo verliep succesvol; na vier jaar waren er al meer dan 60.000 toetsen, zonder noemenswaardige problemen, afgenomen. Studenten bleken, na een korte introductie over de bediening van de beeldschermtoets en de eigenschappen van adaptieve toetsing, goed overweg te kunnen met de applicatie. Slechts af en toe kwamen er meldingen, bijvoorbeeld van stress, die veroorzaakt zou worden door het adaptieve karakter van de toets. Zo maakte de onmogelijkheid om een gepresenteerde opgave even over te slaan, om die op een later tijdstip te beantwoorden, sommige studenten zenuwachtig. Nerveus werden sommigen ook als na één of enkele foute antwoorden de moeilijkheid van een volgende opgave opvallend veel lager was dan van vorige opgaven. Voor die studenten was dit kennelijk hét signaal dat men op weg was naar een onvoldoende resultaat. Vaker echter hadden studenten (en ook hun docenten) moeite met de scoring. Dat hetzelfde aantal correct beantwoorde opgaven niet altijd leidt tot dezelfde vaardigheidsscore en soms ook tot een verschillende beslissing over zaken/slagen, is moeilijk te begrijpen en te accepteren. Bij de invoering van de WISCAT-pabo bleek het daarom noodzakelijk om alle betrokkenen uitvoerig voor te lichten over de werking van deze adaptieve toets en ook over alle verschil-

len met de voor hen meer bekende klassieke toetsen (Straetmans & Eggen, 2009).

Voortgangstoets Geneeskunde

In de jaren 70 van de vorige eeuw werd in de pas opgerichte medische faculteit van de Universiteit van Maastricht de Voortgangstoets geïntroduceerd. Deze innovatie was bedoeld om een halt toe te roepen aan het tentamengerichte studiegedrag van studenten, dat in samenhang met het absolverende karakter van de tentamens een bedreiging vormde voor de diplomakwaliteit. Voortgangstoetsing zou studenten ertoe aanzetten om al verworven kennis blijvend te onderhouden. Om dat te bereiken maakten studenten uit elk studiejaar van de bachelor- en masteropleiding Geneeskunde op vier tijdstippen een toets, die vaststelde in hoeverre het vereiste kennisniveau van de geneeskundeopleiding bereikt was. De scores werden vooral gebruikt voor formatieve beslissingen (Is er progressie gemaakt ten opzichte van een eerder toetstijdstip? Welke studieonderdelen verdienen extra aandacht?) over studenten, maar telden ook mee voor summatieve beslissingen (Voldoet het kennisniveau aan de gestelde eisen voor de bachelor of masteropleiding?). Verder leverden ze informatie op om het onderwijsprogramma of onderdelen daarvan mee te evalueren. In de jaren daarna kregen steeds meer geneeskundeopleidingen in het land belangstelling voor de Voortgangstoets. Inmiddels hebben acht opleidingen zich verenigd in het samenwerkingsverband interuniversitaire Voortgangstoets Geneeskunde (iVTG).

Tot voor kort was de Voortgangstoets een gestandaardiseerde paper-based test (PBT). Alle studenten van alle deelnemende opleidingen kregen op elk van de vier toetsmomenten dezelfde toets te maken, bestaande uit 200 meerkeuzevragen over het toepassen, analyseren en evalueren van medische kennis. Bij het construeren van de vragen stemden de ontwikkelaars de moeilijkheidsgraad ervan in principe af op het vereiste kennisniveau van masterstudenten in het laatste studiejaar. Daarom was er voor studenten de mogelijkheid om op vragen die ze door nog ontbrekende kennis niet konden beantwoorden te responderen met een '?'. Niet beantwoorde vragen leverden geen punten op en voor fout beantwoorde vragen werden punten afgetrokken. Omdat alle studenten dezelfde toets maakten, was het noodzakelijk om de toetsafname voor alle deelnemers steeds op exact hetzelfde moment te laten plaatsvinden. Deze opzet had tot gevolg dat men in de uitvoering te kampen had met problemen van psychometrische en logistieke aard. Zo kon men voor de score-interpretatie geen gebruik maken van een absolute cesuur, omdat de moeilijkheidsgraad van de toetsen varieerde. In plaats daarvan werd een relatieve cesuur gebruikt, die afhing van de gemiddelde score van de groep studenten. Daarmee kon echter geen zuiver beeld worden verkregen van de mate waarin een individuele student de einddoelen bereikt had en ook niet van de voortgang van de kennisverwerving. Problematisch was ook de lengte van de toets; 200 vragen beantwoorden in vier uur tijd is zeer vermoeiend en kan gemakkelijk leiden tot concentratieverlies en daardoor bij sommige studenten tot lagere toetsscores dan verwacht. Nog een ander probleem betrof de moeilijkheidsgraad van de toets. Die was eigenlijk alleen voor vergevorderde masterstudenten goed afgestemd op hun kennisniveau, wat de zuiverheid van de toetsscores en de relevantie van de feedback op de kennisprogressie benadeelde. In logistiek opzicht betekende de eis van gelijktijdige toetsing van minstens 1500 studenten per opleiding een uitdaging voor de facilitaire dienstverlening zoals lokaalbeheer, drukkerij, surveillance en toetsverwerking (Donkers et al., 2024).

Genoemde problemen waren voor iVTG aanleiding om te besluiten een toetsvorm te kiezen waarbij niet langer alle studenten op hetzelfde moment getoetst hoefden te worden. Om dit te realiseren zouden er op een

toetstijdstip verschillende toetsversies moeten worden gebruikt, terwijl de toetsscores daarvan onderling vergelijkbaar zouden blijven. Uiteindelijk leidde dit tot de beslissing om de toetsvorm van de Voortgangstoets te wijzigen van PBT naar CAT, met belangrijke voordelen als gevolg (Donkers et al., 2024):

- *Kortere toets*. De IRT-gekalibreerde opgavenbank telt ongeveer 6.000 meerkeuzevragen, waaruit het algoritme steeds toetsen samenstelt van 135 vragen waarvoor studenten 3 uur de tijd krijgen. Door de grotere efficiëntie van adaptieve toetsing was de verwachting dat de aanzienlijke kortere toets geen grote nadelige gevolgen zou hebben voor de betrouwbaarheid. Een pilotstudy bevestigde dit door aan te tonen dat de test-hertest betrouwbaarheid, waarbij studenten zowel de PBT als de CAT maakten, 0,83 bedroeg.
- *Variabel afnametijdstip.* De opleiding mag voortaan zelf bepalen, binnen een voorgeschreven periode van anderhalve week, op welke dagen en tijdstippen er getoetst wordt. Het toetsen van studenten op verschillende tijdstippen wordt mogelijk gemaakt doordat het bij CAT nagenoeg zinloos is om toetsinhouden aan elkaar door te geven, aangezien elke student in principe een andere toets maakt.
- *Vergelijkbare toetsscores.* Dat toetsversies, zowel binnen een toetstijdstip als tussen toetstijdstippen, van elkaar verschillen in inhoud en moeilijkheidsgraad is geen belemmering voor de vergelijkbaarheid van de behaalde toetsscores. Op grond van het aantal correct beantwoorde opgaven wordt een schatting gemaakt op een onderliggende IRT-vaardigheidsschaal die relevant is voor elke toets die uit de opgavenbank wordt samengesteld. Op deze schaal kunnen ook absolute cesuurscores worden bepaald, wat bijdraagt aan meer nauwkeurige beslissingen over het beheerste kennisniveau.
- *Nauwkeuriger feedback.* Doordat de toetsen qua moeilijkheidsgraad voortaan beter aansluiten bij het kennisniveau van de individuele student, is de feedback (in de vorm van leerstofverwijzingen) relevanter en begrijpelijker voor de student.
- *Eenvoudiger logistiek.* De administratieve werklast die gepaard gaat met het samenstellen, vermenigvuldigen en verspreiden van (12.000) toetsboekjes, het organiseren van voldoende grote toetsruimtes, de verwerking van de antwoordbladen, de administratie en rapportage van de toetsscores en het bijwerken van de metadata van de opgavenbank, is nu grotendeels geautomatiseerd.

Donkers et al. (2024) schrijven dat er tot dan zes reguliere adaptieve toetsafnames zijn geweest en dat die zonder noemenswaardige problemen zijn verlopen. Ervaringen van studenten werden niet gerapporteerd.

Tot slot

De snelle ontwikkeling en implementatie van computergestuurde adaptieve toetsing is sterk aangejaagd door de vooruitgang op twee cruciale gebieden, namelijk de ontwikkeling van de itemresponstheorie en de opkomst van computertechnologie in het onderwijs. IRT is vooral in theoretische zin belangrijk geweest. Niet voor niets wordt er gezegd dat adaptieve toetsing 'le raison d'être' van IRT is (Wainer et al., 2000, p. 9). De grote winst is dat toetsen die geconstrueerd zijn op basis van IRT-principes en -procedures zowel nauwkeuriger als efficiënter zijn dan toetsen op basis van de KTT. De grootschalige introductie van computersystemen in het onderwijs heeft de toepassing van CAT in praktische zin mogelijk gemaakt. Daardoor werd het mogelijk om de complexe berekeningen van IRT-modellen in real-time uit te voeren tijdens toetsafnames.
In de 21ste eeuw hebben de ontwikkelingen niet stilgestaan. Naast de CAT bestaat nu ook de Multi-Stage Toets (MST). Een MST past zich net als een CAT aan het vaardigheidsniveau van een kandidaat aan. Het verschil is echter

dat een MST geen individuele opgaven, maar sets van gekalibreerde opgaven aanbiedt. Een MST start met een voor alle kandidaten identieke 'routing module' waarna een reeks van volgende stadia volgt, waarin kandidaten op basis van hun prestatie op de eerdere module een volgende krijgen. Een voordeel van de MST ten opzichte van de CAT is dat een MST eenvoudiger is om samen te stellen. Een nadeel is dat een MST minder efficiënt is. Zie voor meer informatie bijvoorbeeld Magis et al. (2017). Ook is er de laatste jaren onderzoek gedaan naar de verdere verfijning van adaptieve toetsing door middel van kunstmatige intelligentie (AI). AI kan verbetering brengen door algoritmes dynamischer en responsiever te maken op basis van real-time data van studenten (De Ayala, 2009). Een ander belangrijk gebied van innovatie is de integratie van multimodale assessments. Hierbij wordt niet alleen naar toetsresultaten gekeken, maar worden ook gegevens zoals oogbewegingen, reactietijd en muisgedrag verzameld om een completer beeld van de student te krijgen (Zapata-Rivera & VanWinkle, 2011).

Deze ontwikkelingen beloven een mooie toekomst voor adaptieve toetsing. De groeiende beschikbaarheid van AI en big data in het onderwijs maakt het mogelijk om CAT-algoritmes verder te verfijnen, waardoor toetsresultaten een beter inzicht kunnen geven in de beheersing van kennis en vaardigheden door onderwijsdeelnemers en meer specifieke feedback kunnen verstrekken om het leerproces te bevorderen. Maar veel belangrijker is dat met de toenemende beschikbaarheid en betere toepasbaarheid van algoritmes adaptieve toetsing ook steeds eenvoudiger te implementeren zal zijn. Het proof of concept dat de WISCAT-pabo en de Voortgangstoets Geneeskunde geleverd hebben, maakt duidelijk dat het voor samenwerkende opleidingen mogelijk is om voor hun gemeenschappelijke kennisbasis adaptieve toetsen te ontwikkelen. Samenwerking tussen overeenkomstige opleidingen zorgt voor voldoende tijd en menskracht. Bovendien maakt dit de hoge ontwikkelings- en onderhoudskosten beter draagbaar, die verbonden zijn met het aanleggen van omvangrijke opgavenbanken en het uitvoeren van grootschalig kalibratie-onderzoek. Uiteraard vraagt dit wel om standaardisering van de kennis- en vaardigheidsbasis van de samenwerkende opleidingen.
De beschrijving van de twee landelijk geïmplementeerde CATs en van enkele veelbelovende, recente ontwikkelingen hebben hopelijk duidelijk gemaakt dat adaptieve toetsing al lang geen gimmick meer is en eerder te beschouwen is als een kroonjuweel van de toetsgemeenschap, dat niet alleen leidt tot efficiëntere toetsing maar ook tot efficiëntere onderwijsprocessen, zoals bijvoorbeeld mogelijk gemaakt door flexibilisering van opleidingstrajecten. ■

Referenties

- Adèr, H. J., Mellenbergh, G. J., & Hand, D. J. (2008). *Advising on research methods: A consultant's companion.* Huizen: Johannes van Kessel Publishing.
- Baker, F.B. (1985). *The basics of item response theory.* Portsmouth, NH: Heinemann.
- Birnbaum, A. (1968). *Some latent trait models and their use in inferring an examinee's ability.* In F.M. Lord & M.R. Novick, Statistical theories of mental test scores (pp. 397-479). Reading, MA: Addison-Wesley.
- De Ayala, R. J. (2009). *The theory and practice of item response theory.* New York: The Guilford Press.
- Donkers, J. Van den Brink, J., Keesom, T. & Bremers, B. (2024). Logistieke organisatie van de computer-adaptieve interuniversitaire voortgangstoets geneeskunde. *Examens 21* (1), 19-25.
- Eggen, T. J. H. M (2004). *Contributions to the Theory and Practice of Computerized Adaptive Testing.* University of Twente.
- Fisher, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics.

Philosophical Transactions of the Royal Society of London. Series A, 222(594-604), 309-368.
- Glas, C.A.W., & Verhelst, N.D. (1993). Een overzicht van itemresponsmodellen. In T.J.H.M. Eggen & P.J. Sanders (red.), Psychometrie in de praktijk. Cito: Arnhem.
- Kingsbury, C. G., & Zara, A.R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*(4), 359-375.
- Lunz, M. E., & Bergstrom, B. A. (1994). An empirical study of computerized-adaptive test administration conditions. *Journal of Educational Measurement, 31*(3), 251-263.
- Magis, D., Yan, D. & Von Davier, A. (2017). Computerized Adaptive and Multistage Testing with R Using Packages catR and mstR. Cham: Springer.
- Mills, C. N., & Stocking, M.L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education, 9*(4),287-304.
- Odendahl, N. V. (2011). *Testwise. Understanding Educational Assessment, Volume 1.* Plymouth: Rowman & Littlefield Education.
- Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests.* Copenhagen: Nielsen & Lydiche.
- Rocklin, T. R. (1994). Self-adapted testing. *Applied Measurement in Education, 7*(1), 3-14.
- Sluijter, C., & Straetmans, G. J. J. M. (2024). Wat betekent die toetsscore? Over voetangels en klemmen bij het geven van betekenis aan toetsscores. *Examens, 21*(4), 6-16.
- Straetmans, G. J. J. M., & Eggen, T. J. H. M. (1998). Computerized Adaptive Testing: What it is and how it works. *Educational Technology, 38*(1), 45–52.
- Straetmans, G. J. J. M., & Eggen, T. J. H. M. (2005). Afrekenen op rekenen: Over de rekenvaardigheid van pabo-studenten en de toetsing daarvan. *Tijdschrift voor Hoger Onderwijs, 23*(3), 123-139.
- Straetmans, G. J. J. M., & Eggen, T. J. H. M. (2007a). WISCAT-pabo: Computergestuurd adaptief toetspakket rekenen. Praktisch artikel. *Onderwijsinnovatie, 3*, 17-27.
- Straetmans, G. J. J. M., & Eggen, T. J. H. M. (2007b). WISCAT-pabo: een gestandaardiseerde toets in een maatkostuum. *Examens, 4*(1), 5-10.
- Straetmans, G. J. J. M., & Eggen, T. J. H. M. (2009). Twee jaar WISCAT-pabo: beschrijving, resultaten en ervaringen. In M. van Zanten (red.), *Leren van evalueren: de lerende in beeld bij reken-wiskundeonderwijs* (pp. 139-160). Verslag van de 27e Panama-conferentie. Utrecht: Freudenthal Instituut.
- Suen, H. K. (1990). P*rinciples of test theories.* Hillsdale (NJ): Lawrence Erlbaum Associates.
- Van der Linden, W. J. (1983). *Van standaardtest naar itembank.* Oratie. Enschede: Universiteit Twente.
- Van der Linden, W. J. & Pashley, P. J. (2010). Item Selection and Ability Estimation in Adaptive Testing. In W.J. van der Linden & C.A.W. Glas (Eds.), *Elements of Adaptive Testing* (pp. 3-30). New York: Springer.
- Verhelst, N. D. (1993). Itemresponstheorie. In T.J.H.M. Eggen & P.J. Sanders (red.), *Psychometrie in de praktijk.* Cito: Arnhem.
- Vispoel, W. P., Rocklin, T. R., & Wang, T. (1994). Individual differences and test administration procedures: a comparison of fixed-item, computerized-adaptive, and self-adapted testing. *Applied Measurement in Education, 7*(1), 53-79.
- Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B.F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000). *Computerized Adaptive Testing: a primer* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Zapata-Rivera, D., & VanWinkle, W. (2011). Multimodal assessments. *ETS Research Report Series*, 2011(1), 1-8.