

Standaardtoets: weg ermee?

□ Wouter Schoonman

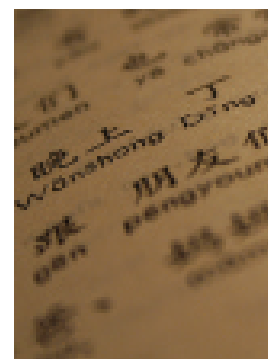
In het oude China werden kandidaten voor overheidsfuncties aan talrijke examens en toetsen onderworpen. Het fenomeen toets of examen is dus al duizenden jaren oud, inclusief fraude die daarbij hoort. Bewaard gebleven is een kimono die aan de binnenzijde volgeschreven was met 'spiekbriefjes'. Hoe het met de eigenaar is afgelopen is niet bekend. De huidige examenpraktijk in Nederland is – althans in het hoger onderwijs – sterk beïnvloed door de Amerikaanse situatie. Het merendeel van onze tentamens en examens bestaat uit meerkeuzevragen, met een vaste sleutel en als het goed is een vaste cesuur. Dit is een verarming wanneer we bijvoorbeeld kijken naar de adaptieve test.

Adaptief toetsen

Alphonse Binet ontwikkelde aan het begin van de vorige eeuw een intelligentietoets op maat: eigenlijk de allereerste adaptieve toets. Niet elke kandidaat hoeft dezelfde opgaven voorgelegd te krijgen, de toets kan zich aanpassen op basis van voorgaand antwoordgedrag. We zien momenteel een hernieuwde belangstelling voor de adaptieve toets, mede onder invloed van *item response* theorie (IRT) en toegenomen computerkracht (zie bijvoorbeeld de adaptieve rekentoets van Cito waar zoveel publiciteit over is geweest). Andere innovaties zijn ook mogelijk, mede

door computer technologie inclusief internet. Een aantal mogelijkheden wordt in dit artikel besproken. Het idee achter adaptief toetsen is nogal eenvoudig: waarom zou je kandidaten lastig vallen met opgaven die of veel te moeilijk of veel te gemakkelijk zijn? In psychometrisch jargon spreek je dan van het optimaliseren van de informatie die je verkrijgt wanneer je een bepaald item bij deze persoon afneemt. Opgaven worden zo gekozen dat er maximale informatie over het beheersingsniveau van de kandidaat wordt verkregen. Hierdoor neemt het aantal af te nemen opgaven af, omdat niet-informatieve opgaven niet worden aangeboden (Schoonman, 1989). Het mogelijke probleem dat kandidaten (dus) met verschillende opgaven te maken krijgen, maar nog steeds onderling vergelijkbaar moeten blijven, is elegant opgelost door de ontwikkelingen op het gebied van *item response* theorie. Zonder te diep op deze 'nieuwe' psychometrie in te willen gaan, kunnen we nu zeker zijn dat kandidaten die verschillende opgaven gemaakt hebben, toch vergelijkbaar blijven. Dit is niet mogelijk wanneer we gebruik maken van klassieke testtheorie. Het tweede voordeel van adaptief toetsen is dat je de betrouwbaarheid van de verkregen scores kunt afstemmen op het belang van de te nemen beslissing (meestal zak/slaag-beslissingen). De betrouwbaarheid van de verkregen score kan hoger worden gemaakt wanneer het om kritische grenzen gaat. Dit verhoogt het rendement (de 'utiliteit') van de toets omdat er minder verkeerde beslissingen worden genomen (de zogenoemde vals positieven en vals negatieven). Dat zijn twee belangrijke voordelen van de adaptieve toets. Uiteraard kleven er ook nadelen aan de adaptieve toets. Het eerste probleem is dat de kwaliteit van de opgaven erg hoog moet zijn. Alle opgaven moeten (vooraf) getest zijn om te zien of zij voldoen aan de strenge eisen van het gebruikte IRT-model, wat geen eenvoudige opgave is. Er zullen veel opgaven moeten worden geschreven, worden getest, worden herzien en opnieuw worden getest, voordat zij aan de itembank toegevoegd kunnen worden. Dit verhoogt de kosten van (de ontwikkeling van) een

Waarom kandidaten lastig vallen met opgaven die of veel te moeilijk of veel te gemakkelijk zijn?



adaptieve toets. Maar stel, we hebben een zogenoemde gekalibreerde itembank met honderden opgaven, zodat adaptieve toetsen volgens IRT inderdaad mogelijk wordt. Wat zal het effect zijn op de kandidaat die adaptief getoetst wordt? Het eerste effect is dat kandidaten in de categorie 'grensgevallen' op hun tenen moeten lopen: immers alle opgaven zijn gericht op de grens tussen wel- en niet-beheersing. Deze kandidaten worden getoetst rond dit cruciale beheersingsniveau. Dit maakt de toets voor deze kandidaten zwaarder. Voor kandidaten boven en onder de kritische grens is zo'n adaptieve toets overigens minder belastend. Het is sneller duidelijk of je boven of onder de cesuur zit. Een tweede effect van een adaptieve toets is dat de afnametijd minder voorspelbaar is. Bij duidelijke gevallen boven of onder de cesuur is de afnametijd korter, bij twijfelgevallen zullen er meer opgaven worden aangeboden, waardoor de afnametijd toeneemt. Qua planning is dit voor de organisatie die de (adaptieve) toets afneemt lastiger. De winst in nauwkeurigheid is daarentegen een belangrijk voordeel bij adaptief toetsen.

Deelkrediet

Iedere opleider weet dat sommige antwoorden op een toets 'fouter' zijn dan andere antwoorden. Bij een meerkeuze toets met vele opgaven zijn sommige alternatieven echt fout, terwijl voor andere alternatieven 'nog iets te zeggen' valt. Een voorbeeld staat hieronder:

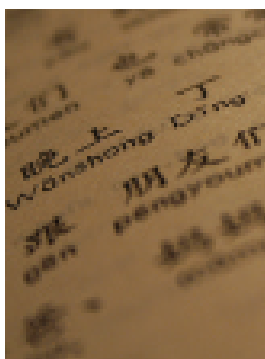
Vraag:

Welke stad is de hoofdstad van Noord Holland?

- A. Amsterdam
- B. Haarlem
- C. Eindhoven

Je zou kunnen zeggen dat alternatief A minder fout is dan alternatief C. Kandidaten die alternatief A kiezen zouden daardoor toch nog punten kunnen krijgen in vergelijking met kandidaten die alternatief C kiezen. Hiermee wordt 'gedeeltelijke kennis' toch beloond, wat een betere reflectie van het vaardigheidsniveau is in vergelijking met een dichotome goed/fout-scoring. Hoeveel 'deelkrediet' aan foute antwoorden moet





De winst in nauwkeurigheid is een belangrijk voordeel bij adaptief toetsen

worden gegeven, kan door experts worden bepaald, maar dit kan ook empirisch worden bepaald. Dat laatste heeft de voorkeur, omdat daarmee minder gelukkige itemconstructie wordt verdisconteerd. Een voorbeeld van de combinatie van deeltkrediet en adaptief toetsen – waarbij zowel partiële kennis als informatie optimalisatie plaatsvindt – is te vinden bij Dodd et al. (1993).

Itemgeneratie

De meeste toetsen maken gebruik van een standaard verzameling van opgaven, in een vaste volgorde. Dit soort toetsen wordt 'serieel' genoemd, omdat elke kandidaat dezelfde opgaven in dezelfde volgorde krijgt aangeboden. Een risico bij dergelijke toetsen is het 'lekken' van de juiste antwoorden. Zowel in het onderwijs als in het bedrijfsleven is dit een bekend gegeven. Een simpele, maar niet afdoende oplossing bij computertoetsen is om de volgorde van de alternatieven bij een meerkeuzevraag *random* te presenteren. Een betere maatregel tegen fraude is itemgeneratie. Bij deze vorm van toetsen bestaan er geen vastliggende opgaven, maar uitsluitend algoritmen die de inhoud en vorm van de opgave bepalen. Elke opgave wordt, door de computer, tijdens de afname gegenereerd volgens vastgelegde regels. Hierdoor wordt elke kandidaat geconfronteerd met een unieke opgave. Ook nu geldt natuurlijk dat elke gegenereerde set items over kandidaten vergelijkbaar moet zijn. Dit is te checken via test-hertestonderzoek.

Kandidaten moeten in zo'n type onderzoek ongeveer dezelfde scores op verschillende, gegenereerde examens behalen. Een van de voordelen is dat op afstand toetsen gemakkelijker is te realiseren. De kandidaat kan de toets doen waar en wanneer hij wil. De toetsende organisatie kan het toetsresultaat valideren door de toets nogmaals af te nemen onder gecontroleerde omstandigheden. De toets zal inhoudelijk hetzelfde zijn, maar de opgaven zullen verschillen.

Conclusie

De standaardtoets met vaste opgaven, vaste alternatieven en een vaste afnamevolgorde en cesuur lijkt achterhaald. Nieuwe technologie en methodologie maakt het mogelijk om al deze kenmerken van de standaardtoets te verbeteren:

- Alternatieven in een meerkeuze toets kunnen in willekeurige volgorde worden aangeboden en gekalibreerde itembanken kunnen gebruikt worden om adaptief te toetsen.
- Adaptief toetsen kan op haar beurt ingezet worden om nauwkeuriger te meten rond de cruciale slaag/faal-grens. Het is een toets op maat, met een hogere nauwkeurigheid.
- Deeltkrediet voor een keuze door kandidaten voor 'minder foute alternatieven' bij meerkeuze vragen biedt een extra mogelijkheid om beter gebruik te maken van de beschikbare informatie die verkregen wordt bij het toetsen. Er wordt beter gebruik gemaakt van beschikbare informatie.
- Itemgeneratie is een fascinerend gebied met tal van voordelen. Er bestaan geen 'vaste items' meer, maar algoritmen die items hier-en-nu, tijdens de afname genereren. Ook hier wordt een toets op maat, dat wil zeggen voor iedere kandidaat verschillend, gemaakt.

De ultieme gevolgen van deze verbeteringen zijn dat toetsen op afstand beter mogelijk wordt, dat de integriteit van toetsen sterker beschermd wordt en dat kandidaten op maat – via internet – worden getoetst. Deze nieuwe mogelijkheden vormen overigens niet het einde van de mogelijke verbeteringen. Toetsen op basis van interactieve video en simulaties met geautomatiseerde scoring, zijn binnenkort haalbaar. Dit resulteert in meer authentieke, efficiënte, gepersonaliseerde, integere en meer valide vormen van toetsen. We moeten natuurlijk wel alert blijven op spiekbriefjes en andere vormen van fraude, want dat gebeurde bij de oude Chinezen ook al.

Literatuur

- Dodd, B.G. et al. (1993). Computerized Adaptive Testing Using the Partial Credit Model: Effects of Item Pool Characteristics and Different Stopping Rules. *Ed. & Psych. Measurement*, 53(1), 61-77.
- Schoonman, W. (1989). *An applied study on computerized adaptive testing*. Lisse: Swets & Zeitlinger.

□ Dr. W. Schoonman is bedrijfspsycholoog en lector aan de Saxion Hogescholen. Hij is tevens voorzitter van de Nederlandse Vereniging voor Examens.