DE STANDAARDMEETFOUT IS BELANGRIJKER DAN DE BETROUWBAARHEID

Studietoetsen: zinvolle inhoud of betrouwbare score?

Klaas Sijtsma

Voor studietoetsen lijken inhoudsvaliditeit en betrouwbaarheid op gespannen voet te staan. Staat de inhoudsvaliditeit centraal, dan is de inhoud van de toets doorgaans heterogeen en kan de toetsscore onbetrouwbaar zijn. Mikt men op een hoge betrouwbaarheid, dan is de toets vaak inhoudelijk eenzijdig. Een beetje van beide levert vlees noch vis, dus wat te doen met dit schijnbare dilemma?

Inleiding

In dit artikel wordt gesteld dat de inhoud van een toets representatief voor de studiestof dient te zijn en beslissingen over studenten moet toelaten die accuraat zijn. Deze twee doelen kunnen gelijktijdig worden bereikt door een groot aantal vragen en opdrachten te gebruiken. Lange toetsen zijn vaak betrouwbaar, ook als ze inhoudelijk heterogeen zijn. Nog belangrijker is dat hun standaardmeetfout ten opzichte van de standaarddeviatie van de toetsscores gering is. Korte toetsen hebben een lagere betrouwbaarheid en een relatief grote standaardmeetfout. Voor het nemen van accurate beslissingen over studenten is een relatief kleine standaardmeetfout vereist, en doorgaans wordt die dus bereikt met een lange toets.

Achtereenvolgens wordt ingegaan op de volgende vragen: Wat is een representatieve toets? Waarom is de standaardmeetfout belangrijker dan de betrouwbaarheid? Wat is de invloed van de toetslengte op de standaardmeetfout?

Summatieve studietoetsen

Uitgangspunt is de *teacher-made*studietoets zoals deze aan universiteiten en hogescholen vaak wordt gebruikt. De examencommissie is wettelijk verantwoordelijk voor de kwaliteit van de toets. De toets wordt meestal *stand-alone*afgenomen. Technisch moeilijk te realiseren alternatieven, zoals adaptief toetsen, zijn hier niet aan de orde. Met behulp van een studietoets stelt men, afhankelijk van de context, vast of studenten voldoende kennis over een vak hebben opgedaan, of zij inzicht in de stof hebben gekregen, en of zij zich de geoefende vaardigheden eigen hebben gemaakt. Het gebruik van de toets is in deze gevallen summatief. Het gebruik van toetsen om het leerproces zelf te beoordelen, het formatieve gebruik, brengt een geheel eigen problematiek met zich mee. Het gaat dan vooral om de mogelijkheid met de toets vooruitgang of het ontbreken ervan te meten. Dit veronderstelt herhaalde meting. Ook kunnen toetsen worden gebruikt ten behoeve van de diagnostiek van kennislacunes en leerproblemen. De meeste studietoetsen zijn summatief, formatieve toetsen worden hier verder buiten beschouwing gelaten.

Wat is een representatieve toets: samenstelling en inhoudsvaliditeit

Validiteit betekent voor een studietoets dat de items (vragen, opdrachten, enzovoort) de studiestof en de eindtermen voor een vak goed representeren. Bij de studiestof is de vraag of de items een representatieve steekproef vormen uit het domein van mogelijke items. Hierover waakt de docent, maar wat vindt hij/zij representatief? Er zijn vele mogelijkheden. Bijvoorbeeld, per hoofdstuk uit het boek worden twee vragen gesteld; over de belangrijke hoofdstukken worden drie vragen per hoofdstuk gesteld en over de minder belangrijke één vraag; het aantal vragen per hoofdstuk is afhankelijk van het aantal pagina's; of over moeilijke onderwerpen worden meer vragen gesteld. Bij de eindtermen kan men bijvoorbeeld denken aan nadruk op technische in plaats van conceptuele kennis, maar ook kan 2/3 van de vragen reproductie van kennis betreffen en 1/3 inzicht

Een inhoudelijk homogene toets met een hoge betrouwbaarheid staat op gespannen voet met onderwijsdoelen

vereisen. Hofstee (1973) stelde ooit het onbenullige item voor ten behoeve van participatiecontrole, dus of iemand het boek heeft gelezen. De docent zou bijvoorbeeld kunnen vragen wat de titel van het boek is. Probleem is hierbij vooral het gebrek aan transparantie: studenten zien de relatie met de studiestof en de eindtermen niet, en voelen zich niet serieus genomen. Het principe is duidelijk: de samenstelling van de toets hangt mede af van de voorkeuren van de docent en is binnen deze beperking arbitrair. Vaak licht de docent de studenten via de studiegids vooraf in over het type vragen (meerkeuze, essay) dat wordt gebruikt en soms ook hoe een toets is samengesteld (welke hoofdstukken vooral? kennis of inzicht?). Verder dient de toets transparant te zijn in de zin dat de studenten de relatie tussen de vragen, de studiestof en de eindtermen zien.

De informatie over de vorm van de toets is bedoeld als geruststelling voor studenten, maar voor het resultaat op de toets doet die informatie er niet zoveel toe. Immers, ongeacht de samenstelling, slagen naar verwachting de betere studenten voor de meeste toetsen en zakken de zwakke studenten. Studenten uit de middengroep zakken voor de ene toets en slagen voor de andere toets, maar het probleem is daar veel meer dat zij zo dicht tegen de cesuur aanzitten, dat toeval een dominante rol speelt. Dan gaat het dus over meetfouten, betrouwbaarheid en standaardmeetfout.

Voor de discussie 'inhoudsvaliditeit of betrouwbaarheid' is van groot belang dat inhoudelijk representatieve toetsen vrijwel altijd heterogeen van samenstelling zijn (de Groot & van Naerssen, 1969). Dit komt doordat het getoetste vak inhoudelijk heterogeen is. 'Menselijke anatomie', 'Inleiding statistiek' en 'Geschiedenis van de psychologie' zijn alle veelzijdige onderwerpen. De naam van het vak suggereert alleen maar homogeniteit. Homogeniteit in de zin van schaalbaarheid langs een enkele meetlat is een buitengewoon restrictieve, wiskundige eigenschap, waarmee men bij de samenstelling van een vak en een representatieve toets geen rekening houdt. Streven naar homogeniteit zou een vak en een toets erover inhoudelijk onaanvaardbaar uitkleden en dus verarmen. In de praktijk zijn representatieve toetsen zelden ééndimensioneel, zelfs niet bij benadering.

Betrouwbaarheid van toetsscores en de standaarmeetfout

De betrouwbaarheid van een toetsscore geeft een antwoord op de vraag: Wat gebeurt er als ik de toets opnieuw doe? Dit spoort met het algemene gebruik van statistiek. Daar wordt de vraag gesteld: Als ik een nieuwe steekproef zou kunnen trekken, in hoeverre zou ik dan tot dezelfde conclusie komen? Verwerp ik opnieuw de nulhypothese? Vind ik dezelfde samenhang? Bij een studietoets wordt de vraag gesteld of iemand op een ander tijdstip of op een gelijkwaardige toets dezelfde prestatie levert.

Betrouwbaarheid is in de klassieke testtheorie gedefinieerd als de correlatie tussen twee toetsen waarvan de toetsscores, op toevallige meetfouten na, geheel inwisselbaar zijn. Technisch gezien gaat het dan om parallelle toetsen. Jan mag dus wel twee verschillende toetsscores hebben, maar het verschil moet geheel aan toeval zijn te wijten, niet aan systematische verschillen in moeilijkheid of inhoud van items.

De betrouwbaarheid (*r*) ligt theoretisch tussen 0 en 1, en toetsconstructeurs vragen begrijpelijk welke waarde voldoende is. Vaak vindt men 0.8 voldoende en minstens 0.9 heel goed. Toch zijn dit maar arbitraire vuistregels die hooguit schijnzekerheid verschaffen. Wat men met een voldoende betrouwbaarheid zou kunnen bedoelen, is een waarde waarbij het aantal studenten dat ten onrechte zakt zo klein mogelijk is. Alleen als $r=1$ en de inhoudsvaliditeit in orde is, is dit aantal gelijk aan 0, maar lagere waarden van *r* geven over dit aantal geen nauwkeurige informatie.

Meer duidelijkheid biedt de standaardmeetfout van de toets. Dit is de standaarddeviatie van de meetfouten, E, die wordt aangegeven met S_E. Daarmee kan de volgende vraag worden beantwoord: Als de cesuur van de toets 24 punten is en Jan heeft 20 punten (zijn toetsscore X is: $X = 20$), is zijn prestatie dan significant lager dan 24 punten, zodat Jan terecht is gezakt en niet door toeval? Men kan statistisch toetsen of Jans *true score*- zijn gemiddelde toetsscore berekend over een groot aantal, denkbeeldige parallelle toetsen - gelijk is aan de cesuur tegen het alternatief dat zijn *true score*eronder ligt. Met Jans toetsscore en de standaardmeetfout is bijvoorbeeld een 90% betrouwbaarheidsinterval voor Jans *true score*te bepalen. Ligt de cesuur boven het interval,

Een inhoudelijk heterogene toets van voldoende lengte kan een hoge betrouwbaarheid hebben

dan is Jan uiteraard gezakt. Ligt de cesuur in het interval, dan is de conclusie dat Jans *true score* gelijk is aan de cesuur en is hij niet gezakt. In de toetspraktijk trekt men die conclusie overigens niet snel en laat men iedereen zakken van wie de toetsscore onder de cesuur ligt.
Het rekenwerk bij het gebruik van betrouwbaarheidsintervallen is als volgt. Neem toetsscore $X = 20$ als schatting van Jans "*true score*" T (notatie $\hat{T}$ geeft aan dat de *true score* is geschat) en veronderstel willekeurig dat $S_E = 1.5$. Bij normaal verdeelde meetfouten is het 90% betrouwbaarheidsinterval

$$[\hat{T}_{\text{Jan}} - 1.645\ S_E; \hat{T}_{\text{Jan}} + 1.645\ S_E];$$

na invullen van de gegevens levert dit [17.53; 22.47]. De cesuur van 24 punten ligt niet in het interval en het is dus voldoende zeker dat Jan gezakt is.
Het zwakke punt van de redenering is dat gedaan wordt alsof de standaardmeetfout voor iedereen gelijk is. De item-responstheorie (van der Linden & Hambleton, 1997) leert dat niet iedereen even nauwkeurig wordt gemeten, en biedt de mogelijkheid de nauwkeurigheid op persoonsniveau te bepalen. Deze benadering is echter onbruikbaar bij kleine steekproeven.

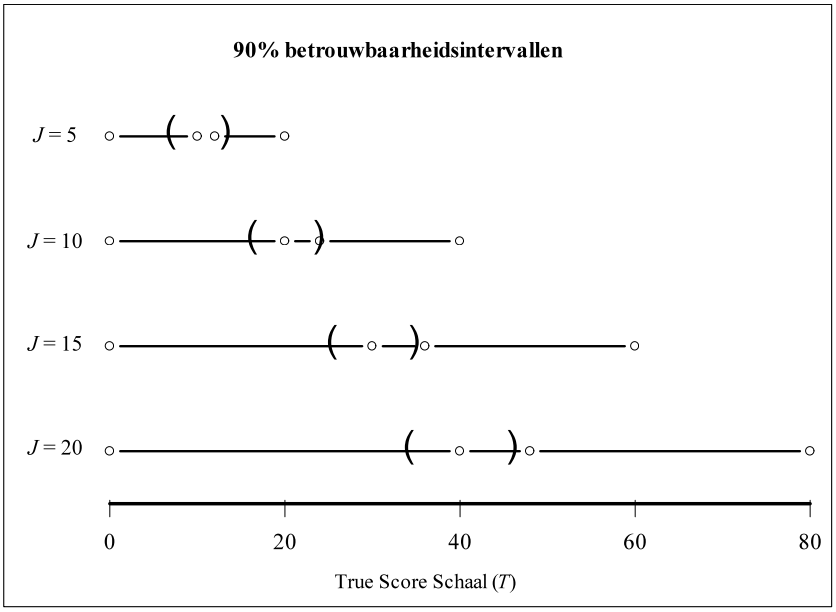
Nauwkeurig meten met een lange toets

Bij een langere toets is de standaardmeetfout S_E ***groter*** en daarmee zijn betrouwbaarheidsintervallen voor T ***langer***. Toch verschilt Jan op een langere toets eerder significant van de cesuur dan op een kortere toetsversie. Dit resultaat is tegenintuïtief, en toch klopt het. Hoe zit dit?
Als een toets langer wordt gemaakt door items toe te voegen, neemt alles toe. De lengte van de schaal neemt toe en de toetsscores nemen toe. Hierdoor liggen de toetsscores verder uiteen en is hun variantie dus groter. Maar ook de varianties van de *true scores* en de meetfouten nemen toe, en dus ook de standaardmeetfout, S_E. Wat vaak slecht wordt begrepen, is dat de variantie van de meetfouten (en dus S_E) langzamer toeneemt dan de varianties van de *true scores* en de toetsscores. Doordat S_E toeneemt, wordt ook het betrouwbaarheidsinterval voor de *true score* langer, maar verschillen tussen *true scores* en verschillen tussen toetsscores lopen dus nog sneller uiteen. Het gevolg is dat bij toetsverlenging verschillen tussen toetsscores van verschillende personen en tussen iemands toetsscore en de cesuur vanzelf significant worden. Als je de toets maar lang genoeg maakt. Het volgende voorbeeld licht dit fenomeen verder toe.

Figuur 1 is gebaseerd op een rekenvoorbeeld en laat zien wat er gebeurt met de standaardmeetfout en het betrouwbaarheidsinterval voor T als het aantal items toeneemt van 5 naar 10, 15 en 20 items (en de standaarddeviatie van de toetsscores, S_X, dus ook toeneemt). Elk item kent scores 0, 1, 2, 3, 4, voor de mate waarin het antwoord juist was. Dus neemt de schaallengte toe van 20, 40, 60 tot 80 punten. Jans toetsscore ligt steeds op 50% van de schaal en de cesuur op 60%, en de afstand in punten tussen die twee posities neemt toe van 2, 4, 6 tot 8 punten. De betrouwbaarheid wordt ook groter (coëfficiënt alfa = .70, .80, .87, .90) maar ook de standaardmeetfout wordt groter (S_E = 1.83, 2.24, 2.76, 3.18). Hierdoor worden de betrouwbaarheidsintervallen langer. Waar het om gaat is dat de snelheid waarmee Jan bij de cesuur wegloopt (als gevolg van een toenemende S_X) groter is dan de snelheid waarmee zijn betrouwbaarheidsinterval toeneemt. Als de toets maar lang genoeg wordt, verdwijnt de cesuur vanzelf uit het betrouwbaarheidsinterval en ligt Jans toetsscore significant onder de cesuur. In figuur 1 is dit bij 15 items voor het eerst zichtbaar.

Hoe kan men de betrouwbaarheid, die nodig is om de standaarmeetfout te schatten, het beste bepalen? Heterogeniteit van de toets kan een lage gemiddelde inter-itemcorrelatie tot gevolg hebben, wat vooral van invloed kan zijn op coëfficiënt alfa (Cronbach, 1951). Een groter aantal items heeft vaak een heilzame werking op alfa, maar er zijn ook methoden die minder gevoelig zijn voor de heterogeniteit van de toets (van der Ark, van der Palm & Sijtsma, 2011; Nunnally, 1978, p. 248).
Men zou ook kunnen overwegen om de heterogeniteit van toetsen te ondervangen door studenten aparte scores te geven voor verschillende, homogene toetsonderdelen. Het eerste argument hiertegen

Studietoetsen dienen vele items te bevatten



Figuur 1 90% betrouwbaarheidsintervallen voor tests bestaande uit resp. 5, 10, 15, en 20 items (J = aantal items); items gescoord 0, 1, 2, 3, 4; cesuur op 60% van de schaal en toetsscore op 50% van de schaal.

is dat elk van die deeltoetsen kort is en dus de hier besproken problemen kent. Het tweede argument tegen is dat uiteindelijk een beslissing over zakken en slagen genomen moet worden, en de deeltoetsscores gecombineerd dienen te worden tot een toetsscore. Dan is men terug bij af.

Een lange toets verkrijgt men het gemakkelijkst met meerkeuze-items. Het maken van goede meerkeuze-items vereist vakmanschap, en veel docenten stellen liever open vragen. Dan is het advies: maak een groot aantal kort-antwoord vragen.

In de psychologie zijn tegenwoordig korte tests in zwang, meestal ingegeven door praktische argumenten als korte testtijd waardoor bijvoorbeeld patiënten minder worden belast. Maar ook financiële overwegingen spelen een rol. Daar is niets mis mee, maar wel blijft staan dat korte tests in vergelijking met langere varianten tot vele foutieve beslissingen kunnen leiden (Sijtsma & Emons, 2007).

Opnieuw: validiteit of betrouwbaar - heid?

Het advies is altijd een representatieve toets te maken op voorwaarde dat deze transparant is en vervolgens via een groot aantal items het aantal onjuiste beslissingen over zakken en slagen binnen de perken te houden. Dus: zorg eerst dat je meet wat je wilt weten en verhoog vervolgens het aantal juiste beslissingen via een groot aantal items. Het minimaal benodigde aantal items is hier helaas niet te geven. Dit minimum aantal hangt af van test- en itemkenmerken, de groep en de cesuur.

Literatuur

Ark, L.A. van der, Palm, D.W. van der & Sijtsma, K. (2011). A latent class approach to estimating test-score reliability. *Applied Psychological Measurement*, 35, 380-392.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.

Drenth, P.J.D. & Sijtsma, K. (2006). *Testtheorie. Inleiding in de theorie van de psychologische test en zijn toepassingen*. Houten: Bohn Stafleu Van Loghum / Springer.

Groot, A.D. de & Naerssen, R.F. van (1969). *Studietoetsen. Construeren. Afnemen. Analyseren*. Den Haag: Mouton.

Hofstee, W.K.B. (1973). Participatie kontrole door "onbenullige" toetsitems. *Nederlands Tijdschrift voor de Psychologie*, 28, 189-198.

Linden, W.J. van der & Hambleton, R.K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Nunnally, J.C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Sijtsma, K. & Emons, W.H.M. (2007). Korte tests: Kostbare tijdwinst en onbetrouwbare beslissingen. *De Psycholoog*, 42, 406-411.

☐ Prof.dr. K. Sijtsma is hoogleraar methoden van psychologisch onderzoek aan de Tilburg School of Social and Behavioral Sciences, Department of Methodology and Statistics. E-mail: k.sijtsma@uvt.nl.