



Een toets die 100%
betrouwbaar is, komt in
de praktijk niet voor.

De relatie tussen toetsbetrouwbaarheid, classificatiefouten en een beoordelingsschaal

■ Henk van Berkel

Dr. mult. H. J. M. (Henk) van Berkel is schrijver en redacteur van verschillende boeken en artikelen op het gebied van toetsen, kwaliteitszorg en probleemgestuurd onderwijs. Sinds 2013 is Henk met pensioen. In 2019 is hij (wederom) gepromoveerd, ditmaal op een juridisch onderwerp: uitspraken van beroepscolleges in het hoger onderwijs. E-mail: henkvanberkel22@gmail.com.

Je construeert een aantal toetsvragen en zorgt ervoor dat deze goed verdeeld zijn over de leerdoelen. Dan heb je een toets. Die neem je af en telt het aantal punten dat de studenten hebben behaald, eventueel na correctie voor raden indien de toets bestond uit meerkeuzevragen. Dat aantal punten zet je om in een cijfer en klaar is Kees. Ziedaar in een notendop het toetsproces. Moet daar nu zoveel heisa om worden gemaakt? Je weet toch hoeveel toetsvragen studenten correct hebben gemaakt? Dat heb je met de toets gemeten. Dus je weet ook hoeveel kennis ze hebben. Daar is toch geen speld tussen te krijgen?

Zoals met veel zaken is niets zo eenvoudig als het lijkt. Natuurlijk kan een docent of examiner makkelijk tellen wat de score is op een toets. Maar het is naïef om die score klakkeloos te interpreteren als hét kennisniveau van de deelnemers. Toetsen zijn vormen van metingen, en net als veel andere metingen zijn ze bijna nooit 100% betrouwbaar. In de toetsliteratuur wordt de betrouwbaarheid, naast de inhoudsvaliditeit, gezien als de belangrijkste beschikbare indicatie voor de kwaliteit van een toets. Als een toets perfect betrouwbaar is, wil dat zeggen dat onwenselijke factoren als vermoeidheid, toetsangst, gokgedrag, een verkeerde verdeling van de vragen over de leerdoelen, nakijkfouten, onduidelijke vragen, storende toetsomstandigheden et cetera, zich niet hebben voorgedaan of niet aanwezig zijn

geweest. Geen enkele toets is echter perfect betrouwbaar. Altijd schort er wel wat aan. Vandaar dat het belangrijk is de betrouwbaarheid van een toets te berekenen. De betrouwbaarheid geeft aan in hoeverre je vertrouwen kunt hebben in de toets als meetinstrument. Meet de toets echt iets, of kun je net zo goed een muntstukje opwerpen om uitspraken te doen over het kennisniveau van de deelnemers? Als een persoon 's ochtend op de weegschaal staat en de meter geeft 85 kilo aan en vijf minuten later met dezelfde persoon erop 90 kilo, is de weegschaal niet betrouwbaar.

Wanneer een toets onbetrouwbaar is, heeft de uitslag geen betekenis. Wanneer een toets 100% betrouwbaar is, is er in ieder geval de zekerheid dat de uitslag iets heeft te beteke-

nen. Alle tussenliggende waarden gaan gepaard met een onzekerheid: hoe dichter de betrouwbaarheid naar het minimum toegaat (de range van de meest gangbare maat voor de betrouwbaarheid, Cronbachs alfa, loopt van 0 tot 1), des te onzekerder is het dat de uitslag wat voorstelt. Met andere woorden, er ligt een bandbreedte, een interval, rond de score: hoe groter de onbetrouwbaarheid, des te breder de bandbreedte. In verkiezingstijd is een dergelijke formulering gebruikelijk. Prognoses van kiesuitslagen kennen een bepaalde (on)betrouwbaarheid. De grenzen worden dan aangegeven. Bijvoorbeeld: het zetelaantal van partij X zal tussen de 40-45 uitkomen. Er is dus steeds sprake van een meetfout. Die meetfout van een toets, en ook de betrouwbaarheid, is te berekenen.

Classificatiefouten

Het zal duidelijk zijn dat de mate van zekerheid vooral belangrijk is voor toetsscores in de buurt van de zak-/slaaggrens, de cesuur. Daar dient het cijfer heel nauwkeurig te zijn. Maar zoals eerder geschreven, iedere meting heeft een bandbreedte om zich heen. Dat betekent dat wanneer een student een 5,4 behaalt (en volgens veel examenregelingen dus is gezakt) dat eigenlijk ook een 5,6 kan zijn geweest, en dus een voldoende. Hier is, zoals dat heet, een classificatiefout gemaakt. Deze redenering werkt natuurlijk ook naar de andere kant: een student is dan ten onrechte geslaagd. Dat is ook een classificatiefout.

Om te voorkomen dat er relatief veel misclassificaties optreden, moet de betrouwbaarheid

zo hoog mogelijk zijn. Een toets die 100% betrouwbaar is, komt in de praktijk niet voor.

Er zal dus een soort compromis moeten worden gevonden ten aanzien van wat nog aanvaardbaar is. Het compromis zal rekening moeten houden met het belang van de beslissing. De vuistregel over de wenselijke hoogte van de betrouwbaarheid in relatie tot de soort beslissing die op basis van de toetsuitslag wordt genomen, staat in tabel 1.

Toetsen die kunnen worden gecompenseerd met een andere toets, mogen op zichzelf een relatief lage betrouwbaarheid bezitten (mits hoger dan 0,70), omdat mag worden verondersteld dat beide toetsen samen genomen een acceptabele betrouwbaarheid bezitten, hoger dan 0,80. In het algemeen geldt namelijk: hoe meer toetsvragen, des te hoger de betrouwbaarheid. Een uitslag van een toets met een betrouwbaarheid van lager dan 0,60 heeft in het algemeen geen waarde.

Cijfers

Cijfers vervullen in het onderwijs een cruciale rol. Cijfers geven studenten inzicht in het resultaat van hun studie-inspanningen. Cijfers weerspiegelen ook de mate waarin studenten de leerdoelen beheersen. Cijfers vormen het uitgangspunt voor talloze beslissingen: zakken of slagen, overdoen of niet, afgewezen worden of geselecteerd, bijleren of ervan af zijn, et cetera. Daarnaast zijn cijfers bruikbaar om studenten te stimuleren en te belonen.

Tabel 1. Minimaal wenselijk hoogte van de betrouwbaarheid in relatie tot beslissing

Beslissing	Minimale betrouwbaarheid
High stake toets	0,90
Toets zonder compensatiemogelijkheden	0,80
Toets met compensatiemogelijkheden	0,70
Formatief toets	0,60

Gezien het belang van cijfers moet het proces van cijfergeven valide zijn. Dat betekent onder andere dat studenten die de leerdoelen het beste beheersen, ook de hoogste cijfers moeten krijgen en dat het cijfer een correcte weergave is van het kennis- of vaardigheden-niveau. Maar dat is lang niet altijd de praktijk, gezien de hoogte van de betrouwbaarheid van een toets en de daaraan gepaarde bandbreedte. De interpretatie die men kan geven aan het cijfer wordt daardoor minder eenduidig.

Het kiezen van een beoordelings-schaal

Iedere behaalde score op een toets kent dus een onzekerheidsmarge. Een student behaalt bijvoorbeeld een score van 70 op een maximum van 100. Doordat de toets niet honderd procent betrouwbaar is geweest, is het niet zeker dat de score inderdaad precies 70 is. Het zou kunnen zijn dat de bandbreedte 64-76 is. Dit is gezien de gebruikelijke betrouwbaarheid beslist geen onrealistische bandbreedte.

Wanneer de docent de score classificeert op een tienpuntsschaal, verkrijgt de student met een score van 70 een '7'. Maar gezien de meetfout kan de (afgeronde) score, statistisch gezien, net zo goed een '6' zijn als een '8'. Er treden dus classificatiefouten op. Als de docent echter een tweepuntsschaal hanteert met als aanduiding voldoende/onvoldoende, is de score van 70 een voldoende, evenals de statistisch mogelijke scores 64 en 76. Er zijn dan geen classificatiefouten gemaakt. Er geldt: het aantal misclassificaties neemt toe naarmate het aantal preciseringen op de beoordelingsschaal toeneemt, tenzij de toets 100% betrouwbaar is. Maar zoals al is geschreven is dat laatste nooit het geval.

Maar hoe gedetailleerd moet een beoordelingsschaal zijn? Een tweepuntsschaal? Een vierpuntsschaal? Een tienpuntsschaal of een honderdpuntsschaal? De laatste categorie komt vaak voor. Het gaat dan bijna altijd om een tienpuntsschaal met een cijfer achter de



Een uitslag van een toets met een betrouwbaarheid van lager dan 0,60 heeft in het algemeen geen waarde.

komma, bijvoorbeeld 6,7. Door een decimaal te hanteren bij een tienpuntsschaal, ontstaat een honderdpuntsschaal.

Het aantal foutieve classificaties is dus evenredig met het aantal preciseringen c.q. cijfers op de beoordelingsschaal. Hoe meer cijfers op de schaal, hoe meer classificatiefouten. Een goede cijferschaal houdt rekening met de te verwachten (on)betrouwbaarheid van de toets. En classificatiefouten moeten zoveel mogelijk worden vermeden. Hoewel de betrouwbaarheid van een toets vóór de toetsafname niet bekend is, kan een docent met behulp van analyses op voorgaande toetsen een indicatie krijgen welke betrouwbaarheid mag worden verwacht.

Welke beoordelingsschaal een docent vervolgens moet hanteren, is af te lezen uit tabel 2.

Tabel 2. Verwachte betrouwbaarheid van de toets en het aantal punten op de beoordelingsschaal

Betrouwbaarheid	Beoordelingsschaal
0,90-1,00	Twintigpunten of minder
0,80-0,90	Tienpunten of minder
0,70-0,80	Vijfpunten of minder
0,60-0,70	Tweepunten
Lager dan 0,60	Geen, gezien de (te) lage betrouwbaarheid

De achterliggende gedachte bij de tabel is dat een docent het aantal classificatiefouten wil beperken. Dat houdt in dat ervoor moet worden gezorgd dat zich bij de verschillende punten op de beoordelingsschaal zo min mogelijk personen bevinden die aan de verkeerde kant zijn terechtgekomen, dus personen die onterecht een te laag of te hoog cijfer hebben gekregen. Zou een docent een beoordelingsschaal met twintig punten hanteren zonder te letten op de betrouwbaarheid van de toets, dan zullen bij een toets met een mindere betrouwbaarheid relatief veel studenten ten onrechte slagen of zakken en/of een verkeerd cijfer krijgen. Een manier om het aantal misclassificaties te verminderen, is het aantal punten op een beoordelingsschaal omgekeerd evenredig te laten zijn met de hoogte van de betrouwbaarheid. In de tabel is dat voorbeeldmatig uitgewerkt. De tabel is geen wet van Meden en Perzen. Er kunnen redenen zijn om ervan af te wijken, maar het principe blijft staan: hoe hoger de betrouwbaarheid, des te meer cijfers een beoordelingsschaal mag bevatten.

In Nederland is het gebruikelijk dat docenten een tienpuntsschaal hanteren. Dat zou geen probleem zijn als de betrouwbaarheid van de toets hoger is dan 0,80. Die waarde is echter niet gebruikelijk. In de praktijk van de kleinschalige toetsen in het onderwijs zullen de meeste betrouwbaarheden daaronder liggen. Dat opleidingen toch kiezen voor een tienpuntsschaal, heeft alles te maken met de

traditie. Dat gold in ieder geval in het (verre) verleden. Maar gaandeweg de tijd is daar verandering in gekomen. Eerst werden aan hele cijfers halfjes toegevoegd, toen kwamen de decimalen en in de havo zijn inmiddels de hectomalen (op een duizendpuntsschaal!) ingevoerd. Het zal duidelijk zijn dat vanuit toetskundig oogpunt dit een ontwikkeling is, die niet verdedigbaar is.

Conclusie

De betrouwbaarheid van de toets is bepalend voor de beoordelingsschaal die men hanteert. Een hoge betrouwbaarheid laat een beoordelingsschaal toe met veel cijfers of decimalen. Toetsen in het onderwijs die zijn geconstrueerd door docenten, hebben doorgaans een betrouwbaarheid die zich beweegt tussen 0,70 en 0,80. Dat impliceert een schaal met maximaal vijf punten, bijvoorbeeld een schaal die loopt van één tot vijf. Het kunnen echter ook letters zijn: A, B, C, D en E, zoals in de Angelsaksische landen gebruikelijk is. Voor andersoortige toetsen, bijvoorbeeld een paper of een vaardigheidstoets, zal de betrouwbaarheid lager zijn. Dat is eigen aan dit soort toetsen. Het ligt voor de hand om voor dit soort toetsen een andere cijferschaal te gaan hanteren, bijvoorbeeld een twee- of driepuntsschaal. In ieder geval moet duidelijk zijn dat vanuit toets-technische overwegingen de meeste schalen die docenten hanteren ten onrechte fijnmazig zijn. Ze suggereren een exactheid die verder gaat dan de toetskwaliteit toelaat. Het aantal classificatiefouten neemt daardoor toe. ■



Congres

Examencommissies in het hoger onderwijs

Woensdag 11 december 2024

Plenaire opening

Actuele wetgeving en fraude (met AI)



2 subsessies naar keuze:

Borgen van toetskwaliteit

A

De rol van examencommissies bij
programmatisch toetsen - **NIEUW**

B

Klachten, verzoeken en
onregelmatigheden juridisch bekeken

C

Jurisprudentie – leren van het verleden

D

Toegankelijk toetsen en de rol van de
examencommissie - **NIEUW**

E