

Statistiek en meten: wat moet je daarover weten?

Onder redactie van Diana Dolmans, Cees van der Vleuten, Albert Scherpbier en Ineke Wolfhagen

Bij het lezen van publikaties over onderzoek van onderwijs worden lezers regelmatig geconfronteerd met allerlei statistische begrippen, waarvan de betekenis veel lezers niet geheel duidelijk is. Daarom heeft de redactie besloten een reeks artikelen te publiceren waarin belangrijke onderwerpen op het gebied van statistiek en meten aan de orde worden gesteld. De redactie van deze reeks wordt gevormd door twee gastredacteuren (Diana Dolmans en Cees van der Vleuten) en twee leden van de BMO-redactie (Albert Scherpbier en Ineke Wolfhagen).

De keuze van onderwerpen is gebaseerd op veel gebruikte statistische technieken en begrippen. De nadruk ligt op de betekenis en interpretatie hiervan. Op deze manier wordt getracht een bijdrage te leveren aan de verdere professionalisering van docenten binnen het medisch onderwijs. In het zevende artikel van deze reeks staat betrouwbaarheid centraal.

Klassieke testtheorie en betrouwbaarheid van meetgegevens

M.C. Pollemans, H. Düsman

Termen die aan bod komen:

betrouwbaarheid, betrouwbaarheidscoëfficiënt, paralleltestmethode, test-hertestmethode, splitsingsmethode, interne consistentie, standaardmeetfout.

In het dagelijks leven wordt veelvuldig gebruik gemaakt van testen en toetsen. Direct na de geboorte worden al APGAR-scores bepaald en bij deze eerste meting blijft het voor de meeste mensen niet. Ontwikkelingstesten, schoolvoorderingstoetsen, toelatingsexamens, eindexamens, intelligentietesten en beroepskeuzetesten volgen elkaar in snel tempo op, om nog maar niet te spreken van de vragenlijsten die men voorgelegd krijgt om een opvatting, houding of belangstelling vast te stellen. Testen en toetsen zijn niets bijzonders en worden wijd en zijd gebruikt.¹ Dat wil niet zeggen dat in de praktijk altijd even goed kan worden beoordeeld of de testen en toetsen wel aan veronderstelde eisen voldoen. Het is niet eenvoudig om goede toetsen te maken. Vanaf het begin van

deze eeuw zijn er theorieën en methoden ontwikkeld om te trachten betere toetsen te maken.

In dit artikel wordt een belangrijk kenmerk van toetsen besproken, namelijk de betrouwbaarheid. De betrouwbaarheid van een toets is een van de onderwerpen van de psychometrie, de leer van het meten van menselijke eigenschappen. Aan de orde komen de betekenis van het begrip meetfout en de definitie van betrouwbaarheid. Verder worden enkele eenvoudige te hanteren maten voor betrouwbaarheid besproken. In dit artikel wordt het begrip betrouwbaarheid aan de hand van toetsen toegelicht, maar het kan uiteraard ook breder toegepast worden.

Toetsen zijn meetinstrumenten die met name in het onderwijs worden gebruikt. Het zijn pogingen om kenmerken van bijvoorbeeld leerlingen, studenten of cursisten op een gestandaardiseerde wijze te representeren. Meestal worden waarden toegekend aan de prestaties die geleverd worden. De prestatie wordt omgezet in een 'score', die een indicatie

is voor het niveau waarop de student de leerstof beheerst of de mate waarin hij of zij een bepaalde vaardigheid bezit.

Toetsen zoals examens, tentamens en dergelijke bestaan meestal uit een aantal onderdelen (vragen, opdrachten, taken). Er zijn verschillende methoden om tot een totaalscore op een toets te komen. In dit artikel wordt uitgegaan van de meest gebruikte methode, waarbij de score de som of het gemiddelde is van de punten die de toetsdeelnemers op afzonderlijke opdrachten hebben behaald. De theorie van de betrouwbaarheid is toepasbaar op vele soorten toetsen. In dit artikel wordt de theorie echter toegelicht aan de hand van voorbeelden van toetsen met meerkeuzevragen. In het taalgebruik van deskundigen worden toetsvragen meestal items genoemd.

Validiteit en betrouwbaarheid

De kwaliteit van toetsen wordt vaak weergegeven aan de hand van de validiteit en de betrouwbaarheid. Validiteit of geldigheid gaat over wát er wordt gemeten: worden er met een rekentoets wel rekenvaardigheden getoetst en niet bijvoorbeeld leesvaardigheid? Is de toets-score wel een weergave van de mate waarin iemand kennis kan toepassen, of is deze volledig afhankelijk van hoe goed iemand kennis weet te reproduceren?² Een betrouwbare toetsscore is consistent, dat wil zeggen dat een herhaling van de meting onder gelijke condities dezelfde resultaten zou opleveren.

Validiteit en betrouwbaarheid zijn verwante begrippen. Ze zijn als volgt met een voorbeeld te verduidelijken. Bij een eindtoets van een cursus geschiedenis worden honderd vragen gesteld, allemaal betrekking hebbend op jaartallen. De scores voor het eerste, tweede, derde en vierde 25-tal vragen zijn allemaal vrijwel gelijk aan de score voor het totaal van honderd vragen: de toets meet dus betrouwbaar. Echter, de cursus was niet gericht op het leren reproduceren van jaartallen, maar op het verwerven van algemene kennis en begrip van

geschiedenis. De toetsinhoud is in dit geval niet representatief voor de inhoud van de cursus; met andere woorden de toets is niet inhoudsvalide.

De betrouwbaarheid is van belang voor de gebruikswaarde van een meetinstrument. Een bruikbaar meetinstrument moet consistente resultaten opleveren: personen die de ene keer hoog scoren, zouden dat onder vergelijkbare omstandigheden, de andere keer ook moeten doen. In het algemeen geldt dat naarmate de resultaten over verschillende metingen consistent zijn, de betrouwbaarheid groter is.

Meetfouten

Hoe goed de toets ook wordt voorbereid, er zal altijd sprake zijn van fouten in het resultaat. Met andere woorden, scores wijken altijd af van de scores die eigenlijk gevonden hadden moeten worden. Deze afwijkingen zijn onder te verdelen in twee typen: systematische fouten en toevallige fouten. Men spreekt van een systematische fout als de score van alle deelnemers evenveel te hoog of te laag is. Systematische fouten kunnen gemakkelijk worden geaccepteerd als de scores van de deelnemers alleen met elkaar vergeleken worden. Alle overige afwijkingen, dus de afwijkingen die niet voor iedere deelnemer hetzelfde effect hebben, worden beschouwd als toevallige fouten. Bronnen van toevallige fouten zijn bijvoorbeeld de omstandigheden waaronder wordt getoetst, zoals het tijdstip van de dag, geluidshinder of de temperatuur in de toetsruimte. Deze omstandigheden kunnen voor de deelnemers aan de toets per persoon een verschillende invloed hebben, bijvoorbeeld tengevolge van iemands gezondheidstoestand, mate van vermoeidheid of honger. Iemand kan ook even een black-out hebben. Toevallige fouten kunnen ook worden veroorzaakt door de instructie die bij de toets wordt gegeven of door slecht geformuleerde items. Zou men toetsen onder precies dezelfde omstandigheden kunnen herhalen, dan zouden in principe voor de deelnemers

dezelfde scores gevonden moeten worden. Verschillen die dan zouden optreden tussen de resultaten beschouwt men als toevallige meetfouten.

Om aan te geven hoe groot het aandeel van de toevallige fouten in een toetsscore is, wordt meestal het begrip 'betrouwbaarheid' gebruikt. Dit begrip vormt een centraal onderdeel van de 'klassieke testtheorie'. Met behulp van de klassieke testtheorie is het mogelijk om de invloed van toevallige meetfouten op de uiteindelijke scores te schatten. De wijze waarop dat precies gebeurt, komt later in dit artikel aan de orde. Eerst wordt ingegaan op wat de klassieke testtheorie is.

Klassieke testtheorie

Er worden meestal drie theoretische benaderingen onderscheiden die inzicht geven in de betrouwbaarheid van toetsscores: de klassieke testtheorie, de generaliseerbaarheidstheorie en de item-respons theorie.³ In deze bijdrage wordt uitsluitend de klassieke testtheorie besproken.

De klassieke testtheorie is vanaf het begin van deze eeuw ontwikkeld. Ze is de oudste en ook de eenvoudigste theorie en voor veel toepassingen nog steeds goed toereikend. Het is niet de bedoeling om de theorie volledig te beschrijven, maar om eenvoudige toepassingen van de theorie voor onderwijssituaties uit te leggen.

Omdat de scores van elke toets door meetfouten worden beïnvloed, kan onderscheid worden gemaakt tussen wat daadwerkelijk wordt gemeten en wat men had willen of moeten meten. Natuurlijk is het de bedoeling dat iemands toetsscore de mate representeert waarin die persoon een bepaald kenmerk werkelijk bezit. De waarde die dit laatste weergeeft, wordt uitgedrukt met het begrip 'ware' score. Dat is echter niet de geobserveerde of 'waargenomen' score. Deze is de som van de ware score en de (toevallige) meetfout. De relatie wordt als volgt weergegeven:

$$O = W + E$$

O = de geobserveerde of waargenomen score

W = de ware score

E = de meetfout, het toevallige deel van de score

De klassieke testtheorie gaat ervan uit dat iemands ware score overeenkomt met de gemiddelde score die deze persoon zou behalen als er een oneindig aantal keer opnieuw getoetst zou worden. De gemiddelde afwijking (meetfout) is dan dus nul. Zowel de ware score als de meetfout zijn hypothetische begrippen. In de praktijk is het immers onmogelijk ooit te beschikken over een oneindig aantal herhaalde metingen. Het is zelfs niet mogelijk een meting één keer te herhalen, zonder dat er iets verandert aan de condities waaronder gemeten wordt.

De betrouwbaarheid wordt uitgedrukt in een *betrouwbaarheidscoëfficiënt*. Deze kan variëren tussen de waarden 0 en 1. Een grotere meetfout zal resulteren in een lagere betrouwbaarheidscoëfficiënt. Naarmate de betrouwbaarheid dichter bij 1 ligt, is de meetfout geringer.

Er zijn verschillende methoden ontwikkeld om de betrouwbaarheid te schatten. Deze schattingsmethoden zijn onder te verdelen in twee groepen: methoden waarbij gebruik gemaakt wordt van de gegevens van herhaalde metingen, en methoden waarbij schattingen van de betrouwbaarheid worden verkregen aan de hand van de gegevens van één toetsafname. Achtereenvolgens wordt kort ingegaan op de paralleltestmethode, test-hertestmethode, splitsingsmethode en interne consistentie.

Bij de *paralleltestmethode* wordt de betrouwbaarheid geschat door de correlatie te bepalen tussen twee parallelle metingen. Een veel gebruikte maat om de samenhang tussen twee metingen uit te drukken is de Pearson's produkt moment correlatie. Deze maat is in een vorig artikel in deze reeks al uitvoerig aan de

orde geweest.⁴ Bij de paralleltestmethode is de betrouwbaarheid gelijk aan de correlatie tussen de scores op twee parallelle toetsen. Deze toetsen hebben beide betrekking op dezelfde inhoud, maar de items van beide toetsen zijn verschillend. Volgens de klassieke testtheorie mag men twee toetsen parallel noemen als de gemiddelden en standaarddeviaties gelijk zijn, en de correlatie van de afzonderlijke toetsscores met een of meer andere variabelen gelijk is.

Om een schatting te maken van de betrouwbaarheid moeten twee parallelle toetsen worden gemaakt en in volledig identieke situaties worden afgenomen, waarna de correlatie tussen de scores op de beide toetsen kan worden bepaald. Het schatten van de betrouwbaarheid met parallelle toetsen levert praktische problemen op. Het is vrijwel onmogelijk twee volledig parallelle toetsen te produceren en deze tevens onder identieke omstandigheden af te nemen.

Bij de *test-hertestmethode* wordt een schatting van de betrouwbaarheid verkregen door tweemaal dezelfde toets af te nemen en vervolgens de correlatie tussen de scores op deze twee toetsen te berekenen. Het belangrijkste probleem hierbij is dat de toetsen onder dezelfde omstandigheden moeten worden afgenomen. Het gebruik van identieke toetsen introduceert het probleem dat met geheugenfactoren rekening moet worden gehouden.

Bij de *splitsingsmethode* wordt een schatting van de betrouwbaarheid verkregen door de toets in twee helften te splitsen en vervolgens de correlatie tussen de scores op de twee helften te berekenen. Een eenmaal afgenomen toets kan achteraf natuurlijk op allerlei manieren in tweeën worden gedeeld. Evenals bij de eerder genoemde methoden is het van belang dat de twee 'nieuwe' toetsen zo gelijk mogelijk aan elkaar zijn. Vaak neemt men de vragen met de even nummers voor de ene toets en die met de oneven nummers voor de andere toets. Omdat er een relatie is tussen de lengte van de toetsdelen en de correlatie van de scores, moet

de schatting van de betrouwbaarheid die volgens deze methode plaatsvindt, worden aangepast aan de lengte van de oorspronkelijke gehele toets. Daarvoor zijn speciale correctieformules ontwikkeld, zoals bijvoorbeeld de Spearman-Brown formule.⁵

Bij de *interne consistentie* wordt een toets opgesplitst in meerdere gelijke delen. Uiteindelijk komt men dan terecht op het niveau van de afzonderlijke items en berekent men in hoeverre van item tot item hetzelfde gemeten wordt. Een veel gebruikte maat hiervoor is Cronbach's α (alfa) en de KR-20.⁵

Normen voor betrouwbaarheidscoëfficiënten

De betrouwbaarheid van toetsen wordt geschat om aan te geven hoe groot de toevallige meetfouten zijn. Bij het beoordelen van individuele prestaties moet een relatief grote meetfout ernstiger worden genomen dan wanneer de toetsscores worden gebruikt om algemene vraagstellingen te beantwoorden. Bij dergelijke vraagstellingen gaat het immers vaker over de betekenis van gemiddelde scores van groepen.

In de literatuur bestaat redelijke overeenstemming over het hanteren van een ondergrens voor de betrouwbaarheid van .80 of .85, wanneer de toets wordt gebruikt voor het beoordelen van individuen.⁶ Valt de geschatte betrouwbaarheid lager uit dan dient hiermee rekening gehouden te worden, zeker als het oordeel belangrijke consequenties heeft voor de deelnemers aan de toets. Worden gegevens van meerdere toetsen meegewogen, dan kan een afzonderlijke toets met een betrouwbaarheid van .60 nog wel een zinvolle bijdrage leveren.

Als toetsgegevens worden gebruikt voor onderzoeksdoeleinden gaat de interesse meestal uit naar groepsgemiddelden. Het groepspectief levert sneller een betrouwbaar resultaat, omdat gemiddelden over groepen stabiel zijn dan gemiddelden over individu-

en. De betrouwbaarheid van gemiddelden neemt met de groepsomvang toe.

Vaak worden toetsen gebruikt om te beslissen of de deelnemers aan een bepaalde norm voldoen. Als wordt gekozen voor een grenscore of cesuur, aan de hand waarvan wordt bepaald of iemand slaagt of zakt, dan is er tengevolge van de meetfout altijd een bepaalde kans dat deelnemers ten onrechte zakken of slagen. De betrouwbaarheid is dus van invloed op het aantal verkeerde beslissingen dat in de praktijk wordt genomen. Als uitgegaan kan worden van een normale scoreverdeling, kan worden berekend op hoeveel foute beslissingen moet worden gerekend bij een bepaald slaagpercentage en een bepaalde betrouwbaarheid. Bij een toets met een betrouwbaarheid van .80 en bij een zakpercentage van 40, is berekend dat 20% van de beslissingen fout is. Dat betekent dat 10% van de deelnemers ten onrechte zakt en 10% ten onrechte slaagt.⁵ Een norm van .80 is dus betrekkelijk en betekent niet dat de toetsscores nauwelijks meetfouten bevatten. Meer inzicht in deze problematiek wordt geboden door de standaardmeetfout.

Standaardmeetfout

De *standaardmeetfout* (= Standard Error of Measurement of SEM) kan gebruikt worden om na te gaan in hoeverre een score die één persoon behaalt op één bepaald moment als 'ware' score beschouwd kan worden. De geobserveerde score wordt immers beïnvloed door allerlei toevallige factoren. Tengevolge van deze toevallige factoren zal een persoon op dezelfde toets de ene keer een lagere score halen dan de andere keer. Indien dezelfde toets bij dezelfde persoon een aantal malen onder gelijke condities zou worden afgenomen, zouden de scores een normale verdeling benaderen. De spreiding die deze verdeling van scores zou vertonen, noemen we de standaardmeetfout. De standaardmeetfout geeft dus de gemiddelde onnauwkeurigheid van een individuele score weer. Als de standaardmeetfout groot

is, dan meet de toets niet betrouwbaar. De standaardmeetfout kan gebruikt worden voor het schatten van het betrouwbaarheidsinterval van een individuele score. Met 68% zekerheid bevindt de ware score zich tussen de gemiddelde geobserveerde score minus één standaardmeetfout en de gemiddelde score plus één standaardmeetfout. Met 95% zekerheid bevindt de ware score zich tussen de gemiddelde geobserveerde score minus tweemaal de standaardmeetfout en de gemiddelde score plus tweemaal de standaardmeetfout. Uitgaande van een waargenomen score van 70 en een standaardmeetfout van 4, bestaat er een kans van 68% dat de ware score ligt tussen 66 en 74. Met een kans van 95% ligt de ware score tussen 62 en 78. Bij de standaardmeetfout wordt dus gekeken naar de betrouwbaarheid van individuele scores.

De standaardmeetfout moet niet verward worden met de standard error (Se), waarvoor geen Nederlands woord bestaat. De standard error heeft betrekking op de nauwkeurigheid van het steekproefgemiddelde en dus niet op een individuele score.⁷ Een belangrijk verschil is dat de standard error wel wordt beïnvloed door het aantal personen dat in de berekeningen wordt betrokken, maar de standaardmeetfout niet.

Factoren die schattingen van de betrouwbaarheid beïnvloeden

Er zijn meerdere factoren die de hoogte van de schatting van de betrouwbaarheid beïnvloeden, zoals de toetsinhoud, de eigenschappen van items, de homogeniteit van de groep en de toetslengte. Deze factoren worden achtereenvolgens kort toegelicht.

Toetsinhoud. Een toets waarin meerdere onderwerpen behandeld worden of waarmee verschillende vaardigheden worden getoetst, is, bij gelijke lengte, meestal minder betrouwbaar dan een toets over één bepaald onderwerp of vaardigheid. Om de kennis van studenten over

een breed domein te toetsen, zijn uiteraard meer vragen nodig dan wanneer wordt beoogd kennis over een smaller en meer afgebakend domein te toetsen. Zolang er sprake is van eenzelfde domein, geldt dat hoe meer items in de toets worden opgenomen, des te betrouwbaarder de uitkomst zal zijn.

Eigenschappen van items. Twee meeteigenschappen van items zijn van belang voor de betrouwbaarheid van een toets: de moeilijkheidsgraad en de discriminatiewaarde van de items. Optimale items zijn niet te moeilijk en niet te gemakkelijk, en onderscheiden de zwakkeren goed van de beteren. Sterk discriminerende items (bijvoorbeeld afgemeten aan item-test of item-rest correlaties) dragen het meest bij aan de betrouwbaarheid.

Homogeniteit van de groep. De samenstelling van de groep respondenten is ook van invloed op de betrouwbaarheid. Als de deelnemers aan een toets sterk verschillen in de mate van beheersing van de leerstof, valt de betrouwbaarheid hoger uit dan in het tegenovergestelde geval.

Toetslengte. Bij de splitsingsmethode is de samenhang tussen de lengte van een toets en de betrouwbaarheid aan de orde geweest. In het algemeen geldt dat hoe langer de toets is, hoe betrouwbaarder de scores zijn. Met behulp van de Spearman-Brown formule is deze relatie vastgelegd. Deze formule is van toepassing als de toets wordt verlengd of verkort met vragen of opgaven die wat betreft inhoud en moeilijkheidsgraad overeenkomen met de items van de oorspronkelijke toets.

Literatuur

1. Suen HK. Principles of test theories. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1990.
2. Van Leeuwen YD. Validiteit. Bulletin Medisch Onderwijs, 1994; 13: 68-72.
3. Drenth PJD, Sijtsma K. Testtheorie. Inleiding in de theorie van de psychologische test en zijn toepassingen. Houten/Zaventem: Bohn, Stafleu, Van Loghum, 1990.
4. Smal JA. Associatie en correlatie. Bulletin Medisch Onderwijs 1993; 12: 110-6.
5. Dousma T, Horsten A. Tentamineren. Groningen: Wolters-Noordhoff, 1989.
6. Eggen TJHM, Sanders PF. Psychometrie in de praktijk. Arnhem: Cito Instituut voor Toetsontwikkeling, 1993.
7. Bender W. Item-analyse. Bulletin Medisch Onderwijs 1994; 13: 37-43.

Aanbevolen literatuur

- Dousma T, Horsten A. Tentamineren. Groningen: Wolters-Noordhoff, 1989.
- Drenth PJD, Sijtsma K. Testtheorie. Inleiding in de theorie van de psychologische test en zijn toepassingen. Houten/Zaventem: Bohn, Stafleu, Van Loghum, 1990.

Opdrachten

1. Waarom zijn onbetrouwbare toetsen in het algemeen niet valide, waardoor het onmogelijk is te meten wat men wil meten?
2. Geef bij de volgende oorzaken van meetfouten aan of ze systematisch of toevallig zijn:
 - a. De toets blijkt zo lang dat een aantal cursisten in tijdnood komt.
 - b. Enkele toetsboekjes missen een pagina.
3. Maakt het voor de hoogte van de betrouwbaarheid uit hoe groot het aantal deelnemers aan de toets is?
4. Een docent heeft een toets gemaakt en afgenomen in twee groepen van verschillend niveau. Hij schat de betrouwbaarheid drie keer. Twee keer aan de hand van de gegevens van de groepen afzonderlijk. En een keer aan de hand van de gegevens van beide groepen tezamen. De laatste schatting blijkt een hogere betrouwbaarheid op te leveren dan de eerste twee. Hoe is dat te verklaren?

DE AUTEURS

M.C. Pollemans, arts / onderwijskundige en H. Düsman, methodoloog, zijn beiden werkzaam bij het Samenwerkingsverband Universitaire Huisartsopleidingen aan de Universiteit Utrecht

Correspondentie-adres:

M.C. Pollemans, SVUH, Universiteit Utrecht, Universiteitsweg 100, 3584 CG Utrecht