

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 om 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 zesde artikel van deze reeks staat de validiteit centraal.

Validiteit

Y.D. van Leeuwen

Termen die aan bod komen:

validiteit, inhoudsvaliditeit, blauwdruk, key features, criteriumvaliditeit, convergente validiteit, divergente validiteit, predictieve validiteit, constructvaliditeit

Validiteit en betrouwbaarheid zijn kernbegrippen waar het toetsen betreft. Het begrip betrouwbaarheid zal in het volgende artikel in deze serie aan de orde komen. In dit artikel staat het begrip validiteit centraal. Validiteit betekent letterlijk 'geldigheid'. Deze vertaling is treffender dan men in eerste instantie zou denken, hetgeen later in deze uiteenzetting zal blijken. Inzichtelijker is echter de uitleg: "een valide toets is een toets die meet wat men bedoelt te meten". Kortom, we willen een antwoord krijgen op de vraag: Welke toets is geschikt als ik dit begrip wil meten? Het zoeken naar een antwoord op deze vraag is validiteitsonderzoek. In feite heeft het begrip 'validiteit' niet alleen betrekking op toetsen zoals die in het onderwijs worden gebruikt, maar ook

op (psychologische) tests en andere meetinstrumenten. Ook bij klinische tests, zoals de bloedsuikertest, kan men van validiteit spreken. In dit artikel gaat het om de validiteit van toetsen binnen het onderwijs.

Er zijn tientallen benamingen bekend, die elk een ander aspect van validiteit benadrukken.¹ Deze diversiteit in terminologie gaat echter ten koste van de overzichtelijkheid. In dit artikel wordt de volgende hoofdingdeling gehanteerd:

- inhoudsvaliditeit
- criteriumvaliditeit
- constructvaliditeit.

Deze vormen van validiteit zullen achtereenvolgens worden besproken. Aangezien voor toetsing van medische competentie de inhoudsvaliditeit van groot belang is, zal daar het accent op liggen.

Inhoudsvaliditeit

De vraag naar de *inhoudsvaliditeit* heeft betrekking op de overeenstemming tussen de inhoud van de toets en de 'inhoud' van hetgeen men beoogt meten. Dit is minder een open deur dan het lijkt.² Voor het meten van fysisch-diagnostische vaardigheden kan men nog redelijk gemakkelijk 'toetssituaties' bedenken die de dagelijkse praktijk weergeven; voor het meten van organisatorische vaardigheden is dat al veel moeilijker; voor het meten van een 'professionele attitude' is het een hele opgave. Wat verstaat men precies onder inhoud?

Laten we als voorbeeld een kennistoets nemen. Allereerst gaat het om het vaststellen van het domein van kennis. Over welk domein gaat het? Over dat van een basisarts of van een chirurg of huisarts? Laten we aannemen dat het over het kennisdomein van een huisarts gaat. Wat is dan het kennisdomein van de huisartsgeneeskunde? Deze vraag is niet zondermeer te beantwoorden. De huisartsgeneeskundige kennis is op tal van manieren te beschrijven en in te delen. Men zal moeten zoeken naar een geschikt indelingsprincipe ofwel classificatiesysteem, dat als 'blauwdruk' voor een kennistoets voor huisartsen kan dienen. Onder *blauwdruk* wordt dan verstaan, een 'mal' of 'sjabloon' waarin vastgelegd wordt welke onderwerpen in welke omvang in iedere toets terugkomen. Een vaste structuur, zou men ook kunnen zeggen. Verschillende blauwdrukken zijn denkbaar, bijvoorbeeld naar orgaansysteem, naar leeftijdsklasse, naar beloop van de aandoening of urgentie van het probleem dan wel een combinatie hiervan (een multidimensionale blauwdruk). Welke keuze men ook maakt, het gaat erom, dat de blauwdruk aan een aantal wensen tegemoetkomt. De blauwdruk moet:

- herkenbaar zijn voor degenen die de toets moeten afleggen, in dit geval huisartsen. De onderwerpen moeten hun 'iets zeggen'; ze moeten hun vak erin herkennen.
- aansluiten bij de bronnen met behulp waarvan men studeert. Als men een onvoldoende

scoort op een onderwerp, moet duidelijk zijn, welk boek men ter hand moet nemen of welke patiënten men vaker moet zien.

Door een zogenaamd 'consensusonderzoek' kan men nagaan in hoeverre de keuze van toetsontwerpen en de 'rationale' daarachter worden onderschreven. Hoe groter het draagvlak, des te stelliger mag men zijn in de uitspraak, dat de toets het beoogde domein dekt.

In de blauwdruk moet ook het aantal vragen per onderwerp worden vastgesteld. Moeten er tien vragen over respiratoire aandoeningen worden gesteld of vier? Ook hiervoor moeten weer geldige argumenten worden aangevoerd. Men kan bijvoorbeeld het aantal vragen per categorie laten bepalen door de frequentie van voorkomen van ziekten (in de bevolking en/of als onderwerp in het curriculum) of juist door de ernst van de aandoening: over ernstige ziekten meer vragen.

Ook de in de toets aangeboden problemen en items dienen op hun validiteit te worden getoetst. Als men heeft vastgesteld dat drie vragen over het onderwerp diabetes moeten gaan, kan dat nog een veelheid aan onderwerpen betreffen, bijvoorbeeld de biochemie (de structuurformule van insuline), de epidemiologie (hoe vaak komt insuline-afhankelijke diabetes voor in Nederland) de diagnostiek (klachten en symptomen) en de therapie (medicatie en dieet). De vraag is nu welke onderwerpen thuishoren in een kennistoets voor huisartsen, om bij ons voorbeeld te blijven. Stel, dat we besluiten dat de drie vragen alle over diagnostiek moeten gaan, dan nog kan het item gericht worden op meerdere aspecten, bijvoorbeeld klachten, bevindingen bij lichamelijk onderzoek en de waarde van diagnostische tests. Bordage, een Canadees onderwijskundige, vestigde de aandacht op het belang van de vraagkeuze en introduceerde het begrip 'keyfeature'.³ Dit begrip houdt in, dat de vraag bij voorkeur moet focussen op hetgeen in de realiteit 'the heart of the matter' of het sleutelprobleem vormt. Bij een polsverstuiking is dat bijvoorbeeld: is ook nog een handwortelbeen-

tje gebroken of niet, bij pijn op de borst: is er een hartaandoening in het spel, bij een tennisselleboog: verdient een injectie in de elleboog de voorkeur boven andere behandelingen?

Ook de wijze waarop het op te lossen probleem wordt gepresenteerd heeft een relatie met 'validiteit'.

Men kan een aantal losse vragen stellen, zoals in het volgende voorbeeld.

Insuline-afhankelijke diabetes mellitus:

- a *komt voor bij 3% van de jongeren tussen 20 en 30 jaar. (juist/onjuist)*
- b *manifesteert zich in de meerderheid der gevallen door het optreden van een coma. (juist/onjuist)*
- c *heeft als behandeling van voorkeur het toedienen van synthetische insuline. (juist/onjuist)*

Men kan ook een patiënt opvoeren, zoals die zich aan de huisarts presenteert:

Een huisarts wordt omstreeks 21.30 uur gebeld voor een spoedvisite bij een 4-jarig kind dat plotseling kortademig wakker is geworden. De diagnosen epiglottitis en laryngitis subglottica worden overwogen. Het kind is altijd goed gezond geweest, maar is sinds enkele dagen verkouden. Er is tevens sprake van een blafhoest. Er is geen koorts en er zijn geen slikklachten. Op grond van deze laatste bevindingen is één van de twee diagnosen het meest waarschijnlijk.

- *Dit is epiglottitis (onjuist).*

De casusvorm is herkenbaar voor huisartsen. Zo krijgen ze de patiënten inderdaad op het spreekuur. Men zegt daarom ook wel dat deze vorm bijdraagt aan de 'face validity' ofwel het realiteitsgehalte van de toets, welke weer van invloed is op de acceptabiliteit.

Tot slot is ook de formulering van een vraag van invloed op de inhoudsvaliditeit: een slecht geformuleerde vraag meet niet wat men wil meten.

Uit het voorafgaande moge duidelijk zijn, dat inhoudsvaliditeit meer omvat dan alleen vaststellen van het domein. Op vele niveaus moet, telkens weer opnieuw, gekeken worden

naar inhoud en vorm. Dit impliceert onder andere, dat de validiteit per nieuw geconstrueerde set van toetsvragen kan verschillen. Dus niet: eenmaal inhoudsvalide, altijd inhoudsvalide!

Criteriumvaliditeit

Criteriumvaliditeit heeft betrekking op de overeenkomst tussen de uitkomsten van de meting met de onderzochte toets met die van een 'andere toets'. Deze overeenkomst wordt meestal uitgedrukt in een correlatiemaat. De 'andere toets' is niet zomaar een willekeurig gekozen toets. Van deze toets weet men dat deze 'meet wat men wil meten'. Met andere woorden, van deze toets is de validiteit voldoende bevonden. Zij fungeert zodoende als 'gouden standaard'. Onmiddellijk rijst dan de vraag: als er een valide toets bekend is, waarom is er dan behoefte aan een andere toets? Inderdaad blijkt de nieuw ontwikkelde toets nogal eens overbodig. Een nieuwe toets kan nuttig zijn als bijvoorbeeld de 'gouden-standaardtoets' zeer moeilijk is af te nemen, bijvoorbeeld doordat deze veel tijd, geld of mankracht vergt. Het is denkbaar, dat men dan zoekt naar een vereenvoudigde vorm. Zo kan men zich voorstellen, dat men een toets waarbij fysisch-diagnostische vaardigheden worden geobserveerd, wil vervangen door een schriftelijke 'kennis-over-vaardigheden-toets'. In onderzoek naar de criteriumvaliditeit vergelijkt men deze schriftelijke toets dan met de observatietoets. Wanneer geen 'gouden-standaardtoets' voorhanden is, en dat is doorgaans helaas het geval, vervalt ook de mogelijkheid van het vaststellen van de criteriumvaliditeit.

Er zijn auteurs die onderscheid maken in 'convergente' en 'divergente' (criterium)validiteit. Bovenstaand voorbeeld van de twee soorten vaardigheidstoetsen is een voorbeeld van convergente validiteit: men vergelijkt twee toetsen die worden geacht hetzelfde te meten. Bij divergente validiteit vergelijkt men een toets met een andere toets, welke geacht wordt

iets geheel anders te meten, bijvoorbeeld niet vaardigheden maar kennis. Men verwacht dan een lage correlatie tussen de twee toetsen. Helaas levert dit soort onderzoek zelden iets op: de uitkomst is meestal een matige correlatie, die evenzeer redelijk hoog als redelijk laag genoemd kan worden (het glas is half vol of half leeg). Dit is ook niet verwonderlijk, daar het om dezelfde kandidaten gaat. Vaak is men goed of slecht in zowel het een als het ander (bijvoorbeeld vaardigheden en kennis). Over validiteit komt men dus weinig te weten.

Een andere veel genoemde vorm van criteriumvaliditeit is de *predictieve validiteit*. In vele situaties is men geïnteresseerd in de vraag of toekomstig gedrag te voorspellen is uit de score behaald op een bepaalde toets. Bijvoorbeeld: voorspellen VWO- eindexamencijfers succes tijdens de medische opleiding, of nog sterker: competentie als arts?

Echter, eindexamencijfers meten iets anders dan 'succes tijdens de opleiding' of 'competentie als arts'. Het gaat dus om het vergelijken van verschillende grootheden hetgeen weinig meer te maken heeft met het vaststellen van de validiteit. Als men de validiteit wil vaststellen, moet men bijvoorbeeld de score op een kennistoets aan het begin van de opleiding vergelijken met de score op een (soort)gelijke kennistoets aan het einde van de opleiding. Dat laat onverlet, dat nagegaan kan worden of hoogscoorders op het eindexamen sneller de medische opleiding doorlopen dan laagscoorders. Helaas is voor een dergelijke voorspelling nog geen goede onderwijskundige naam gevonden, een equivalent van de epidemiologische term 'voorspellende waarde'.

Constructvaliditeit

Constructvaliditeit doelt op de overeenkomst tussen de 'theorie' omtrent het begrip of 'construct' dat men denkt te meten en gegevens verkregen uit onderzoek.

Als de opgestelde hypothesen worden bevestigd draagt dit bij aan de constructvaliditeit.

Men kan dit goed illustreren aan de hand van de IQ-test. Men veronderstelt, gezien de aard van het begrip intelligentie, dat gymnasium-leerlingen gemiddeld hoger scoren op deze test dan MAVO-leerlingen. Indien dit niet zo blijkt te zijn, is òf de theorie onjuist òf de test. Meestal is dit laatste het geval. Een andere hypothese is bijvoorbeeld, dat de intelligentie ongeveer gelijk blijft gedurende het leven, dus dat de score op de IQ-test weinig intra-individuele variatie vertonen zal. Ook deze hypothese kan getoetst worden. Hoe meer hypothesen getoetst worden en juist blijken te zijn, hoe meer aanwijzingen er zijn die de constructvaliditeit van de test ondersteunen.

Constructvaliditeit hoort strikt genomen niet in het bovengenoemde rijtje van drie thuis. Het omvat in feite de andere twee soorten validiteit: een construct-valide toets representeert immers de inhoud van wat men wil meten zo nauwkeurig mogelijk en is dus inhoudsvalide. Tevens is een van de hypothesen ten aanzien van een construct-valide toets, dat de scores hoog correleren met die van een 'construct-verwante' toets, hetgeen criteriumvaliditeit veronderstelt.

Bij inhoudsvaliditeit is reeds opgemerkt: eenmaal valide, niet altijd valide. Variërend hierop, kan voor constructvaliditeit worden opgemerkt dat onderzoek hiernaar in feite nooit ophoudt. Telkens nieuwe hypothesen afgeleid uit de theorie, vragen om telkens nieuw hypothesetoetsend onderzoek. Afname van de toets bij een andere doelgroep bijvoorbeeld, kan verrassend andere gegevens opleveren. Een Europese toets in het kader van de studie bestuurskunde bleek bij Oostaziatische studenten tot vreemde uitkomsten te leiden. Dit lag aan het feit, dat de in Europa als juist geldende oplossingen voor bepaalde problemen in Oost-Azië niet als juist golden. Blijkbaar had de toets geen universele constructvaliditeit.

Geldigheid

We komen nog eens terug op de letterlijke vertaling van validiteit: geldigheid. Het voorbeeld van de bestuurskundetoets laat zien, dat het niet om de toetsuitkomst op zich gaat, maar om de wijze waarop deze wordt geïnterpreteerd, met andere woorden, om de geldigheid van de conclusies.⁴ De conclusie ten aanzien van de score van Oostaziatische bestuurskundigen is niet: zij zijn niet goed in hun vak, maar: zij beheersen de westerse aanpak van bestuurskundige problemen niet (en dat hoeft ook niet, want zij hebben hun eigen aanpak).

Zo noopt ook een lage score van ervaren huisartsen op een kennistoets voor huisartsen-in-opleiding tot de vraag: is de uitspraak 'huisartsen hebben een lager kennisniveau dan huisartsen-in-opleiding' geldig? Wellicht is het zo, dat de toets een soort kennis meet die in het latere beroepsleven steeds minder belangrijk wordt. De resultaten van onderzoek als leidraad gebruiken om zowel de theorie als de wijze van toetsen bij te stellen vormt de kern van valideringsonderzoek.

Validiteit en betrouwbaarheid

De begrippen betrouwbaarheid en validiteit staan niet los van elkaar. Het valt gemakkelijk te begrijpen, dat uitspraken gedaan op grond van toetsuitkomsten die niet betrouwbaar zijn, in hun geldigheid worden aangetast. De uitspraak, "het doet er minder toe of de toets betrouwbaar is, als zij maar valide is", is dus een contradictio in terminis.

Conclusie

De validiteit van een test kan onderzocht worden aan de hand van de volgende vraag: zijn de uitspraken die ik wil doen op grond van deze toets, afgenomen bij deze specifieke groep van mensen, geldig. Om op deze vraag antwoord te kunnen geven is het nodig om:

- de inhoud van de toets zeer zorgvuldig samen te stellen en liefst voor te leggen aan een panel van experts en representanten van de doelgroep (bijvoorbeeld aan een vaste toetsbeoordelingscommissie).
- vooraf exact te formuleren welke betekenissen aan de toetsuitkomsten toekent op grond van welke theoretische veronderstellingen.
- na te gaan - door middel van onderzoek - of deze veronderstellingen juist zijn.⁵

Literatuur

1. Berkel HJM. De diagnose van toetsvragen. Academisch proefschrift. Universiteit van Amsterdam, 1984.
2. Ebel R. The practical validation of tests of ability. *Educational Measurement: Issues and Practice* 1983; 2: 7-10.
3. Bordage G, Page G. An alternative approach to PMP's: the 'key features' concept. In: Hart IR, Harden RM, eds. *Further developments in assessing clinical competence*. Montreal: Can-Heal Publications Inc., 1987: 59-75.
4. Kane MI. The validity of licensure examinations. *American Psychologist* 1982; 37: 911-8.
5. Carmines EG, Zeller RA. *Reliability and validity assessment*. Beverly Hills/London: SAGE Publications, 1985.

Opdrachten

1. Iemand ontwerpt een toets voor het meten van het bewaren van het evenwicht. Koorddansers scoren op deze toets echter lager dan een steekproef uit de bevolking. Hoe formuleert u uw conclusie ten aanzien van validiteit?
2. Onderzoek heeft aangetoond, dat het eindexamencijfer behaald op het vak Nederlands positief correleert met de mate van succes tijdens de medische opleiding. Wat zegt deze onderzoeksuitkomst over de predictieve validiteit van het examen Nederlands?

Aanbevolen literatuur

Drenth PJD, Sijtsma K. *Testtheorie. Inleiding in de theorie van de psychologische test en zijn toepassingen*. Houten: Bohn, 1990.