Construct- en criteriumvaliditeit

In het maartnumer van *Huisarts en Wetenschap* wordt uitvoerig aandacht besteed aan de kwaliteit van het instrumentarium voor het meten van de gezondheidstoestand. *Zaat & Schellevis* stellen daarbij terecht vast dat er nogal wat onduidelijkheid en spraakverwarring bestaat over de definitie van begrippen die verschillende aspecten van de validiteit weergeven. Zij – en ook *König-Zahn & Furer* – kiezen daarbij echter voor een definitie die naar onze mening mede aanleiding geeft tot de spraakverwarring.[1,2] Verwarrend is de keuze om de onderlinge samenhang tussen vragenlijsten die hetzelfde beogen te meten, zonder meer op te vatten als een vorm van constructvaliditeit. Dit lijkt immers niet principieel te verschillen van criteriumvaliditeit: zowel de te valideren test als de gouden standaard beogen uiteraard hetzelfde te meten. Wanneer twee instrumenten hetzelfde beogen te meten, zijn er grofweg twee mogelijkheden:

- Een van de instrumenten meet (per definitie) het te meten verschijnsel op adequate wijze (gouden standaard); het nieuwe instrument wordt geijkt met behulp van deze gouden standaard (het goudgehalte van deze standaard staat niet ter discussie).
- Een 'harde' gouden standaard ontbreekt. Er moet in dat geval noodgedwongen genoegen genomen worden met een 'vergulde' standaard. Dat betreft dan een meetinstrument waarvan de waarde, bijvoorbeeld met behulp van constructvalidering, reeds is aangetoond en dat door de wetenschappelijke gemeenschap als zodanig is geaccepteerd op grond van veelvuldige toepassing in onderzoek. Op grond van deze status geldt een dergelijk instrument als een extern criterium waaraan het te valideren instrument wordt getoetst.

Tussen een gouden en een vergulde standaard bestaat uiteraard slechts een gradueel verschil: het gaat om een verschil in de mate waarin het echt goud is wat er blinkt. Bij de besproken mogelijkheden gaat het om vergelijking van twee instrumenten die hetzelfde beogen te meten. Als er onvoldoende reden is om aan te nemen dat de standaard de werkelijkheid beter benaderd dan het te valideren instrument, kan beter afgezien worden van (criterium)validering: het valt dan immers nooit hard te maken welk van de twee instrumenten het beste is.

Bij voorkeur moet het begrip constructvaliditeit gereserveerd worden voor de vergelijking van twee instrumenten die juist niet pretenderen om hetzelfde te meten. Het door *Zaat & Schellevis* gegeven voorbeeld van de vergelijking tussen de geriatrische depressieschaal en de Zung-depressieschaal is derhalve geen goed voorbeeld van (convergente) constructvaliditeit. De eventuele samenhang tussen deze twee operationalisaties van depressie kan slechts adequaat een licht op de (criterium)validiteit werpen wanneer de Zung-depressieschaal een acceptabele vergulde standaard zou zijn, bijvoorbeeld door een goede samenhang met uitvoerige klinische diagnostiek conform de DSM-III-criteria.

Constructvaliditeit betreft de validering van een instrument aan de hand van de toetsing van een theoretische verwachting. Bij ontstentenis van een acceptabele gouden standaard, zo is de redenering, wordt nagegaan of een theoretisch te verwachten samenhang inderdaad ook bestaat. De uitkomsten van metingen met behulp van het te valideren instrument en die van een instrument dat een ander concept beoogt te meten, en waarvoor de criteriumvaliditeit idealiter vaststaat, worden gecorreleerd. Overigens hoeft het niet altijd om een ander instrument te gaan, maar kan het ook om variabelen als leeftijd, geslacht of de ernst van een ziekte gaan. Essentieel is dat het om twee inhoudelijk verschillende variabelen gaat.

Deze procedure wordt uiteraard alleen uitgevoerd als er geen acceptabel instrument is waaraan de criteriumvaliditeit kan worden getoetst. In die zin is het inderdaad een verlegenheidsvalidering. De gevonden samenhangen zijn meestal niet maximaal. Anders dan bij criteriumvaliditeit hoeft er bij constructvaliditeit dus lang niet altijd sprake te zijn van uitzonderlijk hoge correlaties. Sterker nog, op theoretische gronden is dat bijna nooit te verwachten. Het gaat bij constructvaliditeit immers primair om de overeenkomst tussen de theoretisch veronderstelde en de feitelijk gevonden richting en sterkte van het verband. Komt de verwachting consistent en duidelijk uit, zo luidt de redenering, dan zal de meting wel valide zijn geweest. De theoretische samenhang kan zowel positief als negatief zijn. Er is dan sprake van convergente validiteit. Wanneer de theoretische samenhang zwak is of ontbreekt, spreekt men van divergente validiteit.

Resumerend: criteriumvaliditeit betreft de vergelijking tussen twee meetinstrumenten die hetzelfde beogen te meten en waarvan er één het predikaat gouden standaard verdient. Dit maakt het mogelijk om de validiteit van een bepaald instrument in absolute termen te kwantificeren. Constructvaliditeit behelst een vergelijking tussen twee instrumenten die elk iets anders pogen te meten. Dit geeft slechts in relatieve zin zicht op de validiteit, hetgeen met name nuttig is wanneer er een keuze tussen verschillende instrumenten moet worden gemaakt. De voorkeur dient dan uit te gaan naar het instrument met de hoogste constructvaliditeit.

Het bepalen van de constructvaliditeit van een instrument is dus op zichzelf een heldere procedure die echter minder 'hard' is dan de validering met behulp van een gouden standaard.

Helderheid over de gebruikte definities voorkomt associaties met de Baron van Münchausen. Het antwoord op de vraagstelling in de titel van de bijdrage van *Zaat & Schellevis* dient naar onze mening dan ook met nee te worden beantwoord.

J.Th.M. van Eijk
L.M. Bouter

1 Zaat JOM, Schellevis F. Aan je eigen haren omhoog? Over betrouwbaarheid en validiteit van instrumenten voor het meten van 'kwaliteit van leven'. Huisarts Wet 1995; 38(3): 105-9.
2 König-Zahn C, Furer JW. De keuze van een vragenlijst. Methodologische en praktische overwegingen. Huisarts Wet 1995; 38(3): 110-6.

Naschrift

Valideren van 'kwaliteit van leven'-schalen heeft blijkbaar niet alleen verband met de haren van de Baron van Münchausen, maar ook met de verwarring die deze figuur zelfs binnen één lokaal weet te stichten. De discussies over wat validiteit nu precies is, hebben een hoog rederijkersgehalte. Het is de vraag, of dit onderwerp voor het begrip van de huisarts die iets over kwaliteit van leven wil weten, nog wel te volgen, laat staan interessant is.

De verschillen tussen onze opvatting en die van *Van Eijk & Bouter* zitten in de beschrijving van het begrip criteriumvaliditeit. *Van Eijk & Bouter* zien het valideren van meetinstrumenten als een heldere procedure: bij het bepalen van de criteriumvaliditeit vergelijk je twee vragenlijsten/meetinstrumenten, waarbij er een de gouden of de vergulde standaard is. Terecht stellen *König-Zahn & Furer* dat er op het terrein van ervaren gezondheid geen foutloze of bijna foutloze meetinstrumenten bestaan,[1] zodat het ons inziens onduidelijk blijft welke van de twee lijsten dan voor goud of verguld moet doorgaan. Ook veel gebruikte en in de wetenschappelijke gemeenschap gebruikte meetinstrumenten kunnen immers onzin meten.[2] De wetenschapsgeschiedenis is vol van anekdotes

van zinloze diagnostische instrumenten, zodat wij op voorhand niet denken dat wij nu voor het eerst in de geschiedenis de 'werkelijkheid' echt kunnen meten. Bescheidenheid past de onderzoeker.

Ons voorbeeld over de vergelijking tussen de Zung- en de geriatrische depressieschaal hebben we letterlijk overgenomen uit het artikel waarnaar wordt verwezen.[3] De auteurs, referenten en editors van *Family Practice*, beschouwen het vergelijken van twee vragenlijsten dus blijkbaar ook meer als het bepalen van construct- dan van criteriumvaliditeit. Het vermijden van de term 'criteriumvaliditeit' bij 'kwaliteit van leven'-onderzoek is dus een heldere vereenvoudiging: op deze manier wekken we geen enkele associatie met gouden of vergulde standaarden. Niemand wordt daar slechter van, integendeel.

Joost Zaat
François Schellevis

1 König-Zahn C. Furer JW. De keuze van een vragenlijst. Methodologische en praktische overwegingen. Huisarts Wet 1995; 38(3): 110-6.
2 Patrick DL, Erickson P. Health status and health policy; allocating resources to health care. Oxford: Oxford Univerisity Press, 1993: 181-212.
3 Marwijk H, Arnold I, Bonnema J, Kaptein A. Self-report depression scales for elderly patients in primary care; a preliminary study. Fam Practice 1993; 10: 63-5.

NOTA BENE

In veel overredingsexperimenten wordt onvoldoende onderkend dat de overtuigingskracht van argumenten afneemt bij toenemende extremiteit van het aanbevolen standpunt.

Stelling bij: Bakker AB. Denk na, vrij veilig: Descriptief en experimenteel onderzoek naar attitudes tegenover condoomgebruik [Dissertatie]. Groningen: Rijksuniversiteit Groningen, 1995.