TOETSINGSSPECIALISTEN IN ACTIE

J. Moll, Emeritus Hoogleraar Anatomie en Algemene Gezondheidszorg, Rotterdam/Rhenen.

Bijna een jaar geleden spraken op een vrije voordrachtendag van de NVMO Cees van der Vleuten en Scheltus van Luyk over "Observatietoetsen, mogelijkheden en valkuilen". Het publiek kreeg daarbij in schriftelijke vorm een samenvatting en conclusies, samen één pagina, plus zeven pagina's aan de literatuur ontleende, grotendeels kwantitatieve gegevens over het onderwerp van de voordracht, waarnemingstoetsen: toetsen waarin de student niet moet praten, schrijven of kruisjes zetten bij (goede) antwoorden, maar moet handelen en waarbij dan dit handelen door waarnemers wordt beoordeeld.
Onlangs kwam het ervan te doen wat ik mij had voorgenomen: dit materiaal te herlezen. Ik had daarbij een steuntje in de rug: ik beschikte ook over de tekst van een andere voordracht van dezelfde auteurs, waarin in sterke mate hetzelfde onderwerp aan de orde is: "Decomposition of OSCE's; some methodological considerations and empirical findings" (OSCE: Objective Structured Clinical Examination, red.). Deze voordracht werd gehouden tijdens de "Second Ottawa Conference on Assessing Clinical Competence", zomer 1987. Ging het in de in de aanhef van dit artikel genoemde voordracht over waarnemingstoetsen in het algemeen, in de zojuist genoemde voordracht gaat het vooral over één variant van de waarnemingstoetsen, de Maastrichtse vaardigheidstoets, hoewel dit niet onmiddellijk duidelijk is uit de titel van de voordracht. In beide studies ligt het accent op betrouwbaarheid van de toetsing. In het navolgende enige bespiegelingen over deze twee studies. Ik spreek van "bespiegelingen", omdat ik mijzelf een vraag stel, maar niet tot een antwoord kom.

Om de lezer enig houvast in feitelijkheden te bieden, volgt hier eerst een korte beschrijving van de waarnemingstoets waarover in deze beide studies het meest uitvoerig wordt gesproken, de vaardigheidstoets die in Maastricht -waar de auteurs werken- sinds 1982 wordt afgenomen en die zij uitvoerig hebben geanalyseerd.

De Maastrichtse vaardigheidstoets

Deze toets wordt eens per jaar afgenomen bij studenten van alle jaarklassen, al verschuift het zwaartepunt met de voortgang van de studiejaren van medisch-technische naar klinische vaardigheden, zoals bijvoorbeeld het begeleiden van een bevalling. De toets bestaat uit een tiental "stations", verdeeld over de diverse medische vakgebieden met daarbij ook therapeutische, sociale en laboratoriumvaardigheden. De tijdsduur van deze toets voor de student is meestal twee uur en gemiddeld 15 minuten per station. De toets is een waarnemingstoets in de al omschreven zin. De student moet vaardigheden tonen, handelen dus, en waarnemers (één of twee) kijken of de student juist handelt. Zij doen dit aan de hand van een lijst waarin tenminste tien onderdelen (items) van de vaardigheid die in een station wordt getoetst, zijn vermeld. Zij geven voor elk item aan of juist of onjuist is gehandeld; soms bestaat de mogelijkheid van "onvolledig". Aan de hand van de zo verkregen gegevens wordt voor elke student een score berekend en uit deze score wordt dan voor de student zijn toetsuitslag afgeleid.
De auteurs richten zich in deze beide studies overwegend op de betrouwbaarheid van dit soort toetsen. Dat wil zeggen, op de vraag: "Zou één en dezelfde stu-

dent dezelfde score hebben behaald - wanneer voor de beoordeling van bepaalde vaardigheden andere stations waren gekozen,
- wanneer voor de beoordeling van de vaardigheid die in één station wordt getoetst, andere items waren gekozen, of
- wanneer voor de beoordelingen andere beoordelaars waren gekozen"?
Van der Vleuten en Van Luyk vergelijken hun analyse van de Maastrichtse vaardigheidstoets met de resultaten van andere studies over waarnemingstoetsen. Voor de Maastrichtse toets beschikken zij over een materiaal van indrukwekkende omvang. Voor een periode van vier jaar verwerkten zij de gegevens van de jaarlijkse toets, afgenomen bij alle zes jaarklassen, gemiddeld 110 studenten omvattende.
Deze studies leiden mij tot een vraag die voortkomt uit interesse voor toetsing gezien niet als zelfstandig probleemveld, maar als element -uiterst belangrijk element- binnen het totaal van het medisch onderwijs. Mijn vraag luidt: "Hoeveel leveren studies als deze op voor de toetspraktijk?". En ik kom tot de conclusie, dat ik zowel "veel" als "weinig" goed verdedigbare antwoorden vind. Wat pleit voor "veel"?

Veel

Van der Vleuten en Van Luyk concluderen uit hun eigen gegevens en die in de literatuur, dat in op vaardigheden gerichte waarnemingstoetsen een hoge graad van betrouwbaarheid bereikt kan worden. Dat hoort bij "veel"! Vaardigheden, het verwerven van vaardigheden, zijn een belangrijke doelstelling, misschien de belangrijkste doelstelling van het medisch onderwijs, zeker wanneer men het begrip "vaardigheden" ruim opvat, reikende van het stellen van een diagnose bij een verwarrende veelheid van klachten en symptomen tot het correct uitvoeren van een bloeddrukmeting en het in een gesprek tot rust brengen van een angstige en verwarde psychiatrische patiënt. En betrouwbaarheid van toetsing is evenzeer van groot belang. Betrouwbaarheid van de toetsing betekent, dat de student een juist beeld van zijn vorderingen krijgt. Dat is op zich zelf van betekenis, maar ook omdat de student zo zekerheid wordt geboden omtrent de vraag waar zijn sterke en zwakke punten zijn gelegen. Ook geeft een betrouwbare toets opheldering omtrent sterke en zwakke punten in het onderwijs. "Kan" kreeg een onderstreping bij de conclusie over de bereikbaarheid van een hoge graad van betrouwbaarheid bij toetsen van de soort die hier aan de orde is. Van der Vleuten en Van Luyk vonden bij vergelijking van de gegevens over uiteenlopende waarnemingstoetsen, dat de betrouwbaarheid nogal kan wisselen.
De Maastrichtse vaardigheidstoets scoort hoog voor betrouwbaarheid. Van der Vleuten en Van Luyk menen, dat deze hoge betrouwbaarheid -die bovendien wordt bereikt bij een vrij bescheiden toetsduur van twee uur- mogelijk berust op de wijze waarop de prestatie van de student wordt gescoord. Per station, waarvoor in verhouding tot soortgelijke toetsen de tijd vrij ruim is bemeten, wordt een student beoordeeld aan de hand van een checklist met relatief veel en zorgvuldig vastgestelde items.
Naast de conclusie omtrent toetsbetrouwbaarheid in het algemeen leverden deze studies ook een aantal andere interessante en belangrijke conclusies op. Hier zijn er enkele.
Aangetoond kon worden, dat bij het hanteren van de bij dergelijke toetsen gebruikelijke beoordelingsmethoden de scores slechts in geringe mate afhankelijk zijn van de beoordelaar, zodat de tijd en extra inspanning verbonden aan het werken met twee beoordelaars beter op andere wijze kunnen worden besteed. Natuurlijk is deze conclusie alleen geldig binnen de gekozen beoordelingsmethode. Dat bij de Maastrichtse vaardigheidstoets de

scores van twee beoordelaars weinig uiteenlopen, is niet verrassend; zij mogen immers alleen ja of neen zeggen aangaande scherp omschreven (onderdelen van) gedrag van een student. Bepaald wat anders dan het lezen van een als examenopgave door een student geschreven tekst en daar dan na enig nadenken een cijfer voor geven. Het is zorgvuldig aangetoond, dat beoordelingen dan hemelsbreed kunnen verschillen.
Van der Vleuten en Van Luyk weten waarschijnlijk te maken, dat de winst aan betrouwbaarheid die wordt verkregen door de tijdsduur van de toets te vergroten, groter is bij uitbreiding van de tijd per station, en dus bij meer beoordelingsitems per station, dan bij uitbreiding van het aantal stations.
Interessant en van direct praktisch belang lijkt mij ook, dat Van der Vleuten en Van Luyk konden aantonen, dat de betrouwbaarheid waarmee individuele beoordelaars vaardigheden van studenten beoordelen, niet toeneemt naarmate de examinator meer ervaring krijgt.
Bij "veel" voor de winst voor de toetspraktijk, afkomstig uit studies als die van Van der Vleuten en Van Luyk, kan ook nog worden genoemd dat dergelijk werk, waarin zo zorgvuldig naar het toetsingsproces wordt gekeken, kan bijdragen tot enige heilzame afbraak van het onverantwoorde zelfvertrouwen van examinatoren die menen, dat zij betrouwbare resultaten kunnen bereiken met hun nauwelijks aan regels gebonden beoordelingen van mondelinge of schriftelijke examens en tentamens, of van hun haastig verzonnen "multiple choice" toetsen. Gevreesd moet echter worden, dat zij van studies als die welke hier aan de orde zijn, geen kennis nemen. En zou niet het jargon van de toetsspecialisten daartoe bijdragen? Het hiermee aangesneden onderwerp is mijns inziens van groot belang. Het hoofdprobleem van de toetsing in het medisch onderwijs is immers niet, dat er geen goede toetsen zijn, maar dat ze niet worden gebruikt.

Weinig

Wat pleit er voor om "weinig" te kiezen als antwoord op de vraag: "Hoeveel leveren studies als deze op voor de praktijk van het toetsen?".
We krijgen geen antwoord op de vraag voor welke doelstellingen dit soort observatietoetsen de aangewezen keuze zijn. Een nog bredere vraag die geen werkelijk antwoord krijgt, is die naar de validiteit van dit soort toetsen. Meten zij wat men wil meten? Of, specifieker geformuleerd, meten deze toetsen hoever de student is gevorderd in zijn vaardigheid om als arts te functioneren? Validiteit komt in "Observatietoetsen, mogelijkheden en valkuilen" wel aan de orde, maar de grondigheid waarmee dit onderwerp wordt besproken, staat ver, ver achter bij die waarmee de betrouwbaarheidsvraag wordt besproken.
In beide voordrachten ging het vooral over betrouwbaarheid. Natuurlijk van groot belang, maar is het juist betrouwbaarheid geheel in isolement te behandelen? De betekenis voor het onderwijs van de betrouwbaarheid van een toetsvorm wordt toch beïnvloed door het totaal van het toetsingsprogramma en ook door het totaal van het onderwijsprogramma? Zou dit aspect niet tenminste één of twee zinnen waard zijn geweest?
Ook aan belangrijke praktische vragen wordt voorbij gegaan. Hoeveel tijd is gemoeid met het voorbereiden en uitvoeren van dergelijke toetsen? Onbeantwoord en ongenoemd blijft ook de voor de praktijk heel belangrijke vraag: dienen de beoordelaars, zoals in Maastricht, artsen te zijn en zelfs artsen met een specialisatie overeenkomende met het terrein van de toetsing, bijvoorbeeld oogartsen voor de beoordeling van oogheelkundige vaardigheden?

Astronomen en de geologie

Maar is dit pleidooi voor "weinig" eigenlijk wel billijk? Is het niet zo iets als tegen een astronoom zeggen dat zijn werk weinig bijdraagt aan onze kennis van de geologie van Nederland? Of -zonder beeldspraak- ben ik niet bezig om de auteurs van deze studies te verwijten, dat zij zich andere vragen hebben gesteld dan ik in hun plaats zou hebben gesteld? Ongetwijfeld een dom en onzinnig verwijt! Zo eenvoudig lijkt het me toch niet. Het onderzoek betreffende toetsing heeft natuurlijk een autonome wetenschappelijke waarde, maar het is toch ook zo verbonden met de praktijk van het onderwijs dat men een directe oriëntatie op die onderwijspraktijk mag verwachten. Uit de aard der zaak zal de onderzoeker die zich op toetsing richt, slechts een heel beperkt deel van het totale veld van toetsproblemen deugdelijk kunnen onderzoeken. Maar hij zou als specialist de niet-specialist die graag zijn hulp ontvangt, veel hulp kunnen bieden door zijn onderzoeksresultaten steeds te presenteren tegen een achtergrond -het mag een uiterst summiere schets zijn- van wat er bekend en onbekend is op terreinen van de toetsing, grenzende aan het noodzakelijk beperkte gebied dat hij heeft onderzocht.

Bladvulling

Wanneer men de som van de frekwenties van rectumcarcinoom en prostaatcarcinoom bij de man vergelijkt met de frekwentie van mammacarcinoom of cervixcarcinoom bij de vrouw, is het verwonderlijk dat er in het afgelopen decennium geen voorstellen zijn gedaan om alle mannen boven 50 jaar op te roepen om zich naar een in hun buurt gereedstaande bus te begeven en zich daar, tegen geringe vergoeding, door speciaal daarvoor opgeleide paramedische krachten rectaal te laten toucheren.

Stelling Proefschrift M.J.J.T. Bogman, Nijmegen, 16 december 1983.