VAN ACTIE NAAR PRAKTIJK: reactie op Moll's "Toetsspecialisten in actie"

C.P.M. van der Vleuten, psychometricus, medewerker Projektgroep Evaluatie Studieresultaten (PES), Rijksuniversiteit Limburg, Maastricht,
S. van Luyk, arts, medewerker Projektgroep Evaluatie Studieresultaten (PES), Rijksuniversiteit Limburg, Maastricht.

Als eersten zullen wij met Moll beamen, dat de studie welke hij onder de loep heeft genomen, weinig bijdraagt aan de toetspraktijk van alledag. Integendeel, de besproken publicatie is betrekkelijk specialistisch, bedoeld en geschreven voor een redelijk psychometrisch geschoold publiek. Het doet ons deugd, dat Moll niettemin heel helder en scherp de inhoud heeft ontrafeld en er praktische conclusies aan heeft verbonden.
De besproken studie is slechts een radertje uit het gehele empirische uurwerk, dat langzamerhand enige Zwitserse precisie begint te krijgen. Naar onze mening heeft zich in de afgelopen jaren een belangrijke ontwikkeling voorgedaan: er is een groeiende, warme belangstelling voor de toetsproblematiek in de geneeskunde ontstaan op het gebied van valide c.q. levensechte metingen van student-prestaties. De start daarvan is al wat ouder. Met de toenemende aantallen studenten en de komst van de computer heeft de efficiënte multiple choice vraag in de zestiger jaren flink om zich heen gegrepen. Met deze opkomst ging echter ook een groeiende ontevredenheid gepaard: er moest toch méér getoetst worden dan kale herkenning van feiten alleen. Het gevoel van onbehagen heeft tot een bloeiende periode van de "papieren patiënt" geleid (b.v. het Patient Management Problem, Portable Patient Management Problem (P4-deck) en, meer recent, de computersimulaties). Ongeveer 10 jaar geleden werd de Objective Structured Clinical Examination (OSCE) geïntroduceerd. Deze levensechtere toetsvorm is, zo moge blijken uit de Ottawa conferenties, wereldwijd aangeslagen. Nu slaat de term OSCE slechts op een *organisatieprocedure,* minder op een inhoudelijke methode, maar kenmerkend voor veel OSCE's is, dat gebruik wordt gemaakt van nagebootste levensechte situaties betrekking hebbend op alles wat in een arts-patiënt contact kan voorkomen. De Maastrichtse vaardigheidstoets is daar één exponent van. De grondgedachte van dit soort toetsen is simpel en zeer valide: examineer studenten op datgene waarmee ze in hun onderwijs (met name in de laatste jaren) maar vooral ook in de praktijk, te maken krijgen en maak dat zo (levens)echt als maar mogelijk is. Wellicht dat deze grondgedachte velen heeft aangesproken en er (mede) voor verantwoordelijk is dat de bovengenoemde groeiende belangstelling voor evaluatie is ontstaan.
Vanaf 1986 zijn de eerste wat grotere empirische studies verschenen over deze nieuwe toetsvorm. Daarvan is het door Moll besproken verhaal er één. Vòòr de Maastrichtse toets zijn er meer verschenen, ook over de onderwerpen waarin Moll meer geïnteresseerd is. Het aantal publicaties neemt de laatste jaren sterk toe, waarbij opvallend is hoe duidelijk de resultaten van deze studies met elkaar overeenkomen. Natuurlijk is er nog een heleboel werk te verrichten, maar de volgende conclusies zijn uit de bevindingen te trekken. Zij kunnen rechtstreekse praktische consequenties hebben:
- de toetsen worden sterk gewaardeerd door zowel studenten als docenten;
- beoordelaars vormen geen grote bron van variantie (overigens geldt dit niet alleen voor dit "geharnaste" formaat, zoals Moll opmerkt, maar ook voor andere instrumenten waarin de beoordeling wat globaler is, c.q. er zijn an-

dere variantiebronnen die veel groter en doorslaggevender zijn);
- simulatiepatiënten vormen eveneens geen grote bron van foutenvariantie;
- goed getrainde leken (b.v. simulatiepatiënten) zijn prima in staat om, afhankelijk van de gebruikte beoordeling, studenten te beoordelen;
- het gebruik van rating scales in plaats van checklists kan voor sommige vaardigheden goed worden aangewend en leidt, mits op goede wijze gemaakt en ingezet in een toets, tot niet veel minder betrouwbare resultaten;
- het blijkt, dat de competentie van een student op de ene vaardigheid (casus) niet veel voorspellende waarde heeft voor de competentie op een andere casus: competentie is nogal inhoudsafhankelijk. Om tot representatieve, stabiele of betrouwbare uitspraken te komen, betekent dit, dat relatief veel vaardigheden (casus) moeten worden getoetst en dit heeft consequenties voor de minimaal benodigde toetstijd. De meeste studies tonen aan, dat een toetslengte van minimaal drie à vier uren nodig is;
- Moll concludeert terecht uit ons verhaal, dat het vergroten van een station, en dus meer beoordelingsitems, meer betrouwbare scores oplevert dan het toetsen van meer stations. Op grond van andere studies is hierin echter een verandering van inzicht ontstaan. Het blijkt, dat lengte en aantal stations tot op zekere, nog onbepaalde hoogte uitwisselbaar zijn. Het meest doorslaggevend is zonder meer de totale toetstijd, waaraan studenten worden onderworpen;
- de studies naar de validiteit van deze toetsvorm laten voorlopig positieve resultaten zien, alhoewel deze tot op heden moeilijk interpreteerbaar zijn gebleken. Het is duidelijk, dat op het gebied van de validiteit nog veel werk moet gebeuren. Docenten en studenten vinden in ieder geval wel, dat met deze toetsen belangrijke zaken worden geëxamineerd;
- heel voorlopige gegevens wijzen op een positieve invloed van deze toetsvorm op het leergedrag van studenten. Ook op dit terrein moet nog veel onderzoek worden gedaan;
- toetsen vereisen standaardisatie, hetgeen aanleiding geeft tot standaardisatie van c.q. discussie over doelstellingen van onderwijs;
- de kosten gemoeid met observatie-toetsen, hoeven geen belemmering te vormen voor introductie (uiteraard geheel afhankelijk van de couleur locale).

De conclusie tot op heden is dan ook, dat dit soort van toetsen een belangrijke aanvulling kan zijn op het huidige arsenaal aan examenvormen. Bij de introductie ervan moet men echter rekening houden met een niet al te hoge betrouwbaarheid, vooral wanneer de toetsduur kort is. Niet onaannemelijk is overigens, dat deze niet al te hoge betrouwbaarheid in elk geval al gauw hoger is dan die van de bestaande klassieke toetspraktijk.

Tot zover de inhoudelijke kant van de zaak. Wanneer we Moll's betoog verder goed begrepen hebben, komt hij eigenlijk tot de conclusie, dat nieuwe toetsinstrumenten, inclusief bijbehorende wetenschappelijke bevindingen, van betrekkelijk geringe waarde zijn als zij niet ook worden vertaald naar de toetspraktijk. Hier heeft de toets-specialist een belangrijke taak in, omdat hij de niet-specialist te hulp zou moeten schieten, die, aldus Moll, "graag zijn hulp ontvangt".
Helaas ligt het niet zo eenvoudig. Kijkend naar de bestaande toetspraktijk, valt er inderdaad nog heel wat te verbeteren. Maar voordat een feitelijke verandering plaats zou kunnen vinden, zal naar onze mening eerst ook iets moeten veranderen in de opvattingen over de toetspraktijk. En dat is geen eenvoudige zaak. Toetsing en examinering worden door veel docenten opgevat als een autonome activiteit, waarvoor men alleen zelf verantwoordelijk is en wat behoort tot of geassocieerd wordt met de eigen professionele competentie.

Bemoeienis met de toetsing door derden wordt dan ook vaak opgevat als een aantasting van de eigen deskundigheid. Bovendien worden toetsingsactiviteiten nogal eens onderschat. Goede toetsen kosten gewoon veel werk (net als goed onderwijs trouwens). Zo lang in deze opvattingen of attitude geen verandering komt, zal elke poging tot verandering door welke toetsspecialist dan ook weinig vruchtbaar zijn. Kortom, de wil moet aanwezig zijn, dan ontstaat een weg.
Toetsspecialisten hebben natuurlijk een belangrijke rol bij een poging tot verandering van deze opvattingen. Maar belangrijker nog dan de toetsdeskundigen, zijn de docenten c.q. medici zelf. Vooral zij kunnen in staat worden geacht om collega's (!) te overtuigen van het belang van een goed toetsprogramma. Verwijzing naar de "specialist-toetsdeskundige" voor de verbetering van de bestaande toetspraktijk is te passief, te makkelijk. Eerder, maar vooral ook veel beter, moet verbetering van binnenuit worden geëntameerd door collega-docenten. De rol van de toetsdeskundigen is slechts ter ondersteuning hiervan.
In Maastricht is het gebleken, dat het ook anders kan. Daar wordt het toetsprogramma door medici vormgegeven en uitgevoerd, ondersteund door en in samenwerking met toets-deskundigen. Er is een toetsings-systeem ingevoerd, waarin systematisch een hele wezenlijke stap is ingebouwd. Toetsmateriaal dat wordt gemaakt door docenten uit individuele vakgroepen, wordt bekeken door *collega's*. Concreet wordt dit gedaan door speciaal daarvoor in het leven geroepen commissie's, maar dat is niet essentieel. Wezenlijk is, dat het door de docent gemaakte materiaal nog eens kritisch wordt bekeken en de vrijheid bestaat daarop commentaar te hebben. Het is onze ervaring, dat deze stap een zeer grote invloed heeft op de kwaliteit van de toetsen, maar bovendien in het geheel niet bedreigend werkt. Integendeel, de docenten waarderen de kritiek en *leren* ervan: zij worden de deskundigen.
Door de aanpak in Maastricht is het bovendien mogelijk om centraal toetsmateriaal op te slaan en er veelvuldig gebruik van te maken. Niet alleen is dat een economische strategie, het verhoogt opnieuw weer de kwaliteit van het materiaal. Met iedere afname wordt het toetsmateriaal bijgesteld en verbeterd.
In zijn algemeenheid heeft de Maastrichtse strategie ertoe geleid, dat er een situatie is ontstaan waarin docenten (en de facultaire geledingen zoals onderwijscommissie en examencommissie) bereid zijn na te denken over het toetssysteem, en daaraan ook consequenties durven verbinden (dat is overigens niet beperkt tot het toetssysteem). Anders was het bijvoorbeeld ook niet mogelijk om tamelijk logistiek belastende observatietoetsen, zoals de vaardigheidstoetsen, te introduceren. In Maastricht ontstaat tot ons groot genoegen langzaam de situatie waarin de medici zelf de toetsspecialisten worden en steeds meer actie willen.