

# Toetsing van medische competentie

## Historie en implicaties voor huidige examenopzet

L.W.T. Schuwirth, C.P.M. van der Vleuten en A.J.J.A. Scherpbier

Een overzicht van de resultaten met hun implicaties voor de meest gebruikte examenvormen om medische competentie te toetsen.

**H**ET GROTE BELANG van een verantwoorde toetsing van medische competentie bij geneeskundestudenten wordt algemeen onderschreven. De drie meest betrokken partijen - de maatschappij, de faculteiten en de studenten - hebben er veel belang bij dat door middel van examens nauwkeurig onderscheid wordt gemaakt tussen competente en niet-competente studenten. Voor alle drie de partijen geldt dat zij nadeel ondervinden van zowel onterechte slaagbeslissingen als onterechte zakbeslissingen.

De maatschappij verwacht dat de artsen die in de gezondheidszorg werken competent zijn, zodat niet alleen zoveel mogelijk het menselijk lijden wordt vermindert maar ook de economische verliezen door de aan ziekte verbonden kosten laag blijven. Anderzijds heeft de maatschappij er belang bij dat studenten snel en efficiënt worden opgeleid. De faculteiten moeten dus hun rendementcijfers zo hoog mogelijk houden bij een gegarandeerde kwaliteit van hun afgestudeerde artsen. Ze moeten ook aantonen dat hun afgestudeerden in staat zijn qua competentie te concurreren met elders afgestudeerde artsen. De vervaging van de binnengrenzen in Europa zal deze druk alleen nog maar opvoeren.

Voor de student heeft een onterechte zakbeslissing grote financiële, motivationele en sociale consequenties. Onterecht slagen kan echter ook leiden tot onoverkomelijke problemen later in de studie, zodat de student alsnog met zijn studie moet stoppen. De problemen die kunnen ontstaan bij onterecht afstuderen mogen overduidelijk zijn.

Op grond hiervan zou men verwachten dat een groot deel van het beschikbare geld voor onderwijs en de beschikbare tijd

wordt besteed aan het optimaliseren van de gebruikte examens. In de onderwijspraktijk is dit echter lang niet altijd het geval: bij veel beslissingen ten aanzien van toetsing en de bepaling van zak-slaaggrenzen wordt aan traditie en intuïtie een groter gewicht toegekend dan aan de uitkomsten van de vele studies over deze onderwerpen. In dit artikel willen wij trachten een korte samenvatting te geven van de resultaten met hun implicaties voor de meest gebruikte examenvormen. Kort schetsen wij eerst een kader waarbinnen de verdere bespreking zal plaatshebben.

### Examens als voorspellers van later medisch handelen

Het uiteindelijke doel van alle examens is bij te dragen aan een prognose voor het latere handelen in de praktijk. Er zijn echter vier belangrijke storingsbronnen:

1. Er is een verschil tussen wat artsen kunnen (competence) en wat zij in hun dagelijks handelen werkelijk doen.<sup>1</sup>
2. Kennis en kunde vertonen de neiging geleidelijk af te nemen indien zij niet telkens worden gereactiveerd.<sup>2</sup>
3. Meet een examen werkelijk wat het bedoelt te meten, en is wat gemeten wordt

van examens, de laatste twee met de werkelijke inhoud van het examen. Het is daarom niet meer dan logisch eventuele oplossingen voor de problemen in de desbetreffende gebieden te zoeken.

### Historie van onderzoek van toetsing

#### Aandacht voor medische competentie

Rond de jaren zeventig zijn er veel studies gehouden waarin werd geprobeerd betere methodes te vinden om medische competentie te toetsen. Men ging uit van het concept dat medische competentie was opgebouwd uit diverse apart meetbare entiteiten. In de loop der jaren zijn veel van dergelijke entiteiten gesuggereerd: feitenkennis, inzicht, probleemoplossend vermogen, klinisch redeneren, oordeelsvorming, beslissingscompetentie, technische vaardigheden, communicatieve vaardigheden, attitudes, etc. Er zijn even zoveel toetssoorten ontwikkeld als aparte meetinstrumenten. Om de validiteit van dergelijke examenvormen vast te stellen, is een groot aantal correlatieve studies verricht waarbij de ene toetsvorm met de andere

*Examinatoren menen vaak al binnen vijf minuten te weten wat voor vlees zij in de kuip hebben*

wel een goede indicator voor medische competentie? Aan de validiteit kan dus worden getwijfeld. In examens krijgt kennis bijvoorbeeld vaak veel aandacht, maar het is maar de vraag of kennis wel zo'n ideale indicator is voor later medisch handelen.

4. Gebleken is dat examens als meting onnauwkeurig zijn. Deze betrouwbaarheidsproblemen zijn vooral van belang bij studenten die ternauwernood zijn gezakt of geslaagd. Bij een onbetrouwbare meting kan over deze studenten geen uitspraak worden gedaan, zij kunnen net zo goed zijn geslaagd als gezakt.

De eerste twee problemen hebben te maken met de planning en reglementering

werd vergeleken. De matige correlaties die werden gevonden, zag men als indicator dat inderdaad iets anders werd gemeten. Het is echter goed mogelijk dat de onbetrouwbaarheid van de gebruikte toetsvormen meer debet is aan deze lage correlaties dan werkelijke verschillen. De geschatte ware correlaties, waarbij gecorrigeerd wordt voor de onbetrouwbaarheid, leveren extreem hoge waarden op. Studies waarin zeer betrouwbare toetsen met elkaar worden vergeleken, tonen dan ook beduidend hogere correlaties. Dit was een moeilijk interpreteerbaar gegeven. Hoog correlerende variabelen hoeven niet perse hetzelfde te zijn (lengte en gewicht correleren hoog, maar zijn duidelijk ande- **2**

re variabelen); aan de andere kant geldt dat dingen die verschillende uitingen zijn van dezelfde variabele hoog met elkaar correleren (de lengtes aangegeven in inches of centimeters correleren hoog met elkaar). Los van deze discussie kan in ieder geval worden gezegd, dat de hoeveelheid unieke informatie die een bepaalde toetssoort geeft (ten opzichte van welke andere toetssoort dan ook) over de competentie van een student relatief gering is. Dit lijkt vooral onlogisch bij toetsvormen die veel van elkaar verschillen (bijvoorbeeld praktische en schriftelijke toetsen). Echter, ook bij verschillende schriftelijke vraagvormen is lang gesuggereerd dat ze een verschillend cognitief niveau meten. Van open vragen dacht men dat ze superieur zijn aan multiple choice-vragen, omdat voor het beantwoorden van een open vraag spontane generatie van kennis noodzakelijk zou zijn, en van een multiple choice-vraag het voldoende zou zijn het juiste antwoord te herkennen. Dit 'cueing'-effect is sinds de jaren zestig onderwerp geweest van veel studies.<sup>3</sup> Ook hier luidt de conclusie, dat de ware correlaties hoog zijn en de unieke informatie van beide vraagvormen zeer klein is. Wel werden er verschillen in gemiddelde scores gevonden ten nadele van de open vraag. De correctheid van de vaak getrokken conclusie dat er dus wat anders wordt gemeten, mag worden betwijfeld. Het voorbeeld met de lengtemeting in inches en centimeters mag dit verduidelijken: de waarde in centimeters is altijd groter dan die in inches, maar de correlatie is perfect.

#### **Domeinspecificiteit**

Om een zo hoog mogelijke validiteit van toetsen te realiseren, zijn in de jaren zeventig en tachtig verschillende casusgerichte toetsvormen voorgesteld en onderzocht. Het was de bedoeling de werkelijkheid zo nauwkeurig mogelijk te simuleren. Daarom moest de student de casus in zijn geheel doorlopen, waarbij alle beslissingen werden gescoord. Naast scoringsproblemen (hoeveel punten moesten worden toegekend aan welke beslissing?) en constructvaliditeitsproblemen (ervaren specialisten scoorden even hoog of lager dan pas afgestudeerden) bestond er een betrouwbaarheidsprobleem. Het vermogen van een kandidaat om de casus op te lossen bleek in hoge mate afhankelijk te zijn van de kennis die hij over het onderwerp van de casus had; bovendien bleek dat die kennis in hoge mate afhankelijk was van de inhoud van de casus.<sup>4</sup> Dit wordt domeinspecificiteit genoemd. Het impliceert dat de score die studenten halen op de ene casus een zeer lage correlatie heeft met

de score die zou worden bereikt op een willekeurige andere casus (ook binnen hetzelfde domein). Om een betrouwbaar examen te krijgen, moeten dus veel verschillende casus worden voorgelegd. Dit lijkt in eerste instantie onlogisch: examinatoren menen vaak al binnen vijf minuten te weten wat voor vlees zij in de kuip hebben. Maar als men bedenkt dat een examen een steekproef is uit alle mogelijke opgaven die hadden kunnen worden gesteld, is het logisch dat het aantal waarnemingen groot genoeg moet zijn om een betrouwbare uitspraak te kunnen doen over de competentie van een kandidaat. Verder geldt dat een meting nooit valide kan zijn als zij niet betrouwbaar is, dus geeft een betrouwbaarheidsprobleem van een toets automatisch een validiteitsprobleem.

#### **Invloed op studiegedrag**

Een belangrijke reden voor veel docenten om hun studenten aan examens te onderwerpen is de angst dat studenten anders niet (voldoende) zouden studeren. Er wordt vaak weinig aandacht besteed aan de invloed die de inrichting van de exa-

middel vormen die de examinator in staat stellen tot een dieper niveau door te dringen. Hierdoor zou het mogelijk zijn een goede indruk te krijgen van de kennis of de probleemoplossende vaardigheden. De betrouwbaarheid van deze examens is echter bijzonder laag. Oorzaken hiervoor zijn makkelijk aan te wijzen: de subjectiviteit van de examinator (streng of mild); door het dieper doorvragen per casus worden slechts weinig casus voorgelegd, wat een grote invloed van domeinspecificiteit geeft; de interpersoonlijke relatie tussen examinator en student; etc. Verder speelt dat de examinator vaak na korte tijd een oordeel heeft gevormd over de competentie van de student en hier ook bij een langere toets niet meer vanaf is te brengen: het halo-effect.

Dientengevolge kan de validiteit van een ongestructureerd mondeling niet hoog zijn. De score representeert in hoge mate allerlei aspecten die weinig tot niets met de *algemene* competentie van de kandidaat te maken hebben, maar met storingsbronnen van de meting.

De oplossing om de steekproef te vergroten door het examen te verlengen, zou de

*Er bestaan grote tegenstellingen tussen wat wordt geloofd en wat empirisch is vastgesteld*

mens zou kunnen hebben op het studeergedrag. Ook in de literatuur zijn relatief weinig studies te vinden waarin dit systematisch is onderzocht.<sup>5</sup> Redenen hiervoor kunnen zijn dat sommige effecten zo logisch zijn dat ze niet tot onderzoekbare vraagstellingen hebben geleid. Bovendien zijn dergelijke studies logistiek vaak moeilijk uitvoerbaar. Tot slot geldt dat verschillen in voorbereiding weliswaar door studenten kunnen worden gepercipieerd, maar vaak moeilijk objectiveerbaar zijn. Overeenkomsten in de conclusies van deze studies zijn dat studenten zich bij hun manier van studeren sterk laten leiden door het type examen dat ze verwachten, maar dat het niet duidelijk tot meetbare verschillen leidt.

#### **Consequenties voor toetsvormen**

De praktische consequenties van deze bevindingen voor de verschillende examens kunnen worden uitgesplitst naar toetsvormen (mondelinge, schriftelijke en praktische examens) en naar toetsplanning en -reglementering.

#### **Mondelinge toetsvormen**

Veel docenten hebben het gevoel dat mondelinge examens een goed toets-

examens zo lang maken (soms tot enkele dagen pure toetstijd) dat vermoeidheid een serieuze rol gaat spelen, of dat ze praktisch onuitvoerbaar worden.<sup>4</sup> Het heeft wel zin het mondeling examen te structureren en in plaats van een hele casus te doorlopen alleen enkele essentiële elementen te bevragen.

Een ander probleem bij het gebruik van mondelinge examens is dat te vaak (ook) feitenkennisaspecten worden getoetst. Gezien de docententijd die nodig is voor mondelinge examens is dit een uiterst inefficiënte methode. De tijd die dan wordt besteed aan het afnemen van mondelinge examens zou veel beter benut kunnen worden, bijvoorbeeld aan een zorgvuldige productie van schriftelijk examenmateriaal.

#### **Schriftelijke examenvormen**

Vaak wordt bij schriftelijke examens de voorkeur uitgesproken voor het gebruik van open vragen. Als grootste nadeel van gesloten vraagvormen wordt gezien dat deze alleen geschikt zijn voor de toetsing van triviale feitenkennis, terwijl open vragen meer op inzicht ingaan. Inderdaad kan gezegd worden dat de gesloten vraagvormen in het verleden nogal specifiek hiervoor zijn gebruikt. De ontwikkelin-

gen van de laatste jaren hebben echter geleid tot een bredere toepassing van gesloten vraagvormen, waarbij men steeds meer heeft geprobeerd beslissingen te toetsen. Indien er vergelijkingen worden gemaakt tussen goed geconstrueerde open vragen en gesloten vragen blijkt ook dat de scores veel overeenkomst met elkaar vertonen.

Een ander voordeel van gesloten vragen is dat ze sneller te beantwoorden zijn, waardoor er meer vragen per uur kunnen worden gesteld. Gezien het probleem van de domeinspecificiteit is een toets met gesloten vragen dus betrouwbaarder dan een toets met open vragen.

Gesloten vragen lijken moeilijker te construeren dan open vragen, maar hierbij wordt vaak geen rekening gehouden met het feit dat een goede open vraag ook een helder uitgeschreven antwoordsleutel bevat en dus ook veel tijd kost bij de constructie. Wel geldt, dat gesloten vragen gemakkelijker en met behulp van een computer te scoren zijn.

De beschreven nadelen gelden natuurlijk in veel ernstiger mate voor de essay-tentamens: slechts een zeer smal gebied van de hele domein wordt gesampled en meestal slechts door één examinator nagekeken. Hier is praktisch nooit een duidelijke vooraf vastgestelde antwoordsleutel te geven. Voor de toetsing van medische competentie zijn deze toetsen dan ook nog sterker af te raden dan open vragen. Alleen indien schrijfvaardigheid het doel van meting is, kan er een indicatie zijn voor deze vorm, mits weer duidelijke scoringsrichtlijnen vastliggen en de examinator zelf voldoende literair geschoold is.

#### Praktische examenvormen

Domeinspecificiteit is eveneens een groot probleem bij examens waarbij één of slechts enkele examenpatiënten worden opgevoerd. Dit levert onbetrouwbare en dus invalide scores op. Onbetrouwbaarheid werkt twee kanten op: men weet niet of iemand terecht is gezakt, maar zeker ook niet of iemand terecht is geslaagd. Structureren van de opgaven en een zo breed mogelijke sampling is een oplossing voor dit probleem. De examenpatiënt is daarom in veel instituten vervangen door een objectief gestructureerd stations-examen (OSCE). Hierbij doorloopt een student een circuit van verschillende kamers met verschillende examinatoren en verschillende opdrachten bij verschillende (simulatie)patiënten of modellen. De examinatoren hebben gestructureerde criterialijsten waarop zij bij ieder item kunnen aangeven of de student dat heeft gedaan. Dit levert een zo breed mogelijke sampling van opdrachten, examina-

toren en patiënten. Men moet ervoor oppassen de lijsten niet te veel te structureren, daar men dan het gevaar loopt ze te trivialisieren.<sup>6</sup>

#### Planning van examens

##### Afsluitende examens

Het grote gevaar van afsluitende examens is dat ze een ongewenst studiegedrag induceren. Een ieder herinnert zich uit zijn eigen studietijd hoe je blokte voor een examen, vervolgens dat examen haalde en het gevoel had voor er voor je verdere leven vanaf te zijn. Dit 'immunitetsprincipe' is verre van ideaal. Gechargeerd zou men kunnen stellen, dat studenten zich oppoetsen voor een examen, het examen halen en dan zo snel mogelijk alles weer vergeten om zich voor te kunnen bereiden op het volgende examen. Afsluitende examens stimuleren derhalve niet het vermeerderen van kennis, maar meer een soort hordenloopgedrag. Over het jaar verspreide examens waarin alle onderwerpen deels worden getoetst en die volgens combineringsregelingen werken, hebben een positiever effect op de kennisvergarig van studenten.

##### Herexamens

Herexamens lijken fair ten opzichte van de student, maar zijn dat niet altijd. Op de eerste plaats is een herexamen een herhaling van een meting, met alle statistische gevolgen van dien. Iedere medicus practicus kent het fenomeen dat indien men maar voldoende laboratoriumbepalingen laat uitvoeren er altijd wel een afwijkende wordt gevonden (bij een 95% betrouwbaarheidsinterval is deze kans 5%). Door de herhaalde meting wordt de kans op een fout-positieve uitslag steeds groter. Een fout-positieve uitslag (een ten onrechte geslaagde student) wordt ook steeds groter bij herexamens. Verder kosten ze de docenten veel extra werk en tijd, die beter in de kwaliteitszorg van de echte examens hadden kunnen worden gestoken. Docenten benadelen in die zin de andere studenten.

Ten slotte is een nadeel dat herexamens een minimalistische studiestrategie induceren. Studenten kunnen gerust proberen met marginale inspanning hun examen te halen omdat er voldoende vangnetten in de vorm van herexamens zijn. Het zou veel beter zijn een student meerdere examens te geven die alle bij elkaar worden opgeteld en compensatoir werken. Dit levert aan het einde van een studieperiode een meting op waarbij in het verloop van de tijd een zeer breed domein gesampled is en waarbij de studenten halverwege voldoende feed-back hebben gehad om hun studiegedrag zodig bij te stellen.

Helaas schrijft de nieuwe Wet op het Hoger en Wetenschappelijk Onderwijs het gebruik van herexamens zelfs voor. Een compromis zou dan kunnen liggen in het plannen van herexamens op voor de student ongunstige tijdstippen, of het optellen van hun scores bij eerder behaalde resultaten (een student moet dan een onvoldoende wegwerken).

##### Tot slot

Het maken van goed examenmateriaal is een tijdrovend en nie: altijd aangenaam werk. Het belang van goede toetsing is groot en de noodzaak een optimale efficiëntie te betrachten overduidelijk. Helaas bestaan er grote tegenstellingen tussen wat wordt geloofd en wat empirisch is vastgesteld. Dit maakt dat reeds lang bekende en belangrijke principes slechts langzaam in het medisch onderwijs zijn doorgesijpeld. Gelukkig vindt er steeds meer uitwisseling plaats tussen de medische faculteiten onderling en wordt ook in Nederland op steeds grotere schaal veel aandacht besteed aan onderzoek en ontwikkeling van toetsmateriaal. •

L.W.T. Schuwirth,

C.P.M. van der Vleuten,

A.J.J.A. Scherpbier,  
vakgroep Onderwijsontwikkeling en  
Onderwijsresearch, Rijksuniversiteit  
Limburg

#### Literatuur

1. Rethans JJ, Sturmans F, Drop MJ, Vleuten CPM van der, Hobus P. Does competence of general practitioners predict their performance? Comparison between examination settings and actual practice. *British Medical Journal* 1991; 303: 1377-80.
2. Norcini JJ, Lipner RS, Benson JA, Webster GD. An Analysis of the Knowledge Base of Practising Internists as Measured by the 1980 Recertification Examination. *Annual Internal Medicine* 1985; 102: 385-9.
3. Schuwirth LWT, Vleuten CPM van der, Donkers HJLM. Open-ended questions versus Multiple Choice Questions. An Analysis of Cueing Effects. In: Harden RM, Hart IR, Mulholland H. Eds. *Approaches to the Assessment of Clinical Competence*. part 2. Norwich: Page Brothers, UK, 1992.
4. Swanson DB, Norcini JJ, Grosso LJ. Assessment of clinical competence: written and computer-based simulations. *Assessment and Evaluation in Higher Education* 1987; 12(3): 220-46.
5. Newble DI, Jaeger K. The effect of assessments and examinations on the learning of medical students. *Medical Education* 1983; 17: 165-71.
6. Luijk SJ van. *Al doende leert men*. Proefschrift, Rijksuniversiteit Limburg Maastricht, Maastricht: Universitaire Pers, 1994.

Een uitgebreide literatuurlijst (37 titels) is op te vragen bij de auteurs: Vakgroep Onderwijsontwikkeling en Onderwijsresearch, Universiteit Limburg, Postbus 616, 6200 MD Maastricht