

TOETSING VAN MEDISCHE COMPETENTIE: STAND VAN ZAKEN EN ONTWIKKELINGEN

C.P.M. van der Vleuten
Projectleider
Project Evaluatie Studieresultaten
Rijksuniversiteit Limburg

Correspondentieadres:
Vakgroep Onderwijsontwikkeling
en -research
Rijksuniversiteit Limburg
Postbus 616
6200 MO Maastricht

Het beoordelen van studenten door middel van toetsen en examens behoort tot de vaste taken van elke docent. Bij het verrichten van deze taak laten docenten zich leiden door hun inhoudelijke deskundigheid, de heersende gewoontes en hun persoonlijke ervaringen met toetsing tijdens hun eigen opleiding. Vaak wordt daarbij uitgegaan van onbewezen, intuïtieve veronderstellingen, bijvoorbeeld over de relatie tussen toetsvorm en doel van de meting (kennis, inzicht, toepassing), over de objectiviteit en betrouwbaarheid van verschillende methodes, en over de zekerheid die nodig is om besluiten over studenten te kunnen nemen. In het medisch onderwijs is in de laatste decennia veel onderzoek gedaan naar een aantal van deze veronderstellingen. Ook is er veelvuldig geëxperimenteerd met nieuwe toetsmethoden. Dergelijk onderzoek heeft nogal eens aanleiding gegeven tot het plaatsen van grote vraagtekens bij een aantal van de uitgangspunten van de 'intuïtieve' benadering. In zijn algemeenheid heeft dit onderzoek vooral geleid tot meer kennis over de evaluatie van studieprestaties van studenten. Het doel van dit artikel is om deze kennis samen te vatten en er enkele consequenties en suggesties voor de praktijk uit te distilleren. Onderdelen van dit artikel komen voort uit de summer course 'New methods in student assessment', gehouden in Maastricht in 1989, in samenwerking met Prof. Dr. G.R. Norman van de McMaster University in Canada.

Het artikel is opgebouwd uit vier onderdelen. In de eerste plaats zal een historische schets worden gegeven van de ontwikkelingen in de toetsing van medische competentie. Daarna zal een aantal bevindingen en ervaringen worden beschreven, die zich vervolgens laten vertalen in enkele regel- of wetmatigheden. Tenslotte zullen hieraan enkele suggesties voor toetsing van medische competentie worden gekoppeld.

EEN HISTORISCH PERSPECTIEF

Figuur 1 geeft in schemavorm een historisch overzicht van de ontwikkelingen in de toetsing van medische competentie in de afgelopen 40 jaar. Voor achtergrondliteratuur bij dit schema en meer informatie omtrent de genoemde instrumenten is een lijst met aanbevolen literatuur opgenomen.

In figuur 1 is verticaal de tijd weergegeven en horizontaal de mate waarin de toetsingsprocedure de werkelijkheid benadert. Dit laatste aspect, het streven naar een meer rea-

listische benadering van de werkelijkheid in toetsen, vormt een rode draad in de ontwikkeling van alternatieve toetsmethoden. Met het zoeken naar methoden die de (beroeps)-praktijk beter benaderen, wordt gestreefd naar een meer valide meting van kennis en vaardigheden die relevant zijn voor de (latere) beroepspraktijk. Deze dimensie wordt ook wel eens aangeduid als de mate van (in-) directheid van de meetmethode.¹ Achtereenvolgens zullen nu enkele belangrijke toetsmethoden die sinds 1950 een rol hebben gespeeld in het medisch onderwijs in dit historisch overzicht besproken worden.

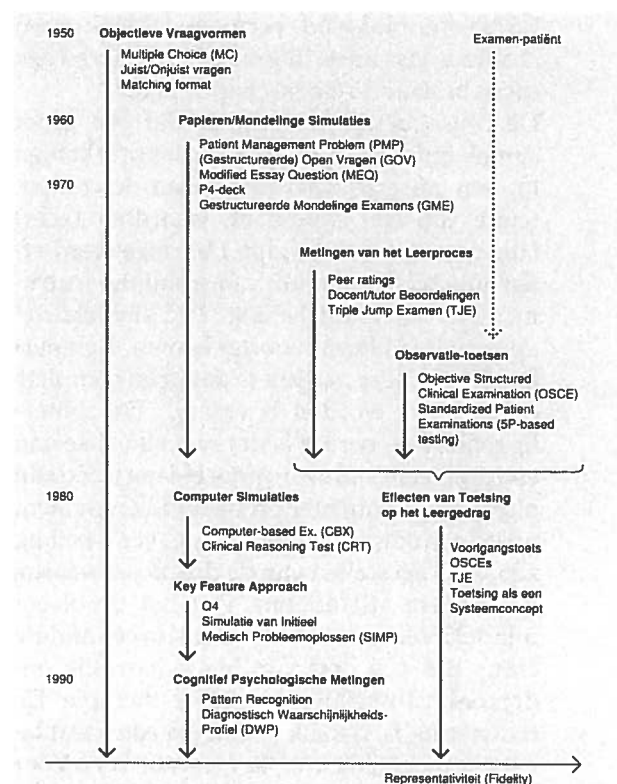
Objectieve vraagvormen

Het overzicht start in de vijftiger jaren, toen, met de schaalvergroting in het onderwijs, de 'objectief scorebare vraagvorm' massaal werd ingevoerd. De multiple-choice vragen en andere soortgelijke gesloten vraagvormen werden op grote schaal geïntroduceerd. Met de opkomst van de computer werd het gebruik van deze vraagvorm nog eens versterkt. De objectieve vraagvorm is tegenwoordig niet meer weg te denken uit het medisch onderwijs.

Met de opkomst van de gesloten vraagvorm groeide een gevoel van onbehagen: dit type vragen zou slechts 'kale' feitenkennis toetsen en geen vaardigheden van een hogere cognitieve orde kunnen meten. Multiple-choice toetsen vereisen 'slechts' herkenning van feiten en geen actieve toepassing van kennis die de student zich eigen heeft gemaakt (in de Angelsaksische literatuur wordt dit fenomeen aangeduid als het probleem van recognition versus recall). Ook inzicht en probleemoplossend vermogen zouden niet of slechts in beperkte mate met deze vraagvorm kunnen worden getoetst. Daarmee zou deze beoordelingsmethode te ver af staan van de praktijk en zou een onvolledig en mogelijk zelfs niet-valide beeld gekregen worden van de competentie van een student.

Simulaties

Deze ontevredenheid leidde tot een stroom van nieuwe instrumenten. Vooral het streven naar het meten van probleemoplossend vermogen kwam daarbij centraal te staan. De basis voor deze nieuwe instrumenten vormde de simulatie. Aan de kandidaat werd een praktijksituatie - doorgaans een patiëntenprobleem - voorgelegd, en het 'probleemoplossend vermogen' van de kandidaat werd beoordeeld op grond van de wijze waarop het probleem werd aangepakt en afgehandeld. De meest verbreide exponent is het Patient Management Problem (PMP), dat in de zestiger jaren ontwikkeld werd. Het PMP is een schriftelijk instrument waarbij de student een casus doorloopt. Sommige PMP's kenden allerlei vertakkingen, afhankelijk van het verloop van de casus. Er werden ingenieuze druktechnieken toegepast (latent images), waarmee de kandidaat opties kon aangeven zonder dat er sprake was van sturing of het geven van aanwijzingen (cueing). In de zeventiger jaren heeft het PMP een grote vlucht genomen en vormde het een belangrijk onderdeel van de Amerikaanse nationale arts-examens. In figuur 1 worden nog andere voorbeelden van soortgelijke instrumenten genoemd. Naast schriftelijke vormen zijn ook mondelinge examens gebruikt voor het toe-



passen van simulatie-methodes. Bij deze examens speelt de examiner - meestal aan de hand van een protocol - de rol van patiënt. Een moderne variant deed zijn intrede met de introductie van de computer. Met de computer was het mogelijk een aantal nadelen te ondervangen (zo waren ingewikkelde en kostbare druktechnieken niet meer nodig) en konden complexere simulaties worden gerealiseerd, die bovendien konden worden uitgebreid met realistische beelden (dia's, video, beeldplaat). Het CBX-project van de National Board of Medical Examiners is een van de meest prestigieuze projecten op dit gebied met zeer complexe en fraaie computersimulaties van patiëntenproblemen.

De scoring van simulatie-instrumenten levert problemen op. Bovendien kwam nog een belangrijk nadeel naar voren, dat doorslaggevend is geweest voor de verdere ontwikkeling van deze toetsvorm. De score verkregen op de ene simulatie of casus bleek nauwelijks voorspellende waarde te hebben voor de score op een andere simulatie of casus. Competentie bleek sterk af te hangen van de toevallige inhoud van een casus en (minimale) verandering van de inhoud leverde reeds een totaal andere waardering op.

Figuur 1. Toetsing van Medische Competentie vanuit een Historisch Perspectief. Zie voor aanbevolen literatuur de lijst aan het eind van het artikel.

Probleemoplossend vermogen bleek geen absolute, los van de inhoud of situatie van een casus bestaande eigenschap te zijn.

De consequentie daarvan is dat een groot aantal simulaties of casus nodig is om te komen tot een correcte waardering van de competentie van een kandidaat, waardoor (zeer) lange toetsen vereist zijn. Daarmee werd efficiëntie bij het gebruik van simulatie-instrumenten van groot belang. De 'key-feature' approach is hieruit voortgekomen. Kenmerkend voor deze aanpak is dat geen complete casus meer worden bevestigd. De achterliggende redenering is dat niet elke fase van een probleem van even groot belang is. Zo kan bij sommige patiënten problemen één element uit de anamnese van doorslaggevend belang zijn voor het stellen van de diagnose, waarbij de verdere afhandeling van het probleem minder interessant of relevant is. In een andere casus kan een deel van het lichamelijk onderzoek van belang zijn, of de therapie. De redenering is, dat elk probleem een karakteristieke moeilijkheid of 'key feature' kent. Voor toetsdoeleinden hoeft niet de volledige casus bevestigd te worden, maar kan men volstaan met deze karakteristieke moeilijkheid. Door bevestiging van alleen de 'key features' wordt veel tijd bespaard en kunnen veel meer casus aan de kandidaat worden aangeboden. Het Q4-project van het instituut voor de nationale medische examens in Canada is daar een grootschalig voorbeeld van.² Een in Nederland ontwikkelde exponent van deze aanpak is de SIMP.³

Niet alleen bleek dat probleemoplossend vermogen geen absolute eigenschap was, bovendien bleek uit onderzoek, dat ervaren personen (bijvoorbeeld klinici) nauwelijks beter - en soms zelfs slechter - presteerden op probleemoplostaken dan minder ervaren personen (bijvoorbeeld basisartsen of studenten).^{4,7} Deze bevindingen hebben ertoe bijgedragen dat er meer fundamenteel (cognitief psychologisch) onderzoek is gedaan naar probleemoplossen. Dit onderzoek heeft ertoe geleid dat er onderscheid gemaakt wordt tussen verschillende competentie-fasen, of verschillende niveaus van expertise.^{8,9} Het voert te ver om op deze plaats de theoretische ontwikkelingen te beschrijven, maar cruciaal hierin is de bevinding dat competentie of expertise een ontwikkeling vertoont van een conceptueel rijk en rationeel kennisbestand naar een ervaringsgericht en niet-analytisch

conceptueel vermogen. De 'klassieke' simulaties zouden vooral een beroep doen op de rationele, analytische kennis, waarin redeneringen en integratie van kenniselementen een belangrijke plaats hebben, maar zij zouden minder geschikt zijn voor de evaluatie van het hoogste niveau van expertise. Het herkennen van patronen en een directe en zeer efficiënte aanpak zijn belangrijke karakteristieken van het hoogste niveau van expertise. Hiermee zijn we aangeland bij de 'cutting edge' van de stand van zaken op dit gebied. De vertaling van deze theoretische inzichten naar de toetspraktijk dient nog plaats te vinden. Er worden pogingen ondernomen om de 'laboratoriuminstrumenten' waarmee het cognitief psychologisch onderzoek wordt uitgevoerd te vertalen naar praktisch bruikbare instrumenten, en er zijn enkele instrumenten ontwikkeld die min of meer aansluiten bij deze theoretische inzichten.¹⁰ Deze laatste worden in figuur 1 vermeld. In de komende jaren zijn verdere ontwikkelingen op dit gebied te verwachten.

Metingen van het leerproces en de effecten van toetsing

Met de opkomst van nieuwe onderwijs-systemen en de toenemende aandacht voor onderwijskundige aspecten in het medisch onderwijs werden in de zeventiger jaren instrumenten ontwikkeld die een rechtstreekse evaluatie beoogden van het leerproces van studenten. Bij probleemgestuurd onderwijs wordt grote waarde gehecht aan het kunnen samenwerken en aan het vermogen om snel informatie te verzamelen. Om de prestaties van studenten op deze terreinen te meten werden verschillende toetsinstrumenten ontwikkeld. De Triple Jump Exercise (TJE) is een toetsinstrument, waarmee de snelheid en kwaliteit van informatieverzameling (en -verwerking) worden geëvalueerd. De TJE bestaat uit drie stappen: 1) een geprotocolleerd mondeling examen over een bepaald onderwerp (meestal één of verschillende patiëntproblemen), 2) een studieopdracht die binnen een etmaal moet worden uitgevoerd en 3) een mondeling examen waarin de studieopdracht wordt getoetst. Het vermogen tot samenwerken in een groep werd gemeten met beoordelingen door medestudenten uit dezelfde groep (peer ratings), met zelfbeoordelingen en docentbeoordelingen. Dergelijke instrumenten hebben echter nooit een grote vlucht

genomen. Wellicht is dat te verklaren uit de poore psychometrische bevindingen van deze toetsen.¹¹

Ondanks de beperkte toepassing van instrumenten voor de evaluatie van het leerproces, waren de ontwikkelingen op dit gebied wel van historisch belang. Ze gaven blijk van een toenemende aandacht voor de effecten van toetsing op het leren van de student (dit vormde overigens ook voor de simulaties reeds een belangrijk motief, met name in opleidingen met volgens onderwijskundige principes vernieuwde curricula). Het werd in toenemende mate duidelijk dat er van toetsen en examens een grote sturende werking uitging op het leerproces. Vernieuwingen in het onderwijsprogramma hadden weinig effect als het toetsingsprogramma daar niet op aansloot. Fragmentarische, soms irrelevante, detaillistische en oppervlakkige kennis, het snel vergeten van leerstof en het 'pieken' van studenten in sommige periodes afgewisseld met 'dalen', waren vaak rechtstreeks het gevolg van het examenprogramma, ongeacht alle goede bedoelingen van de docent en het onderwijsprogramma.¹² Dit besef, laten we zeggen van toetsing als onderwijskundig fenomeen, is een belangrijk gegeven geworden. Deze notie werd in toenemende mate 'instrumenteel' gebruikt om gewenste effecten op het leergedrag te bewerkstelligen. De Maastrichtse Voortgangstoets is een voorbeeld van een instrument waaraan deze gedachte ten grondslag ligt.¹³ De voortgangstoets is een kennistoets bestaande uit een steekproef uit alle einddoelen van de opleiding, die periodiek op hetzelfde tijdstip bij alle studenten van de opleiding wordt afgenomen. Specifieke voorbereiding is daardoor haast niet mogelijk en dit bevordert dat studenten regelmatig studeren op geleide van hun persoonlijke studiedoelen.

De toetsing als 'systeemconcept', waarbij doelbewust strategieën worden gevolgd die gewenst studiegedrag stimuleren en ongewenst leergedrag tegengaan, is een logische volgende stap waarmee aanzienlijke onderwijskundige winst zou kunnen worden geboekt. Dit veronderstelt echter een meer centrale aanpak van toetsing dan gebruikelijk is in de meeste onderwijsprogramma's. Op het aspect van toetsing en de invloed daarvan op het studiegedrag van studenten zal nog worden teruggekomen.

Observatietoetsen

Bij de ontwikkeling van observatietoetsen speelden niet alleen de effecten op het leergedrag een rol, maar ook het streven naar een voor de beroepspraktijk representatieve evaluatie. Observatietoetsen confronteren studenten met (gesimuleerde) praktijksituaties, waarin hun prestaties worden beoordeeld. Deze wijze van toetsing kan worden gezien als een evolutie van het traditionele examen, waarin een kandidaat aan de hand van een concrete patiënt werd geëxamineerd. Deze examenprocedure had echter sterk subjectieve kanten en bleef bovendien doorgaans beperkt tot één casus. Observatietoetsen of stationsexamens combineren de hoge realiteitswaarde van de oorspronkelijke examenvorm met de gewenste toetscondities van standaardisatie en objectiviteit. Een stationsexamen bestaat uit een serie zogenaamde 'stations', afzonderlijke ruimtes of kamers waarin een kandidaat een opdracht moet uitvoeren (bijvoorbeeld, het afnemen van een anamnese, of het verrichten van een lichamelijk onderzoek). De prestaties worden geregistreerd door een examiner, doorgaans aan de hand van geëxpliciteerde criterialijsten. Na een vastgestelde periode gaat de student naar een volgend station. Een examen bestaat uit een circuit van stations, dat door een aantal kandidaten tegelijk wordt doorlopen.¹⁴ De inhoud of de aard van de opdrachten kan nogal verschillen van station tot station of van toets tot toets. Opdrachten kunnen bestaan uit het afhandelen van complete patiëntproblemen (anamnese, lichamelijk onderzoek, differentiaal diagnose, beleid, voorlichting), maar ook uit onderdelen hiervan (bijvoorbeeld één bepaald lichamelijk onderzoek of het voeren van een slecht-nieuws gesprek). Soms moeten studenten laten zien dat zij basale psychomotorische vaardigheden beheersen, zoals het toedienen van injecties, soms krijgen zij opdracht bepaalde laboratoriumtesten uit te voeren of röntgenfoto's te beoordelen. In deze toetsen staat de *handeling* altijd centraal (in het Engels: hands-on behaviour) en wordt de nadruk gelegd op de competentie in een toepassingscontext.

In het medisch onderwijs hebben observatietoetsen de laatste tien jaar een enorme vlucht genomen. Ondanks de logistieke inspanningen en de hoge kosten zijn zij zeer populair geworden. Wellicht is deze populariteit te

verklaren uit hun hoge werkelijkheidsgehalte. Deze toetsvorm combineert de voordelen van gestandaardiseerde toetsing met een realistische benadering van de beroepspraktijk. Van de student wordt niet alleen beheersing van kennis verlangd, maar ook daadwerkelijke toepassing hiervan in representatieve situaties. Ook dit heeft consequenties voor de manier waarop studenten zich voorbereiden: hun leerstijlen en competentie worden hierdoor beïnvloed.^{15,16} In de laatste jaren is betrekkelijk veel onderzoek gepubliceerd over observatie-toetsen en inmiddels is vrij veel bekend over hun meeteigenschappen en de daaruit voortvloeiende praktische consequenties.¹⁷

Zoals uit dit historische overzicht blijkt, zijn er in de afgelopen decennia nogal wat ontwikkelingen geweest in de toetsing van medische competentie. Het streven naar een meer realistische benadering van de werkelijkheid, het verfijnen van bestaande en het ontwikkelen van nieuwe toetsprocedures, en de toenemende aandacht voor toetsing als onderwijskundig concept hebben daarbij centraal gestaan. De meeste van deze ontwikkelingen gingen gepaard met wetenschappelijk onderzoek. Daar waar dit van belang was voor de verdere evolutie van instrumenten zijn de bevindingen van dit onderzoek hierboven reeds aangegeven. In de navolgende paragraaf zullen we de consistenties in dit onderzoek samenvatten.

BEVINDINGEN VAN ONDERZOEK

De bevindingen van onderzoek kunnen worden ingedeeld in een drietal gebieden. In de eerste plaats zijn er de resultaten die betrekking hebben op de validiteit en betrouwbaarheid van toetsen. Op dit gebied heeft verreweg het meeste onderzoek plaatsgevonden. Ten tweede zijn er de bevindingen ten aanzien van de relatie tussen onderwijsprogramma en toetsing. En tenslotte zijn er uitkomsten van onderzoek naar de invloed van toetsen op het leergedrag van de student.

Betrouwbaarheid en validiteit

Scores generaliseren slecht over problemen, taken en situaties.

Dit is in eerste instantie gebleken uit onderzoek naar simulatie-instrumenten, met name

voor de toetsing van probleemoplossen. Het bleek dat de score van een kandidaat op het ene probleem nauwelijks voorspellende waarde had voor de score op een ander probleem; de intercorrelatie tussen scores op verschillende simulaties was laag. Hieruit werd geconcludeerd dat probleemoplossend vermogen sterk afhangt van de inhoud van een probleem. Dit fenomeen werd aangeduid als het probleem van de 'inhoudsspecificiteit' of 'casusspecificiteit'.¹⁸

Het probleem van de inhoudsspecificiteit bleek echter niet beperkt tot het terrein van het probleemoplossen, maar deed zich voor bij alle mogelijke metingen van medische competentie, zoals computersimulaties,¹⁹⁻²¹ open vragen,³ mondelinge examens,²² statusbeoordelingen²³ en stationsexamens.¹⁷

Dit betekent dat relatief veel problemen, taken of situaties in een toets moeten worden opgenomen, voordat men uitspraken kan doen over de competentie van een kandidaat. De in een toets opgenomen taken vormen slechts een steekproef uit alle mogelijk op te nemen taken en uiteraard wil men uitspraken kunnen doen die niet afhankelijk zijn van de toevallige toetsinhoud. Korte toetsen met weinig taken zullen dan ook onbetrouwbare metingen opleveren. Van de meeste hierboven genoemde instrumenten kan men stellen dat de bereikte betrouwbaarheid onvoldoende was. Voor het bereiken van een adequate betrouwbaarheid zijn veel taken, problemen of situaties noodzakelijk, hetgeen leidt tot lange toetstijden. De benodigde toetslengte heeft natuurlijk ook te maken met de efficiëntie van een taak. Er zijn bijvoorbeeld stationsexamens waarin een enkel station meer dan een half uur in beslag neemt, waardoor de benodigde toetstijd kan oplopen tot enkele dagen!²⁴

Scores generaliseren goed over testen/of vraagvormen.

Geheel tegengesteld aan de vorige bevinding is gebleken dat er hoge correlaties worden gevonden tussen scores verkregen met verschillende instrumenten en/of vraagvormen. Hoewel men intuïtief geneigd is te veronderstellen dat bijvoorbeeld multiple-choice vragen niet hetzelfde meten als open vragen, blijkt uit onderzoek steeds weer dat de intercorrelaties tussen scores verkregen met beide vraagvormen hoog tot zeer hoog zijn. En dit geldt wederom niet uitsluitend bij deze specifieke vraagvormen, maar ook bij andere

(zeer) uiteenlopende vraagvormen en instrumenten zoals open en gesloten vragen,^{25 26} Patient Management Problems en gesloten vragen,²⁷ mondelinge examens, computersimulaties en schriftelijke instrumenten.²⁸⁻³⁰ Zelfs tussen schriftelijke metingen en observatie van medische vaardigheden zijn hoge correlaties gevonden.³⁰

Opgemerkt dient te worden dat een hoge correlatie niet automatisch betekent dat hetzelfde gemeten wordt. Lengte en gewicht zijn bijvoorbeeld sterk aan elkaar gecorreleerd, maar beide zijn duidelijk andere entiteiten. Het is geheel afhankelijk van de doelstelling van een evaluatie, in hoeverre een hoge correlatie betekenis heeft. Vanuit een beslistkundig perspectief, waarbij men alleen geïnteresseerd is in de uitkomst van een beslissing en de ermee gepaard gaande efficiëntie danwel kosten, betekent een hoge correlatie tussen twee instrumenten dat weinig unieke informatie wordt verzameld en dat het weinig uitmaakt met welk instrument het besluit genomen wordt. In dat geval zal de voorkeur uitgaan naar het efficiëntste of goedkoopste instrument. De hoge correlaties die werden gevonden tussen gesloten vragen en Patient Management Problems was aanleiding voor de Amerikaanse National Board (makers van nationale artsexamens) om de kostbare PMP's uit het examenprogramma te schrappen. Deze leverden namelijk nauwelijks andere beslissingen op.

Vanuit een onderwijskundig perspectief liggen de zaken echter geheel anders. Wanneer bijvoorbeeld een stationsexamen zou worden vervangen door schriftelijke metingen, ligt het voor de hand dat studenten zich anders zullen gaan voorbereiden. De student zal bij de voorbereiding meer aandacht besteden aan het cognitieve aspect dan aan de handeling zelf.³² Dat zou een zeer ongewenst gevolg zijn. Vanuit een onderwijskundig perspectief zijn hoge correlaties dus van geringere betekenis. De belangrijkste conclusie die kan worden getrokken uit de bevinding dat scores in het algemeen goed generaliseren over test- en vraagvormen is de ontmanteling van de intuïtieve veronderstelling dat de vraagvorm en het oogmerk van de meting rechtstreeks aan elkaar gerelateerd zijn. Niet de vorm van een toets bepaalt wat er gemeten wordt, maar de inhoud of de specifieke taak die de kandidaat moet uitvoeren.³³ Het is bijvoorbeeld mogelijk om met een gesloten vraag inzicht of

toepassing van kennis te bevragen, maar het is ook mogelijk om met een open vraag kale feitenkennis te toetsen. Een en ander is geheel afhankelijk van de gestelde taak.

Een beetje structuur en standaardisering heeft reeds een gunstig effect.

Aanmerkelijke winst kan worden geboekt met het aanbrengen van enige structuur en standaardisering in toetsen. Het is algemeen bekend dat de betrouwbaarheid laag is bij globale beoordelingen in ongestructureerde situaties (bijvoorbeeld stagebeoordelingen),³⁴ bij mondelinge examens³⁶ en bij open vragen (tengevolge van de variatie tussen verschillende correctoren³⁵). Met het aanbrengen van enige structuur kan daarin een goede verbetering worden bereikt.³⁷ Het verstrekken van antwoordsleutels bij open vragen³⁸ en het volgens een protocol standaardiseren van mondelinge examens zijn hiervan voorbeelden.³⁹

Te veel structuur werkt averechts.

Hoewel het aanbrengen van structuur een positieve invloed heeft op de betrouwbaarheid van de meting, kan een teveel aan structuur averechts werken.⁴⁰ Als de beoordeling van communicatieve vaardigheden geschiedt aan de hand van gedetailleerde concrete checklist-items bestaat het gevaar dat de inhoud triviaal wordt. Sommige Patient Management Problems zijn dermate ingewikkeld (en ook sommige computersimulaties zoals bijvoorbeeld CBX) dat zij aan realiteitswaarde inboeten. Ervaren beoordelaars hebben vaak weinig waardering voor sterk gestructureerde beoordelingen vanwege het keurslijf karakter ervan.⁴¹ Ook studenten ervaren een teveel aan structuur en detaillering als een hinderlijk keurslijf. Van Luijk et al. rapporteren dat studenten die beoordeeld worden met sterk gedetailleerde criteria lijsten in een stationsexamen de neiging hebben lijsten uit het hoofd te leren zonder dat er sprake is van enig begrip van de gedemonstreerde vaardigheden.⁴² Een te sterke structurering brengt het risico met zich mee dat de negatieve effecten de positieve overschaduwen.

Objectiviteit is niet hetzelfde als betrouwbaarheid. Gerelateerd aan de vorige twee bevindingen is de wellicht tegen-intuïtieve bevinding dat betrouwbaarheid niet hetzelfde is als objectiviteit. Met de objectiviteit van een meet-

instrument wordt doorgaans de betrouwbaarheid van een instrument bedoeld, maar men dient zich te realiseren dat deze twee begrippen niet identiek zijn.⁴³ Een 'subjectief' instrument *kan* betrouwbare resultaten opleveren en een 'objectief' instrument *kan* onbetrouwbare resultaten geven. Een korte multiple-choice toets zal onbetrouwbare uitspraken opleveren en een toets met voldoende open vragen kan zeer betrouwbaar zijn.²⁶ In het laatste geval wordt de onbetrouwbaarheid tengevolge van verschillen tussen de correctoren gecorrigeerd door het aantal vragen. Een ander voorbeeld is dat globale beoordelingschalen niet noodzakelijkerwijs minder betrouwbare scores opleveren.⁴¹ Een en ander is volstrekt afhankelijk van de verhouding tussen de foutenbronnen die de toets situatie beïnvloeden.⁴⁴

Opgemerkt dient te worden dat de redenering niet zo maar omgedraaid mag worden: subjectieve metingen zijn niet even betrouwbaar als objectieve instrumenten. De toets situatie (mate van standaardisatie, instructie aan kandidaten, periode waarover beoordeeld wordt, toewijzing of indeling van beoordelaars aan kandidaten, etcetera) bepaalt of, en in hoeverre, subjectieve metingen de betrouwbaarheid negatief beïnvloeden.^{43 40} Van belang is het besef dat objectiviteit niet identiek is aan betrouwbaarheid, en dat de keuze van een meetinstrument niet automatisch afhankelijk dient te zijn van een veronderstelde inherente superioriteit van een instrument op grond van de mate van subjectiviteit of objectiviteit.

Toetsing en het onderwijsprogramma.

Er zijn twee ingangen denkbaar als men de relatie tussen toetsing en het onderwijsprogramma bestudeert. Op micro-niveau kan de vraag gesteld worden in hoeverre toetsing kan worden gebruikt om veranderingen in het onderwijsprogramma te bewerkstelligen. Op macro-niveau kan de vraag gesteld worden of resultaten van toetsing kunnen worden gebruikt om opleidingen of verschillende onderwijsprogramma's met elkaar te vergelijken.

Wordt toetsing (of gebrek aan resultaten) gebruikt voor veranderingen in onderwijsprogramma's?

Er zijn weinig publikaties over dit aspect te vinden. Het komt blijkbaar weinig voor dat toetsresultaten als programma-evaluatie

worden gebruikt. Voorzover studieresultaten worden gebruikt voor programmamodificatie lijkt dat alleen het geval te zijn wanneer sprake is van excessief lage rendementen, bijvoorbeeld wanneer een grote meerderheid van een jaargroep zakt. In het algemeen bestaat er een tendens om slechte studieresultaten eerder toe te schrijven aan de (falende competentie van) studenten, dan aan het onderwijsprogramma. Het is niet uitgesloten dat deze conclusie vaak (ten dele) terecht is, maar een impliciete keuze in die richting is op zijn minst een eenzijdige benadering. Ook docenten zouden kunnen leren van de resultaten van hun studenten.

Wordt toetsing gebruikt voor vergelijking van onderwijsprogramma's?

Ook bij deze vraag kan gesteld worden dat publikaties op dit gebied betrekkelijk zeldzaam zijn. Er zijn her en der vergelijkingen gemaakt tussen nieuwe en meer traditionele onderwijsprogramma's (bijvoorbeeld met betrekking tot probleemgestuurd onderwijs), maar zeker in Nederland bestaat er nauwelijks een cultuur op dit gebied.⁴⁵ Toch ligt het voor de hand dat studieresultaten belangrijke informatie kunnen opleveren over de kwaliteit van onderwijsprogramma's. Of zoals De Groot het stelde: "Kwaliteit van onderwijs moet blijken".⁴⁶ Wellicht dat met de toenemende belangstelling voor kwaliteitsbewaking in het hoger onderwijs op dit gebied in de toekomst meer te verwachten is.

Toetsing en de student.

Ook wanneer men kijkt naar de relatie tussen toetsing en de student zijn twee verschillende vragen relevant.

Leren studenten als gevolg van toetsing?

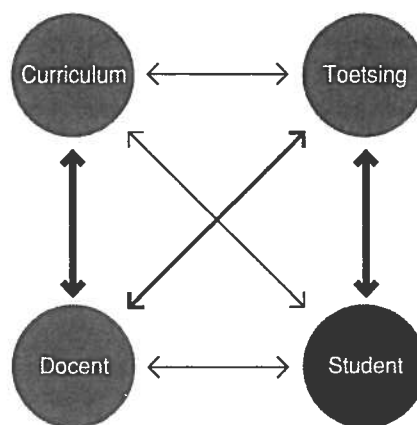
Deze vraag kan zonder verwijzing naar onderzoek met een volmondig ja worden beantwoord, omdat elke docent dat uit de praktijk kent. Studenten leren zelfs vóór al tengevolge van toetsing. Het zijn met name de tentamens en examens die bepalen of een student studiesucces heeft. Elke student is dan ook vooral geïnteresseerd in hoe deze hordes het beste kunnen worden genomen en - enigszins gechargeerd - de rest is bijzaak. De toenemende maatschappelijke druk om de studieduur te beperken zal deze invloed alleen maar versterken.

De vraag is echter of toetsgericht leren ook

daadwerkelijk effectief is. Vanuit de leerpsychologie zijn de snelle vergeetcurves bekend.⁴⁷ De vraag is in hoeverre het studeren ten behoeve van toetsing een beklijvend effect heeft op de lange termijn. Het antwoord daarop hangt natuurlijk mede af van de aard van de toetsing (dat is het volgende aspect dat zal worden besproken), maar doorgaans geldt dat de voorbereiding voor tentamens en examens gericht is op de korte termijn. Studenten studeren vlak voor een examen enorm hard om hun slaagkans te vergroten, om zich vervolgens volledig te kunnen concentreren op het volgende examen.¹² Het 'pieken' van studenten op bepaalde momenten is echter geen garantie dat de leerstof behouden blijft. Ook in vergelijkingen tussen onderwijsprogramma's zijn dit soort pieken geconstateerd, maar deze lijken niet te resulteren in lange-termijn effecten.^{45 48}

Heeft de wijze van toetsing invloed op de wijze van studeren?

Veel auteurs hebben gewezen op het belang van de relatie tussen de wijze van toetsing en het studeergedrag van studenten.^{49 50 15 51 52 31} Newble en Jaeger hebben bijvoorbeeld aangetoond dat studenten na de introductie van het stationsexamen veel meer tijd gingen besteden aan het leren van klinische vaardigheden dan aan het voorbereiden van theoretische examens¹⁵ en dit effect bleek consistent over de jaren heen.¹⁶ Met name in de laatste jaren begint men zich dit in het medisch onderwijs te realiseren. Onderwijskundige veranderingen in het onderwijsprogramma, die niet gepaard gaan met overeenkomstige veranderingen in het toetsingssysteem zijn bij voorbaat gedoemd te mislukken. Er bestaat bijna geen verband dat duidelijker is dan dat tussen wat studenten doen in een onderwijsprogramma en de wijze van toetsing. Het is dan ook merkwaardig dat dit besef nog zo weinig is doorgedrongen. Figuur 2 geeft een overzicht van de min of meer geldende situatie. Als docenten of onderwijsmakers zijn we geneigd veel energie te stoppen in het curriculum, doorgaans veel meer dan in het toetsingsprogramma (weergegeven door de dikte van de pijl). Er wordt gesproken over curriculumherzieningen, didactische methodieken, onderwijsmethoden, study load, etcetera. De studenten daarentegen, laten zich vooral leiden door het examenprogramma. De inhoud en de vorm daarvan bepalen het leerproces in



Figuur 2.
Aard van de relaties
tussen de elementen van
een onderwijsprogramma

veel sterkere mate dan het onderwijsprogramma. Op zich is daar niets op tegen, voorzover er een duidelijke relatie bestaat tussen het onderwijsprogramma en de toetsing. Maar juist bij dat aspect kan men vraagtekens plaatsen. Doorgaans besteden docenten veel minder aandacht aan het toetsprogramma, en worden toetsen en examens als iets vanzelfsprekends beschouwd waardoor zij vaak de sluitpost op de (individuele) onderwijsbegroting vormen. Dit geldt in nog sterkere mate voor toetsen met een formatief karakter (gericht op het proces en de voortgang van leren voor feedbackdoeleinden en minder op besluitvorming). Toetsen met een ander doel dan het nemen van besluiten over de studievoortgang lijken per definitie van minder belang. Bij de auteur is ook geen enkele majeure onderwijsverandering bekend waarin het toetsingsgedeelte niet pas op de laatste plaats of in het geheel niet aan bod kwam. Wanneer we de relaties uit figuur 2 zien, is dat toch een merkwaardige situatie.

Niettemin is de relatie tussen toetsen en studeren de afgelopen jaren meer in de belangstelling komen te staan. In toenemende mate wordt ingezien dat er tussen de doelstellingen van het curriculum en de aard van het toetsingsprogramma een direct verband behoort te zijn. Door toetsing als een systeemconcept op te vatten kan men recht doen aan deze relatie.⁵³ In dit concept wordt een toetsingsprogramma opgevat als een systeem met een stelsel van subsystemen. Onder de subsystemen kunnen alle elementen van een examenprogramma worden verstaan: de verschillende toetsmethoden (bijvoorbeeld

schriftelijk of mondeling), de aard van de toets (bijvoorbeeld de gerichtheid op specialistische feiten of op grotere gehelen), het tijdstip van toetsing (begin, eind, concurrentie met andere examens, etcetera) en - zeer belangrijk - het hele stelsel van regels en richtlijnen rondom het toetsingsprogramma (bijvoorbeeld de zwaarte van de onderdelen). Van wezenlijk belang bij deze systeemnotie is het besef dat een verandering in één subsysteem veranderingen in andere delen teweeg kan (en zal) brengen. Het (studeer-) gedrag van de student vormt voortdurend de afhankelijkke variabele.

Op deze wijze kunnen toetsen worden gebruikt om gewenste veranderingen in het leerproces te bevorderen. Sommige auteurs gaan nog een stapje verder en stellen onderwijsveranderingen voor waarbij uitsluitend het toetsprogramma gewijzigd wordt.⁵⁴

Uit het onderzoek naar de toetsing van medische competentie is, zoals hierboven is gebleken, een aantal duidelijke, consistente bevindingen naar voren gekomen. Deze 'regelmatigheden' zouden we tentatief in enkele 'wetmatigheden' kunnen uitdrukken.⁵⁵ Drie wetmatigheden vloeien rechtstreeks uit het bovenstaande voort. Enkele praktische suggesties zullen eraan worden verbonden.

ENKELE WETMATIGHEDEN EN SUGGESTIES

De 'Black-Box' wet

Ongeacht welk acroniem door de instrument-ontwikkelaars gehanteerd wordt:

OSCE PMP MCQ MEQ CPMP SMP PBLM
CASE MONDELING SHORT-CASE
LONG-CASE MSE SIMP Q4 SPT TJE TFO
VGT etcetera.....

en wat men ook maar wil meten in de 'Black Box':

clinical reasoning, problem-solving, diagnostic skill, doctor patient relationship, cognitive skills, attitudes, history-taking, data-gathering, medical expertise, etcetera

1) *bedenk dat er waarschijnlijk grotere variaties binnen methoden bestaan dan tussen methoden, en*

2) *bedenk dat wat er gemeten wordt waarschijnlijk sterker wordt bepaald door de specifieke inhoud van de meting, dan door de eigenschappen van de methode.*

Kortom, de methode zelf is niet belangrijk, wel de inhoud. De opdrachten die de kandidaat moet uitvoeren, bepalen wat er gemeten wordt, waarbij de methode in principe ondergeschikt is. Dezelfde taken kunnen vaak met behulp van meerdere methoden worden gepresenteerd. De variatie tussen de resultaten die met één methode verkregen worden, is daarentegen groot: een score op de ene taak zegt weinig over de score op een andere taak. Een toets moet dus uit een groot aantal opdrachten bestaan.

Een tweetal praktische suggesties kan hieraan worden verbonden. In de eerste plaats: wees niet getrouwd met één bepaalde toetsmethode. Doorgaans zijn verschillende methoden noodzakelijk voor een goede evaluatie en - aangezien de inhoud belangrijker is dan de methode zelf - *maak de methode afhankelijk van de inhoud* en niet andersom. Bijvoorbeeld, het is niet goed mogelijk communicatieve vaardigheden schriftelijk te meten en dus ligt een observatiemethode meer voor de hand. Het is echter zeer moeilijk gedetailleerde criteria-lijsten voor een observatie te maken zonder dat de inhoud triviaal wordt. In zo'n geval zijn meer globale beoordelingsvormen te prefereren. In een taak waarbij van studenten wordt verlangd om hypothesen te genereren, zouden bij gebruik van multiple-choice vragen cueing effecten kunnen optreden (herkenning). Open vragen zijn hier dus geschikter. Ook binnen vraagvormen dient de inhoud voorop te staan. Wie verlangt eigenlijk dat bij elke vraag van een multiple-choice toets uit vier alternatieven moet worden gekozen? Met een dergelijk harnas is het geen wonder dat het zo moeilijk is zinvolle alternatieven te vinden. Waarom zou het aantal alternatieven geen rechtstreekse functie kunnen zijn van het aantal zinvolle mogelijkheden?

Een tweede suggestie betreft de samenstelling van een toets. Wanneer een groot aantal taken noodzakelijk is, dient men zich te realiseren dat een toets al snel een zeer behoorlijke omvang moet hebben. Een en ander is natuurlijk geheel afhankelijk van de grootte en de heterogeniteit van het te meten domein, maar de praktijk leert dat enkele uren toetstijd een vrij normale zaak zou moeten zijn.

Wet van het behoud van energie

Het rendement van een toetsmethode is gelijk aan de tijd die men eraan besteedt.

Het maken van goede toetsen kost veel tijd, het maken van slechte toetsen weinig tijd. Een goede toets wordt gekenmerkt door een hoge kwaliteit van de gestelde taken (inhoud, vorm, relevantie), door standaardisatie en structurering, en door een adequate steekproef uit alle taken die mogelijk zijn. Het streven naar hoge kwaliteit is zeer tijdrovend. En ook hier geldt weer dat de toetsmethode op zich niet zaligmakend is. Een goede multiple-choice toets is te prefereren boven een slechte openvragen toets; een stationsexamen is niet per definitie goed omdat het een stationsexamen is. De tijdsinvestering levert echter winst op: de kwaliteit wordt aanzienlijk verhoogd als men kritisch naar de inhoud kijkt, enige structuur aanbrengt, meer standaardisatie toepast, een blauwdruk ontwerpt, gebruik maakt van statistische gegevens van voorgaande toetsen, etcetera. Het rendement daarvan wordt nog eens onderstreept als we ons bewust zijn van de betekenis van toetsen in het totale leerproces van studenten (figuur 2).

Wet van oorzaak en gevolg:

Elke evaluatie roept een (evenredige of grotere en soms zelfs tegengestelde) onderwijskundige reactie op.

In het licht van de hierboven beschreven relatie tussen toetsen en het studiegedrag van studenten behoeft deze wetmatigheid geen verdere uitleg. Van belang is echter dat men bij de planning of verandering van een toetsingsprogramma voortdurend rekening moet houden met onderwijskundige neven-effecten en, een stap verder, dat men deze wetmatigheid doeltreffend kan gebruiken om onderwijskundige veranderingen te bewerkstelligen.

DE UITDAGING

Het is duidelijk dat in de afgelopen decennia de kennis over toetsing in het medisch onderwijs aanzienlijk is uitgebreid. Uit onderzoek en ervaringen zijn consistente bevindingen naar voren gekomen. Een aantal intuïtieve veronderstellingen blijkt bij empirische toetsing onjuist te zijn, terwijl andere juist zijn onderbouwd. Kortom wij beschikken over een ruime know-how op het gebied van toetsing van medische competentie. Het is aan de onderwijsmakers en toetsconstructeurs om deze kennis in praktijk te brengen.

LITERATUUR

1. Rethans J, Van Leeuwen YD, Drop M, Van der Vleuten CPM, Sturmans F. Competence and performance: two different constructs in the assessment of quality of medical care. *Family Practice* 1990; 7: 168-74.
2. 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: Heal-Publications, 1987; 59-75.
3. De Graaff E, Post G, Drop M. Validation of a new measure of clinical problem-solving. *Medical Education* 1987; 21: 213-8.
4. Friedman R et al. Experience with the simulated patient physician encounter. *J Med Educ* 1978; 53: 825-30.
5. McLeskey C, Ward R. Validity of written examinations. *Anesthesiology* 1978; 49: 224.
6. Grant J, Marsden P. Primary knowledge, medical education, and consultant expertise. *Medical Education* 1988; 22: 746-53.
7. Schmidt H, Boshuizen H, Hobus P. Transitory stages in the development of medical expertise: the 'intermediate effect' in clinical case representation studies.

- Proceedings 10th Conference Cognitive Science Society. Hillsdale, New Jersey: Erlbaum, 1988; 139-45.
8. Norman GR, Allery L, Berkson L et al. Research in the psychology of clinical reasoning: implications for assessment. In: Jolly B, ed. *New directions in the assessment of clinical competence*. Proceedings of Cambridge Conference IV. Cambridge: Madingley Hall. (Ter perse)
 9. Schmidt H, Norman GR, Boshuizen H. A cognitive perspective on medical expertise: theory and implications. *Academic Medicine* 1990; 65: 611-21.
 10. Norman GR. Reliability and construct validity of some cognitive measures of clinical reasoning. *Teaching and Learning in Medicine* 1989; 1: 194-9.
 11. Swanson D, Case S, Van der Vleuten CPM. Student assessment in problem-based learning curricula. In: Boud D, Feletti G, eds. *Using problem-based learning*. London: Kogan Page. (Ter perse)
 12. Van der Drift J, Vos P. Anatomie van een leeromgeving: een onderwijsseconomische analyse van het universitaire onderwijs. Lisse: Swets en Zeitlinger, 1987.
 13. Wijnen WHFW, Van der Vleuten CPM. Toetsing: hordenloop of voortgangscontrole? Universiteit en

Hogeschool 1985; 31: 270-9.

14. Van der Vleuten CPM, Van Luijk SJ. Ontwikkelingen in de toetsing van praktische vaardigheden. *Onderzoek van Onderwijs* 1990; 19: 24-7.

15. Newble D, Jaeger K. The effect of assessments and examinations on the learning of medical students. *Medical Education* 1983; 17: 165-71.

16. Newble D. Eight years' experience with a structured clinical examination. *Medical Education* 1988; 22: 200-4.

17. Van der Vleuten CPM, Swanson D. Assessment of clinical skills with standardized patients: state of the art. *Teaching and Learning in Medicine* 1990; 2: 58-76.

18. Elstein A, Shulman L, Sprafka S. *Medical problem solving: an analysis of clinical reasoning*. Cambridge, Massachusetts: Harvard University Press, 1978.

19. Swanson D, Norcini JJ, Grosso L. Assessment of clinical competence: written and computer-based simulations. *Assessment and Evaluation in Higher Education* 1987; 12: 220-46.

20. Norcini JJ, Meskauskas J, Langdon L, Webster G. An evaluation of a computer simulation in the assessment of physician competence. *Evaluation in the Health Professions* 1986; 9: 286-304.

21. Norcini JJ, Swanson D. Factors influencing testing time requirements for simulation-based measurements: do simulations ever yield reliable scores? *Teaching and Learning in Medicine* 1989; 1: 85-91.

22. Swanson D. A Measurement framework for performance-based tests, 1987. In: Hart IR, Harden RM, eds. *Newer developments in assessing clinical competence*. Montreal: Heal Publications, 1987; 13-45.

23. Erviti V, Templeton B, Bunce J, Burg F. The relationships of pediatric resident recording behavior across medical conditions. *Medical Care* 1980; 18: 1020-31.

24. Williams R, Barrows H, Vu N et al. Direct, standardized assessment of clinical competence. *Medical Education* 1987; 21: 482-9.

25. Norman GR, Smith E, Powles A, Rooney P, Henry N, Dodd P. Factors underlying performance on written tests of knowledge. *Medical Education* 1987; 21: 297-304.

26. Stalenhoef-Halling BF, Van der Vleuten CPM, Jaspers T, Fiolet J. The feasibility, acceptability and reliability of open-ended questions in a problem-based learning curriculum. In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra RP, eds. *Teaching and assessing clinical competence*. Groningen: BoekWerk Publ, 1990; 552-7.

27. Norcini JJ, Swanson D, Grosso L, Shea J, Webster G. Reliability, validity and efficiency of multiple-choice question and patient management problem item formats in the assessment of physician competence. *Medical Education* 1985; 19: 238-47.

28. Maatsch J. Model for a criterion-referenced medical specialty test. Office of medical Education Research and Development Michigan State University, 1980. Final Report Grant No. HS-02038-02.

29. Maatsch J, Huang R. An evaluation of the construct validity of four alternative theories of clinical competence. *Proceedings of the Twenty-fifth Annual Conference on Research in Medical Education*. Washington DC, 1986.

30. Maatsch J. Theories of clinical competence: the construct validity of objective tests and performance assessments. Paper presented at the International Conference on Evaluation in Medical Education. Beer Sheva: Israel, 1987.

31. Van der Vleuten CPM, Van Luijk SJ, Beckers H. A written test as an alternative to performance testing. *Medical Education* 1989; 23: 97-107.

32. Van Luijk SJ, Van der Vleuten CPM, Van Schelven RM. The relation between content and psychometric characteristics in performance-based testing. In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra RP, eds. *Teaching and assessing clinical competence*. Groningen: BoekWerk Publ, 1990; 202-7.

33. McGuire C. Written methods for assessing clinical competence. In: Hart IR, Harden RM, eds. *Further developments in assessing clinical competence*. Montreal: Heal-Publications, 1987; 46-58.

34. Streiner D. Global rating scales. In: Neufeld V, Norman GR, eds. *Assessing clinical competence*. New York: Springer, 1985; 119-41.

35. Neufeld V. Written tests. In: Neufeld V, Norman GR, eds. *Assessing clinical competence*. New York: Springer, 1985; P4-118.

36. Muzzin L, Hart L. Oral examinations. In: Neufeld V, Norman GR, eds. *Assessing clinical competence*. New York: Springer, 1985; 71-93.

37. Frijns P, Van der Vleuten CPM, Verwijnen GM, Van Leeuwen YD. The effect of structure in scoring methods on the reproducibility of tests using open-ended questions. In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra RP, eds. *Teaching and assessing clinical competence*. Groningen: BoekWerk Publ, 1990; 461-71.

38. De Gruijter D. Beoordelen met open vragen. In: Van Berkel H, Bax A, eds. *Beoordelen in het onderwijs*. Almere: Versluys, 1990; 34-41.

39. Van Berkel H, Bax A. Beoordelen met een mondelinge toets. In: Van Berkel H, Bax A, eds. *Beoordelen in het onderwijs*. Almere: Versluys, 1990; 49-55.

40. Norman GR, Van der Vleuten CPM, De Graaff E. Pitfalls in the pursuit of objectivity: issues of validity. *Medical Education* 1991; 25(2): 119-26.

41. Van Luijk SJ, Van der Vleuten CPM. A comparison of checklists and rating scales in performance-based testing. In: Hart IR, Harden RM, eds. *More developments in assessing clinical competence*. (Ter perse)

42. Van Luijk SJ, Van der Vleuten CPM, Van Schelven RM. Opinions about performance-based tests. In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra RP, eds. *Teaching and assessing clinical competence*. Groningen: BoekWerk Publ, 1990; 497-502.

43. Van der Vleuten CPM, Norman GR, De Graaff E. Pitfalls in the pursuit of objectivity: issues of reliability. *Medical Education* 1991; 25(2): 110-8.

44. Van der Vleuten CPM, Wijnen WHFW. Niets praktischer dan een goede theorie: generaliseerbaarheidstheorie als instrument voor betrouwbaarheidsstudies. *Bulletin Medisch Onderwijs* 1991; 10(1): 2-13.

45. Verwijnen GM, Van der Vleuten CPM, Imbos TJ. Comparing an innovative medical school with traditional schools: an output analysis in the cognitive domain. In: Khattab T, Schmidt H, Nooman Z, Ezzat E, eds. *Innovation in medical education: an evaluation of its present status*. New York: Springer Publishing Company, 1990; 40-49.

46. De Groot A. Is de kwaliteit van het onderwijs te beoordelen? In: Creemers B, Hoebe W, Koops K, eds. *De kwaliteit van het onderwijs*. Groningen: Wolters-Noordhoff, 1983.

47. Ebbinghaus, geciteerd in Klausmeier H, Ripple R. *Learning and human abilities*. New York: Harper & Row, 1971.

48. Besseling C, Tromp J, Van der Vleuten CPM, Verbraeck A. De voortgangstoets in het hoger onderwijs. Tijdschrift voor Hoger Onderwijs 1986; 4: 107-15.

49. Frederiksen N. The real test bias: influences of testing on teaching and learning. American Psychologist 1984; 39: 193-202.

50. Entwistle N. Styles of learning and teaching. Chichester: John Wiley & Sons, 1981.

51. Stillman P, Swanson D. Ensuring the clinical competence of medical school graduates through standardized patients. Archives of Internal Medicine 1987; 147: 1049-52.

52. Bouhuijs P, Van der Vleuten CPM, Van Luyk SJ. The OSCE as a part of a systematic skills training approach. Medical Teacher 1987; 9: 183-91.

53. Van der Vleuten CPM, Verwijnen GM. Assessment in problem-based learning. In: Van der Vleuten CPM, Wijnen WHFW, eds. Problem-based learning: perspectives from the Maastricht approach. Amsterdam: Thesis-publ, 1990; 27-50.

54. Stillman P. Let the tail wag the dog. Ongepubliceerde bijdrage aan Cambridge Conference I. Cambridge, 1986.

55. Norman GR, persoonlijke communicatie.

AANBEVOLEN LITERATUUR BIJ FIGUUR 1.

Toetsing van Medische Competentie - Algemeen: Neufeld V, Norman GR, eds. Assessing clinical competence. New York, Springer, 1985. / Hart IR, Harden RM, Walton H, eds. Newer developments in assessing clinical competence. Montreal: Heal Publications, 1987. / Hart IR, Harden RM, eds. Further developments in assessing clinical competence. Montreal: Can-Heal, 1987. / Bender W, Hiemstra RJ, Scherpier AJA, Zwierstra RP, eds. Teaching and assessing clinical competence. Groningen: BoekWerk Publ, 1990.

Schriftelijke methoden - Algemeen: McGuire C. Written methods for assessing clinical competence. In: Hart IR, Harden RM, eds. Further developments in assessing clinical competence. Montreal: Heal-Publications, 1987; 46-58.

GOV: Knox J. How to use modified essay questions. Medical Teacher 1980; 2: 20-4.

PMP: McGuire C, Solomon C. Construction and use of written simulations. Chicago: The Psychological Corporation, 1976.

MEQ: Feletti G, Engel C. The modified essay question for testing problem-solving skills. Medical Journal of Australia 1980; 1: 79-80.

P4-deck: Barrows H, Tamblyn R. The portable patient problem pack (P4). A problem-based learning unit. J Med Educ 1977; 52: 1002-4.

CBX: Norcini JJ, Meskuskas J, Langdon L, Webster G. An evaluation of a computer simulation in the assessment of physician competence. Evaluation in the Health Professions 1986; 9: 286-304.

CRT: Williams R, Vu NV, Barrows HS, Verhulst S. Profile of the Clinical Reasoning Test (CRT): an objective measure of problem solving skills and proficiency in using medical knowledge. In: Schmidt H, De Volder M, eds. Tutorials in problem-based learning. Assen: Van Gorcum, 1984.

Key feature approach: Norman GR, Bordage G, Curry L, Dauphinee D, Jolly B, Newble DI, Rothman A, Stalenhoef-Halling BF, Stillman P, Swanson DB, Tonesk X. A review of recent innovations in assessment. In: Wakeford R, Bashook P, Jolly B, eds. Directions in clinical assessment. Cambridge: Office of the Regius Professor Of Physics Cambridge University School of Clinical Medicine, 1985.

Q4: 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: Heal-Publications, 1987; 59-75.

SIMP: De Graaff E, Post G, Drop M. Validation of a

new measure of clinical problem-solving. Medical Education 1987; 21: 213-8.

Pattern recognition: Case S, Swanson D, Stillman P. Evaluating diagnostic pattern recognition: the psychometric characteristics of a new item format. Proceedings of the Twenty-seventh Annual Conference on Research in Medical Education (RIME). Chicago, USA, 1988.

DWP: Van Rossum HJM, Briët E, Bender W, Meinders AE. The transfer effect of one single patient demonstration on diagnostic judgment of medical students: both better and worse. In: Bender W, Hiemstra RJ, Scherpier AJA, Zwierstra RP, eds. Teaching and assessing clinical competence. Groningen: BoekWerk Publ, 1990; 435-40.

Cognitief (psychologische) metingen: Norman GR. Reliability and construct validity of some cognitive measures of clinical reasoning. Teaching and Learning in Medicine 1989; 1: 194-9.

Triple jump exercise: Powles A, Wintrup N, Neufeld V, Wakefield J, Coates G, Burrows J. The triple jump exercise: further studies of an evaluative technique. Proceedings of the 20th Annual Conference on Research in Medical Education. Washington: American Association of Medical Colleges, 1981.

Gestructureerde mondelinge examens: Grava-Gubins I, Khan S, Rainsberry P. Factor analysis of simulated office oral examinations in family medicine. In: Hart IR, Harden RM, eds. Further developments in assessing clinical competence. Montreal: Heal Publications, 1987; 406-17.

OSCE: Harden RM, Gleeson F. Assessment of clinical competence using an objective structured clinical examination (OSCE). Medical Education 1979; 13: 41-54.

Standardized patient-based testing: Stillman P, Ruggill J, Rutala P, Sabers D. Patient instructors as teachers and evaluators. J Med Educ 1980; 55: 186-93. / Van der Vleuten CPM, Swanson D. Assessment of clinical skills with standardized patients: state of the art. Teaching and Learning in Medicine 1990; 2: 58-76.

Effecten van toetsing op het leergedrag: Newble D, Jaeger K. The effect of assessments and examinations on the learning of medical students. Medical Education 1983; 17: 165-71. / Frederiksen N. The real test bias: influences of testing on teaching and learning. American Psychologist 1984; 39: 193-202.

Voortgangstoetsen/toetsing als systeemconcept: Van der Vleuten CPM, Verwijnen GM. A system for student assessment. In: Van der Vleuten CPM, Wijnen WHFW, eds. Perspectives from the Maastricht experience. Amsterdam: Thesis-publ, 1990; 27-50.