

## Toetsen van vaardigheden door middel van het stationsexamen

S.J. van Luijk

### Samenvatting

*Het toetsen door middel van stationsexamens heeft de laatste vijftien jaar een grote vlucht genomen. De basis van deze ontwikkelingen ligt bij twee toetsvormen: de 'Objective Structured Clinical Examination' en de 'Standardised Patient-based Test'. Inmiddels zijn er vele tussenvormen ontwikkeld. Dit artikel bevat een overzicht van de belangrijkste gegevens over stationsexamens. Informatie wordt verstrekt met betrekking tot belangrijke betrokkenen bij het stationsexamen: beoordelaars, simulatiepatiënten, studenten, en docent en onderwijsorganisatie. Daarna wordt ingegaan op de betrouwbaarheid en validiteit van het stationsexamen en worden mogelijkheden tot verbetering van stationsexamens beschreven. Het artikel besluit met praktische suggesties voor diegenen die betrokken zijn bij organisatie, afname en onderzoek rondom stationsexamens.*

### Inleiding

Aan medische faculteiten in Nederland worden sinds de jaren tachtig stationsexamens afgenomen. Een stationsexamen is een toetsvorm waarbij studenten bij het uitvoeren van handelingen worden geobserveerd aan de hand van voorgestructureerde beoordelingslijsten die door docenten worden ingevuld. Er worden globaal twee vormen onderscheiden: de Objective Structured Clinical Examination en de Standardized Patient-based Test.

Tot laat in de jaren zeventig gold voor vrijwel alle instellingen van medisch onderwijs dat toekomstige artsen vrijwel uitsluitend getoetst werden op kennisbeheersing. Het toetsen van medische competentie en vaardigheden als

onderdeel van het medisch handelen gebeurde op ad hoc basis tijdens de stages, bij zich toevallig aandienende patiënten, door een beoordelaar met eigen impliciete criteria. Deze waarneming was voor Harden en Gleeson aanleiding om te zoeken naar een betrouwbare methode om medische competentie te meten die bovendien praktisch toepasbaar was in de dagelijkse routine van de co-assistenten schappen. Zij ontwikkelden als toetsvorm de *Objective Structured Clinical Examination* (OSCE).<sup>1</sup>

In een OSCE moet een student een aantal opdrachten uitvoeren. Sommige opdrachten worden geobserveerd door een beoordelaar, andere opdrachten moeten schriftelijk afgehandeld worden. Alle opdrachten zijn van beperkte tijdsduur (maximaal 4 à 5 minuten). Soms zijn opdrachten inhoudelijk aan elkaar gerelateerd (tabel 1), maar ook andere varianten zijn mogelijk (tabel 2). Elk onderdeel wordt in een aparte ruimte uitgevoerd, zodat meerdere studenten na elkaar over dezelfde patiëntencasus kunnen worden getoetst. Deze aparte ruimten worden 'stations' genoemd. Bij een dergelijk stationsexamen kunnen zowel kennis als vaardigheidselementen aan bod komen. Voorts kunnen zowel het proces als het product van het handelen getoetst worden.

Aan de hand van voorgestructureerde beoordelingslijsten kunnen beoordelaars onafhankelijk van elkaar op eenduidige wijze sterke en zwakke kanten van studenten vaststellen. Deze informatie kan aangewend worden om examenbeslissingen te nemen, maar ook als educatieve feedback voor de student en de onderwijsinstelling. Het gebruik van OSCE's voor het toetsen van medische competentie heeft wereldwijd een grote vlucht genomen,

**Tabel 1.** Voorbeeld van een circuit van OSCE-stations waarbij alle stations inhoudelijk aan elkaar gerelateerd zijn

## Beschrijving

voer een gesprek met deze patiënt met maagpijn\*  
 vat de belangrijkste elementen van het gesprek samen \*\*  
 voer het buikonderzoek uit \*  
 beschrijf de bevindingen bij lichamelijk onderzoek \*\*  
 welk aanvullend onderzoek zou verricht moeten worden \*\*  
 beoordeel de laboratoriumuitslagen van de patiënt \*\*  
 beoordeel de X-foto \*\*  
 geef aan welk beleid op dit moment het meest zinvol is \*\*  
 voer een gesprek met de patiënt en deel hem mede dat hij een ulcus ventriculi heeft \*

Alle stations duren 4-5 minuten; Bij de met \* gemerkte stations is een observator aanwezig; De met \*\* gemerkte stations worden schriftelijk afgenomen

vooral omdat zij, vergeleken met kennistoetsing, een hoog realiteitsgehalte bezitten.

Ongeveer tegelijkertijd met de OSCE ontwikkelde zich ook een andere vorm van competentiemeting als een variant op de OSCE.<sup>3</sup> Deze *Standardized Patient-based Test* (SP-based test) is evenals de OSCE een stationsexamen, dat wil zeggen dat studenten opdrachten moeten uitvoeren in afzonderlijke ruimten (stations), waarbij zij beoordeeld worden aan de hand van voorgestructureerde lijsten. Een belangrijk verschilpunt is echter dat bij de SP-based tests simulatiepatiënten ingezet worden voor het beoordelen van studenten. Simulatiepatiënten worden specifiek getraind voor die taak. Een ander verschilpunt betreft de omvang van de opdrachten. Bij de OSCE wordt een patiëntencasus vaak opgedeeld in een grote reeks gedetailleerde deelopdrachten van zeer korte duur. Bij de SP-based tests hebben opdrachten over het algemeen een meer globaal karakter en duren langer (15 tot 45 minuten) waarbij grotere delen van het medisch consult aan bod komen. Het betreft veelal ook inhoudelijk aparte entiteiten (tabel 3). De langere tijdsduur bij stations van de SP-based tests leidt in het algemeen ook tot veel langere toetstijden (2 tot 6 uur) dan bij de OSCE.

Deze verschillen weerspiegelen zich ook in de beoordelingslijsten. De beoordelingslijsten van de OSCE bevatten voornamelijk gedetailleerde items, die veelal op het proces van de handeling gericht zijn. De beoordelingslijsten van de SP-based tests zijn over het algemeen wat globaler van aard en zijn vaker productmatig, dus op uitkomsten van handelingen gericht.

### Ontwikkeling in de afgelopen vijftien jaar

De klassieke OSCE en de SP-based tests zijn als het ware twee uitersten op een schaal. Er zijn de afgelopen jaren veel variaties ontwikkeld; de term stationsexamen beschrijft dan ook alleen maar het organisatiemodel van de toets. Er zijn bijvoorbeeld verschillen wat betreft:

#### a. het doel van de meting:

- toetsing van tussentijds gegeven onderwijs
- einddoelgerichte toetsing

#### b. de uitvoering van de meting:

- het aantal stations
- de stationsduur
- mate van detaillering van beoordelingslijsten

**Tabel 2.** Voorbeeld van een circuit van OSCE-stations waarbij vaardigheden aan bod komen die niet noodzakelijkerwijs aan elkaar gerelateerd zijn

Stationsbeschrijving
voer een gesprek met een patiënt die kortademig is*
beantwoord vragen naar aanleiding van dit station**
voer de reflexen uit van de bovenste extremiteiten*
beantwoord vragen naar aanleiding van dit station**
beoordeel een röntgenfoto (X-thorax)**
beantwoord vragen naar aanleiding van dit station**
bepaal de hoeveelheid eiwit en glucose in de urine**
beantwoord vragen naar aanleiding van dit station**
beoordeel een electrocardiogram
beantwoord vragen naar aanleiding van dit station**
voer een gesprek met een patiënt met klachten over lage rugpijn
beantwoord vragen naar aanleiding van dit station**
beoordeel het netvlies van de patiënt aan de hand van een dia
beantwoord vragen naar aanleiding van dit station**

Alle stations duren 4-5 minuten; Bij de met \* gemerkte stations is een observator aanwezig; De met \*\* gemerkte stations worden schriftelijk afgenomen

- de verhouding tussen proces- en product-items
- specificiteit van beoordelingsschalen
- inbedding van de toets in het geheel van het facultaire onderwijs en evaluatieprogramma, dat wil zeggen:
- frequentie van toetsafname
- aansluiting van de toets op het onderwijs
- zorg besteed aan kwaliteitsbewaking van de toets
- betrokkenheid facultaire medewerkers bij constructie en afname

*c. de 'status' van de toets:*

- formatief of summatief

Naast verschillen bestaan er ook overeenkomsten tussen de diverse stationsexamens. Zo is men grotendeels teruggekomen van het klassieke OSCE-model waarbij studenten zich van het ene naar het andere station spoeden en vervolgens onderdelen van vaardigheden moesten demonstreren die geheel los staan van enige

**Tabel 3.** Voorbeeld van problemen aangeboden in een standardized patient-based test

Stationsbeschrijving
pijn op de borst *
vragen naar aanleiding van het station**
bloed bij de urine*
vragen naar aanleiding van het station**
kortademigheid*
vragen naar aanleiding van dit station**
intermenstrueel bloedverlies*
vragen naar aanleiding van dit station**
pijn onder in de buik*
vragen naar aanleiding van dit station**
slechthorendheid*
vragen naar aanleiding van dit station**
duizeligheid*
vragen naar aanleiding van dit station**
depressiviteit*
vragen naar aanleiding van dit station**

Alle stations duren 20 minuten; Bij de met \* gemerkte stations is een simulatiepatiënt aanwezig; De met \*\* gemerkte stations worden schriftelijk afgenomen

context. Een andere trend is de toename van de tijdsduur per station. Gemiddeld genomen ligt deze nu tussen 10 tot 15 minuten per station met uitschieters naar 30 minuten.

Een laatste ontwikkeling is dat items minder gedetailleerd en meer globaal geformuleerd worden. Hierdoor wordt het voor studenten minder efficiënt om criterialijsten uit het hoofd te leren. Bovendien wordt een globalere beoordelingsschaal door beoordelaars meer gewaardeerd. Het stationsexamen als organisatie-model wordt ook toegepast bij nascholing, waarbij al of niet wordt getoetst.<sup>4</sup>

### De beoordelaar

Beoordelaars bij OSCE's worden ingezet om observeerbare gesprekstechnische en/of fysisch-diagnostische vaardigheden te beoordelen. Beoordelaars bij SP-based tests beoordelen meestal uitsluitend het gemaakte 'schriftelijke

materiaal' van de student. De simulatiepatiënt beoordeelt de handelingsvaardigheden van de student. Bij de inzet van beoordelaars bij stationsexamens zijn verschillende overwegingen van belang.

#### *Wie moet beoordelen?*

Het ligt voor de hand dat degenen die het onderwijs verzorgen ook degenen zijn die de studenten beoordelen en op deze wijze toetsen of het gegeven onderwijs effectief is geweest. Anderzijds kan men stellen dat degenen die het onderwijs verzorgen juist niet betrokken moeten worden bij de toetsing om een eventuele bias ten aanzien van studenten of het eigen functioneren te voorkomen. De meest praktische oplossing is dat degenen die het onderwijs verzorgen samen met andere beoordelaars in de toets participeren.

Regelmatig wordt de vraag gesteld hoe inhoudsdeskundig een beoordelaar dient te zijn. De ervaring leert dat dit afhangt van de detaillering van de criteria op de beoordelingslijst. Hoe gedetailleerder de criteria worden weergegeven, des te minder hoeft er een beroep te worden gedaan op de specifieke inhoudsdeskundigheid van de beoordelaar. Globale criteria vereisen een grotere mate van inhoudsdeskundigheid. Een en ander hangt natuurlijk ook samen met de mogelijkheid beoordelaars vooraf te trainen, dan wel de ervaring die de beoordelaar al bij andere stationsexamens heeft opgedaan.

#### *Heeft trainen van beoordelaars zin?*

Training van beoordelaars gericht op de inhoud van het te toetsen domein heeft zin, hoewel het geen spectaculaire verbeteringen in de kwaliteit van de observatie oplevert. Een training gericht op de inhoud van een station is vooral nuttig voor beoordelaars die niet inhoudsdeskundig zijn op het te toetsen domein.<sup>6</sup> Uit ervaring blijkt dat dergelijke trainingen het meest efficiënt verlopen als de te trainen vaardigheid gedemonstreerd wordt en de beoorde-

laar aan de hand van de demonstratie de beoordelingslijst moet invullen. Vervolgens worden de beoordelingen nabesproken. Een alternatief is dat nieuwe beoordelaars eerst ingezet worden als co-beoordelaars en hun beoordelingen met een ervaren beoordelaar nabespreken. Een voordeel van deze laatste methode is dat men tegelijkertijd ervaring opdoet met de procedures van het stationsexamen. Dit laatste is belangrijk omdat blijkt dat er juist in dat opzicht veel misgaat (te laat komen, studenten te lang in de kamer laten zitten, opdrachten niet, of op de verkeerde momenten, aan studenten overhandigen, er een eigen scoringsmethode op na houden, zich laten oppiepen tijdens de toets en dergelijke).

#### *Hoe betrouwbaar kunnen beoordelaars beoordelen?*

Vastgesteld is dat de interbeoordelaarsbetrouwbaarheid bij stationsexamens over het algemeen voldoende hoog is ( $\geq 80$ ).<sup>7</sup> Beoordelaars kunnen betrouwbaar beoordelen. Er zit echter wel een adder onder het gras. Studenten hebben weinig boodschap aan hoge betrouwbaarheidsindices. Voor hen wordt de kwaliteit van de beoordeling veel meer bepaald door de attitude van de beoordelaar, tot uiting komend in de bejegening van de student tijdens de toets. Beoordelaars die in hun bejegening van studenten blijf geven van een extreem ongeschikte attitude - bijvoorbeeld door ze geen hand te geven, de student niet aan te kijken, nauwelijks iets te zeggen, aanvullende (voor het examen irrelevante) vragen te stellen - kunnen weliswaar soms betrouwbaar beoordelen, maar door hun houding ook de prestatie van de student sterk negatief beïnvloeden. Uiteraard is dit effect sterker naarmate de student onzekerder is en/of de toets als bijzonder stressvol ervaart.

Ook simulatiepatiënten kunnen een rol spelen bij de beoordeling van studenten. Hierop wordt ingegaan in de paragraaf over simulatiepatiënten.

*Hoe lang kan een beoordelaar geconcentreerd beoordelen?*

De tijd gedurende welke een beoordelaar wordt ingezet bij een stationsexamen varieert. Er bestaan bij de stationsexamens, ongeacht het aantal observaties, grote verschillen in observatietijd tussen beoordelaars (2 tot 7 uur). Lange observatietijden komen voor en het effect daarvan op het gedrag van de beoordelaar en de wijze van scoring is onbekend.

Vanuit de praktijk wordt gemeld dat kortdurende stations - waarbij de beoordelaar veel studenten ziet per tijdseenheid en waarbij beoordeeld moet worden aan de hand van gedetailleerde criterialijsten - aanzienlijk meer energie en concentratie van de beoordelaar vergen dan stations die lang duren en waarbij de criteria globaler zijn geformuleerd. Dit geldt in nog sterkere mate indien de beoordelaar bij de kortdurende en gedetailleerde stations ook nog materialen moet klaarzetten of opruimen.

*Moet een beoordelaar feedback geven tijdens de toets?*

Hoewel strijdig met de ideeën die stafleden hierover hebben, blijkt systematisch dat studenten feedback van beoordelaars tijdens de toets erg prettig vinden.<sup>8</sup> Een belangrijke verklaring hiervoor kan zijn dat het voor studenten vaak de enige mogelijkheid is om discrepanties tussen de verwachtingen over het eigen presteren en de uiteindelijke beoordeling te bespreken. Voor de beoordelaar kan een dergelijke nabespreking ook nuttig zijn. Zeker als beoordelingslijsten pas worden ingevuld na het demonstreren van de vaardigheid, komt het vaker voor dat een beoordelaar niet exact meer weet of een bepaald item al of niet is uitgevoerd. Helaas blijkt in de praktijk dat er vaak geen tijd is voor een nabespreking. Het is zinvol hier specifiek tijd voor te reserveren tijdens de toets.

*Wat is het nut van co-beoordelaars?*

In een aantal stationsexamens worden ook co-beoordelaars ingezet. De feitelijk bedoeling hiervan is om de betrouwbaarheid van de observatie te verhogen. Dit biedt de mogelijkheid om uitspraken te doen over de interbeoordelaarsbetrouwbaarheid.

In de praktijk blijkt echter dat door de complexe logistiek van stationsexamens vaak reservebeoordelaars nodig zijn om ad hoc te kunnen vervangen of tijdelijk waar te nemen. Co-observatoren kunnen daarvoor ingezet worden. De praktijk leert dat de reservefunctie van de co-observator de belangrijkste is. Een ander voordeel van co-beoordelaars is dat dubbele waarnemingen nuttig kunnen zijn in het geval van ernstige onenigheid tussen student en observator. Voorts heeft het inzetten van co-beoordelaars mogelijk een preventief (positief) effect op het gedrag van de reguliere beoordelaars. Vooraf dient duidelijk afgesproken te worden met de (co)observatoren wie de student feitelijk beoordeelt.

Studenten reageren verdeeld op de aanwezigheid van co-beoordelaars. De aanwezigheid van co-beoordelaars bij toetsen met complexe stations wordt meer gewaardeerd dan bij stationsexamens waarbij de nadruk ligt op eenvoudige technische handelingen.<sup>8</sup>

*Welke knelpunten ervaren beoordelaars bij stationsexamens?*

Beoordelaars worden ingezet om studenten te beoordelen aan de hand van voorgestructureerde lijsten. Dat aspect van de beoordelaarsrol is duidelijk. Minder helder is hoe een dergelijke taak precies moet worden vormgegeven.<sup>9</sup> Moet een beoordelaar zich actief of terughoudend opstellen tijdens de toets? Moet ingegrepen worden indien de student op het verkeerde spoor zit? Welke mate van sturing is toegestaan gedurende het examen? Kan commentaar of feedback gegeven worden terwijl de student de opdracht uitvoert? Regelgeving is sterk situationeel bepaald en derhalve bijzonder lastig in

algemene adviezen te verwoorden. Gelet op de kritiek van studenten dienen observatoren zich eerder te richten op het geruststellen van de student dan op het bieden van actieve hulp.

## De simulatiepatiënt

Simulatiepatiënten zijn personen die ingezet worden om bepaalde ziekten met de bijbehorende verschijnselen te simuleren. Ook hier zijn verschillende overwegingen van belang.

*Welke specifieke eisen moeten gesteld worden aan simulatiepatiënten in toetssituaties?*

Simulatiepatiënten die tijdens toetssituaties worden ingezet, dienen over andere vaardigheden te beschikken dan simulatiepatiënten in het onderwijs. In het onderwijs worden simulatiepatiënten veelal ingezet om anamnestiche en fysisch-diagnostische verschijnselen te simuleren. Alle andere informatie kan de simulatiepatiënt meestal zelf invullen, waarbij gebruik gemaakt kan worden van eigen ervaringen, hetgeen de echtheid en overtuiging waarmee de rol gespeeld wordt ten goede komt.

Onder toetsomstandigheden dienen simulatiepatiënten geheel gestandaardiseerd te zijn, zodat alle studenten 'dezelfde' patiënten zien. Er moeten ook afspraken gemaakt worden over de minder relevante zaken van het ziektebeeld. Eveneens moeten afspraken gemaakt worden over de hoeveelheid informatie die verstrekt dient te worden en het tijdstip waarop dat gebeurt. Ook dient te worden bepaald welke informatie spontaan en welke informatie alleen bij navraag verstrekt mag worden. Kortom, de eisen die aan simulatiepatiënten gesteld worden, zijn bij toetssituaties veel hoger dan bij het reguliere onderwijs.

*Kunnen simulatiepatiënten hun rol gedurende langere tijd op dezelfde wijze blijven spelen?*

Simulatiepatiënten kunnen op uitstekende wijze (na training en selectie) een ziekerol uit-

beelden. Zelfs zo goed dat praktizerende artsen geen verschil merken tussen een echte patiënt en een simulatiepatiënt wanneer die op hun spreekuur komt.<sup>11</sup>

Onderzoeken hebben aangetoond dat de 'rolvastheid' van simulatiepatiënten gemiddeld genomen goed is. De overgrote meerderheid weet de klachten en symptomen op de juiste manier te presenteren, mits voldoende aandacht besteed is aan een uitgebreide training van deze personen.<sup>12 13</sup> Zo er al verschillen bestaan, lijken deze nauwelijks van invloed op de uiteindelijke beoordeling van studenten.<sup>14</sup>

*Hoe adequaat kunnen simulatiepatiënten studenten beoordelen?*

In de Nederlandse situatie worden simulatiepatiënten alleen binnen onderwijssituaties ingezet om feedback te geven aan studenten. In de praktijk blijkt deze feedback vaak niet optimaal. In het algemeen vinden simulatiepatiënten het moeilijk om zich kritisch uit te laten over een gesprek. Het is de vraag of dit aan de patiënten ligt of dat er meer geïnvesteerd moet worden in de selectie van simulatiepatiënten en feedbacktrainingen voor hen.

In de Verenigde Staten, waar simulatiepatiënten als beoordelaar bij toetsen worden ingeschakeld, blijkt dat hun beoordelingen redelijk overeenkomen met die van stafdocenten, mits voldaan is aan twee voorwaarden.<sup>15</sup> De simulatiepatiënt moet goed getraind zijn en de beoordeling moet plaatsvinden aan de hand van duidelijk omschreven criteria. Dit laat onverlet dat er ook redenen kunnen zijn om géén simulatiepatiënten als beoordelaars in toetssituaties in te zetten. Dit kan te maken hebben met de benodigde inhoudsdeskundigheid bij toepassing van globale beoordelingslijsten of met de geloofwaardigheid van de beoordeling naar studenten toe, zeker daar waar het certificerende examens betreft.

## De student

Studenten zijn de 'consumenten' van stationsexamens. Naast de mening van de consument over bepaalde aspecten van stationsexamens worden tevens de belangrijkste vragen rondom het gedrag van deze consumenten besproken.

### *Welke effecten hebben stationsexamens op het gedrag van studenten?*

Dat toetsen een sterk sturend effect hebben op het gedrag van studenten is reeds lang bekend.<sup>2</sup>  
<sup>16</sup> Voor stationsexamens betekent dit meestal dat studenten voor een toets beoordelingslijsten uit hun hoofd proberen te leren. De leerstijl is sterk op reproductie van kennis over de vaardigheid gericht, met de 'onderwijskundige' zekerheid dat deze kennis weer snel vergeten zal worden. Een dergelijke voorbereidingsmethode zorgt er eveneens voor dat het onderwijsprogramma ten tijde van deze examens slechts marginaal gevolgd wordt. Deze effecten zijn bepaald niet wenselijk. Deels hangen deze gedragspatronen echter samen met het feit dat het een 'examen' is, waarbij het niet uitmaakt of het een stationsexamen of een kennistoets betreft. De studenten denken daarbij ook aan hun studievoortgang en aan het aantal studiepunten dat met het examen behaald moet worden. Stationsexamens hebben een belangrijk positief effect: studenten gaan frequent oefenen voor de toets.

### *Wat zijn de kritiekpunten van studenten tijdens stationsexamens?*

De meeste kritiek leveren studenten op het functioneren van beoordelaars, de ervaren discrepantie tussen onderwijs en toetsing, en de tijd waarbinnen een bepaalde opdracht moet worden uitgevoerd.<sup>17</sup> Ten aanzien van het functioneren van beoordelaars oogst de houding van de observator de meeste kritiek. Observatoren blijken soms ongeïnteresseerd, stellen studenten niet gerust en onderbreken studenten vaak tijdens de uitvoering van een

opdracht. Kritiek op onvoldoende deskundigheid van de beoordelaar komt (gelukkig) vrijwel niet voor.

Veel kritiek bestaat er ook op gepercipieerde discrepanties tussen de criteria die gehanteerd worden bij de toetsing en de criteria die in het vaardigheidsonderwijs worden gehanteerd. Meestal is de kritiek te herleiden tot één of twee docenten die in het onderwijs andere accenten leggen bij het (lichamelijk) onderzoek of de vaardigheden gedurende het onderwijs minder gedetailleerd aanleren dan van studenten tijdens de toets verwacht wordt.

Een laatste belangrijk kritiekpunt betreft de tijd die studenten krijgen om een bepaalde opdracht uit te voeren. In de praktijk blijkt het moeilijk om vooraf in te schatten hoe lang een bepaalde opdracht duurt. Dit heeft ten minste twee oorzaken. In de eerste plaats is de gerichtheid van de toetsing vaak niet expliciet duidelijk. Moet de beschikbare tijd gericht zijn op het afronden van de opdracht door een zwak presterende (maar niet onvoldoende) student of moet de opdracht alleen door een goede student in de benodigde tijd kunnen worden afgerond. Kiest men voor het laatste dan zal een aanzienlijk deel van de studentenpopulatie in tijdnood komen. De keuze is afhankelijk van het doel van de toetsing. In het algemeen zal men onderscheid willen maken tussen de 'niet voldoende' student en de overige studenten en niet tussen de 'goede' student en de anderen. Vanuit deze optiek verdient de eerste benadering de voorkeur. Een tweede reden waarom studenten tijd tekort komen bij de uitvoering van hun opdracht is dat makers van examenmateriaal zich moeilijk kunnen verplaatsen in het vaardigheidsniveau van de student en de ervaring die een student heeft in de toepassing van bepaalde vaardigheden. Over het algemeen wordt het niveau van studenten overschat, waardoor al snel de neiging bestaat te veelomvattende opdrachten te maken per tijds-eenheid.

## De docent en de onderwijsorganisatie

Ook docenten en de onderwijsorganisatie spelen een belangrijke rol bij stationsexamens; zij stellen het examenmateriaal op, bepalen de zak/slaaggrens en zorgen voor een correcte, valide en betrouwbare toetsafname.

### *De constructie van stations*

Docenten maken het onderwijs en zijn derhalve de eerst aangewezenen om ook de toetsproductie ter hand te nemen. Dat kost veel tijd. Belangrijk bij de constructie van stations zijn criteria op basis waarvan studenten beoordeeld worden. Deze criteria kunnen door de onderwijsinstelling zelf worden geformuleerd; waar nodig zal men teruggrijpen op leerboeken. Desondanks zal het vrijwel nooit voorkomen dat criteria zonder meer uit (eigen) literatuur kunnen worden overgenomen als toetsingscriteria. Regelmatig ontstaan er discussies over de aard en de mate van detaillering van de criteria en mogelijke alternatieven. Dit is echter niet de enige reden waarom het maken van examenmateriaal voor observatietoetsen veel tijd kost. Een andere reden is dat, behalve aan de toetsingscriteria, ook tijd besteed moet worden aan het maken van de opdracht, aan een schriftelijke rol-instructie voor simulatiepatiënten, aan een schriftelijke instructie voor observatoren en aan schriftelijke informatie ten behoeve van de student. Tevens dient men zich ervan te vergewissen dat voor een toetsstation een onderwerp gekozen wordt dat aan de volgende eisen voldoet: het moet observeerbaar zijn (dus geen inwendig onderzoek), het moet niet beperkt zijn tot het opzeggen van rijtjes gegevens behorend bij personen met normale bevindingen (zoals vaak gebeurt bij het inzetten van gezonde proefpersonen), de noodzakelijke materialen dienen in voldoende mate aanwezig te zijn (bijvoorbeeld fantomen, verbanden, injectienaalden) en het onderwerp moet in de beschikbare tijd kunnen worden getoetst.

### *Het vaststellen van zak/slaagnormen bij stationsexamens*

Bij de vaststelling van normen zijn globaal genomen twee vragen van belang: In welke mate moet alléén de inhoud van het examen het cijfer van de toets bepalen? En in welke mate moet men compensatie toestaan bij het bepalen van het eindcijfer?

Ten aanzien van de eerste vraag bestaan er twee stromingen. Vertegenwoordigers van de ene stroming vinden dat de inhoud allesbepalend moet zijn bij de eindbeoordeling en vertegenwoordigers van de andere vinden dat - naast inhoud - ook andere factoren verdisconteerd moeten worden in het eindcijfer. Te denken valt bijvoorbeeld aan onvolledig/slecht onderwijs en niet optimaal examenmateriaal of niet optimale examenomstandigheden.

De eerste groep kan zich meestal vinden in een of andere vorm van absolute normering, de tweede groep bestaat meestal uit voorstanders van een relatieve normering, waarbij de prestaties van de groep kandidaten als geheel van invloed zijn op de zak/slaagnorm.<sup>18</sup> Er zijn verschillende methoden om tot absolute en relatieve normen te komen, waarvan sommige ook bij stationsexamens kunnen worden toegepast.<sup>19</sup> Wanneer men binnen de instelling over voldoende mogelijkheden beschikt, verdient de volgende wijze van normbepaling de voorkeur. Men selecteert meerdere inhoudsdeskundigen, die ieder de zak/slaaggrens van het examen bepalen. Het gemiddelde van deze zak/slaaggrenzen wordt de voorlopige cesuur. Men neemt vervolgens de toets af en bepaalt enkele psychometrische indicatoren (gemiddelde scores per item en per station en tevens het aantal onvoldoendes per station en per toets). Er wordt eveneens voor gezorgd dat men de beschikking heeft over commentaarformulieren van beoordelaars en studenten over de desbetreffende toets. De psychometrische indicatoren en de commentaarformulieren worden door dezelfde inhoudsdeskundigen bekeken en deze informatie kan vervolgens



leiden tot bijstelling van de oorspronkelijk bepaalde cesuur.

Los van bovenstaande benaderingswijzen speelt ook de vraag in hoeverre men bereid is studenten compensatie toe te kennen bij het berekenen van de eindscore per station (compensatie tussen items onderling), per toets (compensatie tussen stations onderling) of tussen toetsen. Hierover bestaan tussen docenten vaak grote verschillen van inzicht. Het verdient aanbeveling om tenminste enige vorm van compensatie binnen en tussen stations toe te staan.

### Betrouwbaarheid

Bij stationsexamens kunnen met behulp van de generaliseerbaarheidstheorie verschillende soorten betrouwbaarheden worden bepaald, afhankelijk van de variantiebronnen van het stationsexamen.<sup>7</sup> De meeste waarden schommelen tussen .40 en .80. Ongewenste variantiebronnen zijn observatoren, simulatiepatiënten en stations. Zoals eerder vermeld, is de overeenstemming in beoordeling tussen twee observatoren in het algemeen hoog. De beoordelaarsvariantie is derhalve geen betekenisvolle foutenbron.

Ook de verschillen in betrouwbaarheid tussen simulatiepatiënten zijn reeds eerder besproken. Hieruit bleek dat, behoudens incidentele uitzonderingen, simulatiepatiënten goed in staat zijn essentiële onderdelen van hun rol betrouwbaar weer te geven. Indien er grote aantallen simulatiepatiënten nodig zijn (bijvoorbeeld bij landelijke examens), is het moeilijk vaste criteria te handhaven en het benodigde trainingsaanbod te geven om de kwaliteit te handhaven. De gegevens die simulatiepatiënten verstrekken in hun rol als beoordelaar blijken eveneens voldoende betrouwbaar te zijn. Variaties tussen simulatiepatiënten lijken derhalve ook geen wezenlijke foutenbron te zijn.

De laatst ongewenste variantiebron, de stations, zorgen voor het grootste deel van de (on)betrouwbaarheid van stationsexamens.

Dit komt omdat een behaalde prestatie op het ene station niets zegt over de prestaties op andere stations. Dit geldt zelfs binnen één vakgebied.<sup>20</sup> Deze inhoudsspecificiteit doet zich bij allerlei soorten toetsen voor. Voor de vaardigheidstoets is het effect van de inhoudsspecificiteit geringer naarmate de toets uit meer, inhoudelijk verschillende stations bestaat, dat wil zeggen een betere dekking geeft van het totale te toetsen domein. Met de huidige stationsduur komt dat gemiddeld neer op toetstijden van minimaal 4 uur om tot de gewenste betrouwbaarheid van .80 te komen. Overigens kan het ook zijn dat competentie uit zoveel specifieke vaardigheden bestaat dat men bij stationsexamens misschien tevreden moet zijn met waarden die een stuk lager liggen dan de gewenste .80.

Betrouwbaarheden worden in de literatuur doorgaans gerapporteerd vanuit een normgeoriënteerd perspectief, dat wil zeggen dat de score van de student pas betekenis heeft in relatie tot de score van de andere studenten. Dit in tegenstelling tot het domeingeoriënteerde perspectief waarbij het gaat om absolute posities van studenten ten opzichte van de leerstof, dat wil zeggen ten opzichte van een tevoren vastgesteld criterium voor de beheersing van de stof. De domeingerichte benadering stelt hogere eisen aan de betrouwbaarheid. De betrouwbaarheid van zak/slaagbeslissingen wordt daarbij groter naarmate de cesuur sterker afwijkt van de gemiddelde score.

### Validiteit

Evenals bij de betrouwbaarheid bestaan er ook verschillende validiteitscriteria. Men onderscheidt onder andere *constructvaliditeit*, *criteriumvaliditeit*, *inhoudsvaliditeit* en *predictieve validiteit*. Evenals bij andere examens is het probleem van stationsexamens dat er geen gouden standaard bestaat waartegen gegevens kunnen worden afgezet. Om inzicht te krijgen in de validiteit van stationsexamens dienen dus

verschillende soorten informatie uit diverse studies bij elkaar genomen te worden.

*Constructvaliditeit* bepaalt men door empirisch te onderbouwen wat theoretisch verondersteld wordt. Vaak gebeurt dat door verschillende groepen studenten van verschillend competentieniveau hetzelfde stationsexamen te laten afleggen, om vervolgens te concluderen of het verwachte verschil in competentie ook door de betreffende toets gemeten is. Dergelijke studies zijn bijvoorbeeld gedaan door Klass, die inderdaad verschillen tussen niveaugroepen vond.<sup>21</sup> Door het ontbreken van een gouden standaard blijft het echter onduidelijk *welk* verschil verwacht moet worden bij de diverse niveaugroepen en welke betekenis deze verschillen hebben. Ook is getracht om onderliggende inhoudelijke factoren te bepalen uit data van stationsexamens. Hierbij vindt men twee te onderscheiden factoren: een cognitieve factor (data-interpretatie, (differentiaal) diagnostiek, aanvragen onderzoek en beleid) en een niet-cognitieve factor (attitude, gesprekstechnische vaardigheden en fysisch-diagnostisch onderzoek). Daarom is het zinvol de resultaten van studenten in deze twee factoren weer te geven.<sup>10 22</sup>

Bij *criteriumvaliditeit* wordt getracht correlaties te vinden tussen scores behaald bij een stationsexamen en andere indicatoren. Men zou verwachten dat er een gering verband bestaat tussen resultaten behaald bij stationsexamens en kennistoetsen, en een betrekkelijk sterk verband tussen resultaten van stationsexamens en stagebeoordelingen. In de praktijk blijkt echter dat er zeer wisselende correlaties bestaan tussen gegevens van stationsexamens en andere relevante indicatoren.<sup>23</sup> Correlatiestudies blijken derhalve een geringe waarde te hebben bij het bepalen van de validiteit van stationsexamens.

*Inhoudsvaliditeit* betreft de mate waarin belangrijke elementen van het vakgebied evenredig worden weerspiegeld in de toets en tevens de vraag of de vormgeving van de toets

een redelijke weerspiegeling vormt van de praktijk. Verder dient de toets relevant te zijn en fair in de beoordeling ten opzichte van studenten. Inhoudsvaliditeit wordt vastgesteld door inhoudsdeskundigen de inhoud van de toets te laten beoordelen op bovengenoemde aspecten. In het algemeen is de inhoudsvaliditeit een sterk punt van stationsexamens. De sterke impuls tot de ontwikkeling van stationsexamens betreft de gelijkenis met wat er in de praktijk gebeurt.<sup>24</sup> Toch valt er nog veel te verbeteren aan de inhoud van de test. Wordt er wel systematisch gekeken of het vakgebied op de juiste wijze gerepresenteerd is in de toets? Is er een afweging geweest met betrekking tot de keuze van het ene vakgebied ten opzichte van het andere vakgebied? Wordt het examenmateriaal op de juiste wijze voorbereid? Zijn de opdrachten aan studenten duidelijk? Bestaat er eenduidigheid over de mate van gedetailleerdheid van items? Geven de items de meest essentiële punten weer van de gevraagde opdracht? Bevat het station niet te veel 'ruis'? Op een groot aantal van deze punten valt nog veel winst te boeken. Er kan derhalve nog fors geïnvesteerd worden in de kwaliteitsbewaking van de stationsexamens.

Voor wat betreft de predictieve validiteit zijn er nog bijzonder weinig studies bekend waarbij resultaten van stationsexamens gebruikt zijn om resultaten van toekomstig functioneren te voorspellen. Alleen Vu geeft aan dat resultaten van stationsexamens een redelijke predictie geven voor het succesvol doorlopen van vervolgopleidingen.<sup>25</sup> Er bestaan echter ook gegevens waaruit blijkt dat stationsexamens in de preklinische fase weliswaar een predictieve waarde hebben voor later functioneren in de klinische fase, maar dat dit nog sterker geldt voor kennistoetsing.<sup>8</sup> Nader onderzoek hierover zal nog nodig zijn.

## Kosten

Van observatietoetsen wordt gezegd dat het

dure toetsen zijn. Een dergelijke uitspraak dient echter gerelateerd te worden aan de te verwachten opbrengst. Wat is duur als men kan en mag verwachten dat de kwaliteit van de aanstaande artsen op het gebied van vaardigheden toeneemt doordat vaardigheden onderwezen en getoetst worden? Hoeveel onnodig specialistisch onderzoek zal later in de beroepsbeoefening mogelijk worden voorkomen door een hoger niveau van vaardigheidsbeheersing? Een antwoord hierop valt vooralsnog niet te geven. Ook hier ligt nog een onderzoeksterrein braak.

Gesteld kan worden dat, indien onderwijs en toetsing op het gebied van vaardigheden belangrijk gevonden wordt in een opleiding, er ook een kostenplaatje bij hoort. De bereidheid om de bijbehorende kosten voor vaardigheids-toetsing te dragen zegt iets over de werkelijke prioriteit die een instelling aan deze onderwijs-activiteit toekent.

Bovenstaande neemt niet weg dat er naar wegen gezocht moet worden om de toetsing zo efficiënt en goedkoop mogelijk te laten verlopen. Een aantal methoden is reeds toegepast. De methoden tot kostenbesparing zijn te verdelen in kostenbesparende maatregelen binnen observatietoetsen zelf en kostenbesparende maatregelen door middel van alternatieve toetsen die qua doelstelling uitspraken kunnen doen over het vaardigheidsniveau van studenten. De volgende kostenbesparende maatregelen zijn binnen observatietoetsen zelf toegepast:<sup>5</sup>

- Zorg ervoor dat alleen die onderwerpen in een observatietoets getoetst worden die niet goedkoper in een andere toetsvorm (bijvoorbeeld kennistoetsing) getoetst kunnen worden.
- Overweeg de mogelijkheid tot sequentiële toetsing, dat wil zeggen neem een relatief korte toets af voor alle studenten en vervolg de toets met die studenten die net op de grens voldoende/onvoldoende zitten.
- Overweeg de mogelijkheid om studenten

van de eigen instelling te gebruiken als simulatiepatiënt.

- In formatieve beoordelingssituaties kunnen studenten elkaar beoordelen en feedback geven aan de hand van de beoordelingsformulieren.
- Laat in bepaalde situaties student-assistenten de rol van docenten overnemen.

De volgende kostenbesparende maatregelen zijn mogelijk door het toepassen van toetsmethoden die (min of meer) vergelijkbaar zijn met observatietoetsen:

- Laat eenmaal aangeleerde vaardigheden in het onderwijs aftekenen en beoordelen; laat vervolgens alleen die studenten die een onvoldoende beoordeling hebben of anderszins onvoldoende vaardigheden hebben een stationsexamen doorlopen.
- Toets de kennis die gerelateerd is aan vaardigheden in een aparte kennistoets over vaardigheden (een dergelijke toets kan de observatietoets echter nooit geheel vervangen).
- Laat studenten videofragmenten beoordelen waarop bepaalde vaardigheden goed of slecht worden uitgevoerd.
- Ontwikkel computertoetsprogramma's waarbij zowel theoretische kennis met betrekking tot vaardigheden als beeldmateriaal gebruikt kunnen worden.
- Beoordeel inspectie door middel van foto's, dia's of videofragmenten.
- Verzamel op gestructureerde wijze praktijk-oordelen van verschillende docenten over studenten en kom op basis hiervan tot een eindoordeel over het vaardigheidsniveau.

### Mogelijkheden tot verbetering

Stationsexamens kunnen op een aantal punten verbeterd worden. Een van de belangrijkste punten die verbetering behoeven, is de kwaliteitszorg vóór, tijdens en na de toetsafname. Tot de kwaliteitszorg voor de toetsafname behoort het bewaken van de validiteit van het

examenmateriaal omdat het ontwerpen van het stationsexamen in feite begint met het maken van adequaat examenmateriaal en zijn vervolg vindt in een correcte afname. Fouten in dit proces moeten worden voorkomen, aangezien die doorwerken in alle data en zowel de betrouwbaarheid als de validiteit aantasten. Ook de zorg na afloop van een toets blijft belangrijk. De nabesprekingen van stationsexamens dienen serieus ter hand te worden genomen met de nodige ondersteunende informatie. Concreet betekent dit dat de aandacht van stationsmakers in de meeste gevallen veel meer gericht zou moeten zijn op de inhoud van de stations. De inhoud zou bekeken moeten worden op relevantie en aansluiting met het onderwijs in de diverse disciplines. Aangezien het onderwijs in vaardigheden in veel gevallen niet ggeëxpliciteerd is in criteria waaraan studenten zich dienen te houden, ligt hier dus ook nog een taak voor het onderwijs.

Eveneens zouden stationsmakers zich van tevoren moeten afvragen welke elementen van het onderzoek getoetst dienen te worden. Moet de toetsing met name gericht zijn op procedures en het procesmatige karakter van de handeling of moet de toetsing gericht worden op uitkomsten van handelingen. Eveneens dient men zich af te vragen tot welk detail men de handeling getoetst wil zien. De meeste gedetailleerde beoordelingslijsten herbergen het gevaar van trivialiteiten en het uit het hoofd leren van rijtjes bij studenten. Globale lijsten geven minder gedetailleerde feedback en vereisen meer inhoudsdeskundigheid van observatoren. Kortom, er valt veel te overdenken met betrekking tot kwaliteitszorg voordat de toets heeft plaatsgevonden.

Naast de inhoud is aandacht voor de vorm waarin het examen wordt afgenomen belangrijk. Punten waarop gelet moet worden zijn de formulering van de opdracht, de aansluiting van de opdracht op de beoordelingslijst, de eenduidigheid van de formulering van items, de te kiezen schaalverdeling enzovoorts.

Echter ook kwaliteitszorg en professionaliteit tijdens de afname kunnen in vele gevallen beter. Weet iedere actor (observator, simulatiepatiënt, student, docent en onderwijsorganisatie) wat van hem/haar verwacht wordt, op welke tijdstippen en op welke locatie? Vaak wordt vergeten dat ook observatoren en simulatiepatiënten gespannen zijn tijdens toetsituaties. Om te zorgen dat deze stress hanteerbaar blijft en niet wordt afgereageerd op staf of studenten zijn een goede organisatie, voorlichting en instructie van wezenlijk belang. Zorg er ook voor dat tijdens de afname iedere actor kritiek kan uiten op de inhoud en organisatie van het stationsexamen. Dit zijn gegevens die nuttige informatie kunnen verschaffen bij het bepalen van de items die meetellen voor het eindcijfer. Bovendien kunnen dergelijke gegevens gebruikt worden bij de constructie van volgende toetsen.

De kwaliteitszorg na de toets is ook een belangrijke fase in het proces tot het verkrijgen van valide toetsgegevens. Commentaren van beoordelaars, simulatiepatiënten, studenten en anderen (organisatie) kunnen samen met de verkregen psychometrische variabelen (ten minste omvattend de studentscores per item en de gemiddelde scores per station) worden gebruikt om tot een afgewogen eindoordeel te komen over de kwaliteit van het examenmateriaal. Vaak leiden dit soort besprekingen tot uitval van items, zelfs soms tot uitval van stations, maar bovendien leiden de besprekingen tot een deskundigheidsopbouw met betrekking tot inhoudelijke en vormtechnische kennis omtrent constructie van stations. Ook wordt op een dergelijke manier de kwaliteit van het onderwijs bewaakt. Uiteraard zijn dit niet de enige punten waarop verbetering mogelijk is. Maar op bovenstaand terrein lijkt nog veel winst te behalen.

## Aanbevelingen

Met stationsexamens is ruim vijftien jaar erva-

ring opgedaan in binnen- en buitenland. Met name de laatste jaren is er veel bekend geworden over de kenmerken van dergelijke examens. In het voorafgaande is slechts een deel van de ervaringen weergegeven. De belangrijkste punten worden hier samengevat in een beperkt aantal praktische aanbevelingen voor diegenen die met de uitvoering van stationsexamens belast zijn. Wat hieronder vermeld wordt kan deels onderbouwd worden met onderzoek, deels zijn het waarnemingen die vaker gerapporteerd zijn bij de uitvoering van stationsexamens.

#### *Beoordelaars*

- Bij de training van beoordelaars is het belangrijker aandacht te besteden aan de procedures tijdens het examen en de invloed van de attitude van de beoordelaar op de student dan aan het kennisniveau dat bij de beoordelaar verondersteld wordt.
- Beoordelaars dienen - indien technisch mogelijk - studenten feedback te geven tijdens de toets, maar *na* het uitvoeren van hun opdracht. Extra tijd zou hiervoor gereserveerd kunnen worden.
- Het beoordelen van beoordelaars door studenten heeft meer effect dan training van beoordelaars.

#### *Simulatiepatiënten*

- Aan simulatiepatiënten die tijdens stationsexamens worden ingezet dienen hogere eisen te worden gesteld dan aan simulatiepatiënten die in het onderwijs worden ingezet. Dit betekent strengere selectie en goede trainingsmogelijkheden.
- Dit geldt zowel voor de simulatiepatiënt in de patiëntenrol als voor de simulatiepatiënt in de rol van beoordelaar.
- Simulatiepatiënten als beoordelaars zijn inzetbaar bij stations waar weinig voorkennis wordt verondersteld.

#### *Studenten*

- Er bestaat een duidelijke relatie tussen de mate van gedetailleerdheid van criteria vermeld in de toets en het 'leergedrag' bij studenten.

#### *Docent/onderwijsorganisatie*

- Het ontwerpen van stationsexamens vergt deskundigheid.
- Het ontwerpen van stationsexamens kost veel tijd (10 tot 20 uur per station).

#### *Normering*

- Het stellen van de zak/slaagnorm is slechts afhankelijk van twee factoren:
  - In welke mate bestaat de bereidheid om andere dan inhoudelijke factoren een rol te laten spelen bij de uiteindelijke bepaling van de norm?
  - In welke mate is men bereid compensatie toe te staan?

#### *Betrouwbaarheid*

- Zorg dat stationsexamens zoveel mogelijk bestaan uit te onderscheiden inhoudelijke entiteiten (stations). Hoe meer stations in de toets hoe beter.
- Zorg ervoor dat beoordelaars zoveel mogelijk (als co-beoordelaar) worden ingezet voor de beoordeling van studenten en niet van collega's.
- Globale beoordelingen kunnen even betrouwbaar zijn als gedetailleerde beoordelingen, maar de beide beoordelingsvormen onderscheiden zich in hun effecten.
- Zorg ervoor dat alleen die elementen in een stationsexamen komen die niet op een andere wijze (schriftelijk, COO) getoetst kunnen worden.

#### *Validiteit*

- Zorg ervoor dat docenten gelegenheid hebben expertise te verwerven in het ontwerpen van toetsstations.
- Stel een toetscommissie in die zorg draagt

voor een adequate kwaliteitsbewaking ten aanzien van productie vooraf, uitvoering tijdens en controle na het stationsexamen.

- Ga voor iedere toets na wat het doel is van de meting (evaluatie van het onderwijs/evaluatie van de praktijk) en op welke wijze het examenmateriaal moet worden vormgegeven opdat het doel van de meting het best wordt benaderd.

## Tenslotte

De verdergaande professionalisering van het vaardigheidsonderwijs en de vaardigheidstoetsing leidt langzaam en gestaag in de richting van bovengenoemde suggesties. Het zal echter op zijn vroegst 'nabije toekomst' zijn voordat alle kennis die tot op heden is opgebouwd zijn weerslag vindt in de praktische toepassingen van stationsexamens.

In de toekomst kan ook gedacht worden aan het ontwikkelen van landelijke centra voor vaardigheidsonderwijs en -toetsing. Plaatsen waar de outillage aanwezig is en door onderzoek de know-how kunnen worden gebruikt om stationsexamens verder uit te bouwen in de richting van landelijke toetsing of toetsing binnen het kader van de nascholing. Studenten van diverse instellingen zouden in dergelijke centra een stationsexamen kunnen afleggen, dat bijvoorbeeld op video wordt opgenomen. De video kan dan weer op de eigen faculteit of hogeschool worden gebruikt als beoordelingsinstrument, volgens de criteria die de instelling hanteert. Toekomstmuziek, wie zal het weten?

## Literatuur

1. Harden R, Gleeson F. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Med Educ* 1976;13:41-54.
2. Newble DI, Jaeger K. The effects of assessments and examinations on the learning of medical students. *Med Educ* 1983;17:165-71.
3. Stillman P, Swanson D. Use of standardized patients for assessment of history-taking, physical examination and communication skills. Final Report on the ABIM Standardized Patient Project. The New England Consortium of Internal Medicine Residency Training Programs, 1989.
4. Foolen CHGM, Verwijnen GM, Luijk SJ van, Beusmans GHMI, Vleuten CPM van der. Assessment integrated in a long-term continuing medical education program in family medicine. Paper gepresenteerd op de International Conference on Evaluation in Medical Education. Beer Sheva, Israel, 1987.
5. Jansen JJM, Tan LHC, Vleuten CPM van der, Luijk SJ van, Rethans JJ, Grol RPTM. Assessment of competence in technical clinical skills of general practitioners using different methods. *Med Educ* 1995;29:247-53.
6. Vleuten CPM van der, Luijk SJ van, Ballegoien MJ van, Swanson D. Training and experience of examiners. *Med Educ* 1989;23:290-6.
7. Vleuten CPM van der, Swanson DB. Assessment of clinical skills with standardized patients: state of the art. *Teach Learn Med* 1990;2:58-76.
8. Luijk SJ van. Al doende leert men [proefschrift]. Maastricht: Rijksuniversiteit Limburg, 1994.
9. Luijk SJ van, Louw A de, Visser K, Scherpbier AJJA, Vleuten CPM van der. De ondankbare taak als observer bij de vaardigheidstoets. In: Houtkoop E, Pols J, Pollemans MC, Scherpbier AJJA, Verwijnen GM, redactie. *Gezond Onderwijs-3*. 's Gravenhage: Haagse Hogeschool, 1993;125-30.
10. Vu NV, Barrows HS. Use of standardized patients in clinical assessments: Recent developments and measurement findings. *Ed Res* 1995;23:23-30.
11. Rethans JJ, Drop R, Sturmans F, Vleuten CPM van der. A method for introducing standardized (simulated) patients into general practice consultations. *Br J Gen Pract* 1991;41:97-9.
12. Tamblyn RM, Klass DJ, Schnabl GR, Kopelow ML. The accuracy of standardized patient presentation. *Med Educ* 1991;25:100-9.
13. Reznick R, Smee S, Rothman A, Chalmers A, Swanson D, Dufresne L et al. An objective structured clinical examination for the licentiate: Report of the pilot project of the Medical Council of Canada. *Acad Med* 1992;67:487-94.
14. Colliver JA, Robs RS, Vu NV. Effects of using two or more standardized patients to simulate the same case on case means and case failure rates. *Acad Med* 1991;66:200-2.
15. Vu NV, Marcy ML, Colliver JA, Verhulst SJ, Travis TA, Barrows HS. Checklist characteristics and length of testing: their effects on standardized patients' simulations. *J Med Educ* 1992;26:390-5.

16. Luijk SJ van, Vleuten CPM van der, Schelven RM van. Observer and student opinions about skills tests. In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra RP, redactie. Teaching and assessing clinical competence. Groningen: BoekWerk Publications 1990:497-502.
17. Visser K, Louw A de, Luijk SJ van, Scherpbier AJJA. De observator geobserveerd. In: Houtkoop E, Pols J, Pollemans MC, Scherpbier AJJA, Verwijnen GM, redactie. Gezond Onderwijs-3. 's Gravenhage: Haagse Hogeschool 1993:119-24.
18. Luijk SJ van, Wijnen W. Cesuurbepaling. In: Metz JCM, Scherpbier AJJA, Vleuten CPM van der, redactie. Medisch onderwijs in de praktijk. Assen: Van Gorcum 1995:238-46.
19. Livingston SA, Zieky MJ. Passing scores. Princetown: Educational Testing Service, 1982.
20. Norman GR. Objective measurement of clinical performance. Med Educ 1985;19:43-7.
21. Klass D, Campbell C, Hassard T, Kopelow M, Schnabl G. Influence of level of training on performance in a standardized test of clinical abilities. In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra RP, redactie. Teaching and assessing clinical competence. Groningen: BoekWerk Publications 1990:327-32.
22. Verhulst SJ, Colliver JA, Paiva REA, Williams RG. A factor analytic study of performance of first-year residents. J Med Educ 1986;61:132-4.
23. Vu NV, Barrows HS, Marcy ML, Verhulst SJ, Colliver JC, Travis TA. Six years of comprehensive, clinical performance-based assessment using standardized patients at the Southern Illinois University School of Medicine. Acad Med 1992;67:42-50.
24. Newble D. Eight years' experience with a structured clinical examination. Med Educ 1988;22:200-4.
25. Vu NV, Distlehorst LH, Verhulst SJ, Colliver JA. Clinical performance based test sensitivity and specificity in predicting first year-residency performance. Acad Med 1993;68:41-5.

#### DE AUTEUR

*S.J. van Luijk is als universitair docent verbonden aan de vakgroep Onderwijsontwikkeling en -onderzoek van de Faculteit der Geneeskunde, Universiteit Maastricht met als speciaal aandachtsgebied de toetsing van vaardigheden.*

#### Correspondentie-adres:

*S.J. van Luijk, Universiteit Maastricht, Vakgroep Onderwijsontwikkeling en -onderzoek, Postbus 616, 6200 MD Maastricht.*