E. DE GRAAFF
H.J.M. VAN BERKEL
M.J. DROP

SIMULATIE VAN INITIEEL MEDISCH PROBLEEMOPLOSSEN

EEN SCHRIFTELIJKE TOETS VOOR VAARDIGHEID IN HET OPLOSSEN VAN MEDISCHE PROBLEMEN: SIMP

E. de Graaff
onderwijsontwikkeling
en -research
H.J.M. van Berkel
onderwijsontwikkeling
en -research
M.J. Drop
medische sociologie
Rijksuniversiteit Limburg
Maastricht

Correspondentieadres:
E. de Graaff
Rijksuniversiteit Limburg
Onderwijsontwikkeling
en Onderwijsresearch
Postbus 616
6200 MD Maastricht
Tel.: 043 - 882303

Net als elders in het onderwijs bestaan medische examens tegenwoordig grotendeels uit objectief scoorbare meerkeuzetoetsen. Aan de Medische Faculteit van de Rijksuniversiteit Limburg wordt de groei van de medische kennis van studenten geëvalueerd met een voortgangstoets bestaande uit ± 250 juist-onjuist items.[1] Een beperking van de voortgangstoets is dat met dit soort items geïsoleerde feitenkennis relatief sterk benadrukt wordt. Voor het toetsen van de beheersing van praktisch medische vaardigheden wordt daarom naast de voortgangstoets een aparte vaardigheidstoets afgenomen.[2] Reeds geruime tijd bestaat het inzicht dat naast de toetsing van kennis en praktisch medische vaardigheden ook vaardigheid in "medisch probleemoplossen" gerepresenteerd dient te zijn in het artsexamen.[3] Aan de Rijksuniversiteit Limburg is in 1979 een project gestart met als doel het bewerken of ontwikkelen van een instrument voor het toetsen van vaardigheid in medisch probleemoplossen.

Medisch probleemoplossen wordt omschreven als: de vaardigheid naar aanleiding van de klachten van een patiënt gegevens te verzamelen door middel van anamnese, lichamelijk onderzoek en laboratoriumtests, het verwerken van deze informatie tot een diagnose, en het instellen van een beleid.[4] Met andere woorden de vaardigheid medische kennis in de praktijk toe te passen bij het oplossen van problemen die patiënten aan een arts presenteren. Uitgaande van beschrijvingen van probleemsituaties uit de praktijk zijn verschillende instrumenten ontwikkeld voor het toetsen van deze vaardigheid. De bekendste hiervan is het Patient Management Problem (PMP).

Een PMP bestaat uit een korte casus beschrijving, gevolgd door vragen met een reeks antwoordmogelijkheden. Na het kiezen van een optie krijgt de kandidaat de bijbehorende informatie en een volgende vraag, tot de casus geheel is doorgewerkt.[5] Onderzoek met PMP's heeft echter twijfel doen rijzen ten aanzien van de validiteit van dit instrument. Zo worden regelmatig lage correlaties aangetroffen tussen PMP's onderling.[6-8] Verder worden de resultaten vertekend door een sturend effect van de antwoord-opties.[9]

Dit laatste probleem is ondervangen bij een ander instrument: de Modified Essay Question (MEQ). Een MEQ bestaat uit een casus gevolgd door een reeks open vragen.[10] Na het beantwoorden van elke vraag wordt feedback gegeven over het verdere verloop van de casus. Ook wat betreft dit instrument geeft recent onderzoek echter aanleiding tot twijfel ten aanzien van de validiteit.[11]

Het voornaamste probleem wordt gevormd door de lage correlaties die regelmatig worden aangetroffen tussen verschillende casus binnen een toets. Door sommige auteurs wordt dit verschijnsel geïnterpreteerd als ondersteuning voor de "inhoudsspecificiteit van medisch probleemoplossen".[5,6] Een veel eenvoudiger verklaring wordt gegeven door Norman, die stelt dat casus kunnen worden opgevat als items in een test.[8] Matige correlaties tussen de items in een test zijn niet ongewoon. Om een betrouwbare meting te verkrijgen moet de test wel uit voldoende verschillende items bestaan. Een consequentie van de lage correlatie tussen casus is dus dat met een toets die uit slechts enkele casus bestaat geen acceptabele betrouwbaarheid gerealiseerd kan worden. De resultaten van zo'n onbetrouwbare meting kunnen vervolgens ook nooit valide zijn. De consequentie van deze redenering is dat een toets voor vaardigheid in medisch probleemoplossen uit een substantieel aantal verschillende casus dient te bestaan.

Aangezien het invullen van een PMP of MEQ gewoonlijk ruim een half uur in beslag neemt, is dat met deze instrumenten moeilijk te realiseren. Een van de uitgangspunten bij de ontwikkeling van een nieuw instrument voor het toetsen van vaardigheid in medisch probleemoplossen, was het terugbrengen van de tijd die nodig is voor het beantwoorden van één casus tot vijf à tien minuten.
Dit artikel geeft een korte beschrijving van dit instrument, gevolgd door een overzicht van de resultaten van onderzoek naar de betrouwbaarheid en validiteit.

SIMULATIE VAN INITIEEL MEDISCH PROBLEEMOPLOSSEN (SIMP)

Onderzoek naar de aard van medisch probleemoplossen heeft uitgewezen, dat de initiële hypothesen die direct in het begin van het arts-patiënt contact ontstaan van cruciaal belang zijn voor het verdere verloop van het contact.[6-8] Ervaren artsen onderscheiden zich van studenten door hun vermogen snel de essentiële aspecten van een casus te onderkennen.[12] Het komt erop neer dat wie aan het begin van een arts-patiënt contact niet op het goede spoor zit, er meestal ook niet in slaagt de juiste aanvullende informatie te vinden om het probleem op te kunnen lossen. Verschillen in de vaardigheid in het oplossen van medische problemen zullen dus al in de initiële fase van het probleemoplossingsproces aan het licht komen.
Tegen deze achtergrond is bij de constructie van SIMP de nadruk gelegd op de aanvang van het contact, zonder te proberen het verdere verloop van de interactie tussen arts en patiënt te simuleren.[13] Voorts is gekozen voor open vragen om het sturende effect van antwoord-opties te vermijden. Weliswaar komt uit een overzicht van de literatuur naar voren dat (na correctie voor attenuatie) doorgaans geen noemenswaardige verschillen kunnen worden aangetoond tussen open antwoord toetsen en dezelfde vragen voorzien van voorgestructureerde antwoordmogelijkheden, juist bij een taak als het formuleren van hypothesen worden echter wel systematische verschillen aangetroffen.[14] De open vraag vorm lijkt beter geschikt om het vermogen van kandidaten te toetsen zelf hypothesen te genereren.
Simulatie van Initieel Medisch Probleemoplossen (SIMP), bestaat uit korte beschrijvingen van casuistiek.

Voorbeeld casus
Situatiebeschrijving: huisarts-spreekuur

Man, 36 jaar, gehuwd, twee kinderen, hoofdonderwijzer van een lagere school in een snel uitbreidende nieuwbouwbuurt.
Medische voorgeschiedenis: twee jaar geleden maagklachten, gastritis of beginnend ulcus; bestreden met antacida, cimethidine en twee weken rust.
Komt nu met de klacht: diarree en krampen in de onderbuik, (wijst hele buikstreek aan); het duurt al ongeveer een week; ontlasting is volgens de patiënt dun en zwart van kleur; vanmorgen twee keer dunne ontlasting met bloed en slijm.
Hij voelt zich de laatste dagen vaak misselijk, geen eetlust, soms een drukkend gevoel in de bovenbuik; is een paar kilo afgevallen; niet in staat te werken; bang voor "ernstige" ziekte.

Bij de casus wordt één enkele open vraag gesteld: "Wat zou u doen, als u als arts in de praktijk met een dergelijk geval werd geconfronteerd?"
De open vraag laat de respondenten geheel vrij in het kiezen van hun eigen formulering. In tegenstelling tot de procedure bij PMP's en MEQ's wordt informatie over het verloop van de casus (het effect van handelingen) echter pas achteraf gegeven. Het antwoord is daardoor beperkt tot hetgeen uit de direct beschikbare informatie kan worden afgeleid.
Voor de beoordeling is een systeem van score-sleutels ontwikkeld. De score-sleutels zijn gebaseerd op een bespreking van de casus met een team ervaren artsen en hebben de vorm van een checklist met omschrijvingen van elementen van een correct antwoord (de score-sleutel horend bij de bovenstaande casus is opgenomen op de volgende pagina).

Score-sleutel

Betekenis van de toegepaste symbolen:
+ beide elementen moeten genoemd worden.
/ gelijkwaardige alternatieven, of het ene of het andere element moet genoemd worden.
// inhoudelijk verschillende alternatieven, voor scoringsdoeleinden gelijkwaardig
* niet exact te omschrijven element

S (anamnese)

- ☐ Bang voor kanker // zelf idee van oorzaak // bij kennissen of familie ooit zoiets meegemaakt
- ☐ Ontlasting: kleur // ook zwarte ontlasting gehad // bloed of slijm //
- ☐ Ooit last van aambeien gehad // pijn van achter (rond anus) // jeuk van achter
- ☐ Eerder gehad // pijn anders of hetzelfde als 2 jaar geleden // nog last van maag gehad*
- ☐ Reis naar tropen / op reis geweest // wat gegeten laatste week / iets bijzonders gegeten / "uit de muur"*
- ☐ Werkdruk / stress / hoe is het thuis c.q. op werk*
- ☐ Medicijngebruik // alcohol
- ☐ Familieleden zelfde klachten // darmziekten in familie

O (lichamelijk onderzoek)

- ☐ Algemene indruk / zieke indruk*
- ☐ Inspectie huidkleur / gele huid // turgor // temp
- ☐ Gewicht

Buikonderzoek:

- ☐ Auscultatie + percussie + palpatie // letten op: verhoogde peristaltiek + zoeken naar dempingen + letten op défense musculaire
- ☐ Inspectie anaalstreek + rectaal toucher // proctoscopie

E (hypothesen)

- ☐ Gastro-enteritis // paratyfus (=salmoneflose) // tropische ziekte
- ☐ Colitis (ulcerosa) / proctitis
- ☐ Hemorroiden
- ☐ Psychische spanningen // psychosomatische aandoening*

P (werkplan)

- ☐ Rust + dieet + afspraak voor revisie
- ☐ Microscopisch onderzoek faeces // faeceskweek
- ☐ BSE + leuco's
- ☐ Verwijzing naar internist / salazopyrine
- ☐ Begeleiding voorstellen voor persoonlijke problemen

De beoordelaars behoeven geen waardeoordeel te geven, maar alleen overeenstemming tussen het antwoord en de score-sleutel te markeren. Om het scoren te stroomlijnen zijn de items ingedeeld volgens het SOEP-schema van Weed (S subjectieve informatie/anamnese; O objectieve informatie/lichamelijk onderzoek; E evaluatie/diagnose; P plan/beleid).[15] Deze structuur wordt echter niet aan de respondenten opgedrongen. Zij zijn bij het formuleren van hun antwoord vrij in het kiezen van een eigen volgorde.

BETROUWBAARHEID

Bij een test met open vragen zijn beoordelaars nodig voor het scoren van de antwoorden. Verschillen tussen beoordelaars hebben een negatief effect op de betrouwbaarheid van de test. Bij SIMP is getracht dit probleem zo veel mogelijk te ondervangen door het structureren van de beoordelingstaak met behulp van score-sleutels. Met deze beoordelingsmethode behoeft de beoordelaar zelf geen medisch expert te zijn. Wel is kennis van medische begrippen en terminologie nodig voor het herkennen van afkortingen, synoniemen of alternatieve formuleringen. Van verpleegkundigen mag op grond van hun opleiding en ervaring verwacht worden dat ze aan dit criterium voldoen.

De betrouwbaarheid van de beoordelaars is onderzocht in twee experimenten. In een eerste onderzoek werden de antwoorden van 25 studenten op 22 casus (door onvolledig invullen een totaal van 500 antwoorden) beoordeeld door zes verpleegkundigen. Ter controle is vervolgens een deel van deze antwoorden (een viertal casus, zoveel mogelijk door alle studenten beantwoord) opnieuw gescoord door twee artsen.[16] De overeenstemming tussen de beoordelaars in deze experimenten bleek hoog te zijn. Met de Intraclass Correlatie Coëfficiënt (ICC) werd de correlatie van beoordeling van één random gekozen verpleegkundige, met het gemiddelde van de populatie verpleegkundige beoordelaars geschat op 0.83. Nadere analyse bracht aan het licht, dat onvolkomenheden in enkele score-sleutels en slordigheid van een der beoordelaars de betrouwbaarheid negatief beïnvloedden.
Berekend werd dat met verbetering van de score-sleutels en een selectie van beoordelaars een betrouwbaarheid (ICC) van 0.93 gerealiseerd kon worden. Deze zeer hoge betrouwbaarheid geeft aan dat verpleegkundigen wat betreft onderlinge overeenstemming goed voldoen als beoordelaars.
Er werden geen significante verschillen gevonden tussen beoordelaars, ook niet tussen artsen en verpleegkundigen. Bij nadere analyse bleek echter dat de onderlinge overeenstemming tussen de twee artsen lager was dan die tussen

de verpleegkundigen (betrouwbaarheidscoëfficiënt 0.72 respectievelijk 0.84 over de vier casus). De lagere overeenstemming tussen de twee artsen kan mogelijk verklaard worden op basis van hun inhoudelijke deskundigheid. Artsen hebben op grond van hun eigen medische ervaring een opvatting over de wijze waarop het probleem zou moeten worden opgelost. Dergelijke verschillen in opvatting ten aanzien van medisch handelen kunnen de scoring beïnvloeden. Bij verpleegkundigen speelt een dergelijke eigen opvatting ten aanzien van medisch handelen niet of in veel mindere mate mee. Verpleegkundigen zien zichzelf niet als medische experts en zullen dus bij verschil van inzicht ten opzichte van de score-sleutels toch geneigd zijn de scoringsinstructie strikt te volgen.

Aan de andere kant bestaat het risico dat verpleegkundigen door gebrek aan inhoudelijke deskundigheid tot een onjuist oordeel komen (bijvoorbeeld: een item missen door onbekendheid met een synonieme uitdrukking). Bij zorgvuldige controle van de beoordelingen werden inderdaad enkele van dergelijke beoordelingsfouten gesignaleerd. Op het totaal van 1540 zowel door de verpleegkundigen als door de artsen gescoorde items betrof dit echter minder dan 1% van de gevallen. De invloed van deze beoordelingsfouten op de betrouwbaarheid is daarmee aanmerkelijk kleiner dan die van de verschillen in inzicht tussen de artsen.

Concluderend kan daarom gesteld worden dat verpleegkundigen als beoordelaars van SIMP over het geheel genomen beter voldoen dan artsen.

Overeenstemming tussen beoordelaars is slechts één van de facetten die de betrouwbaarheid van een meetinstrument bepalen. De betrouwbaarheid van een test kan worden gedefinieerd als de mate waarin het herhalen van metingen met die test tot hetzelfde resultaat leidt. Verschillen tussen herhaalde metingen kunnen veroorzaakt worden door diverse facetten van de testsituatie, zoals respondenten, beoordelaars, casus en items.

Generaliseerbaarheidstheorie biedt een kader, waarin al deze facetten gelijktijdig geanalyseerd kunnen worden.[17-19] Met behulp van variantie- analyse wordt bepaald hoe groot de bijdrage van elk facet aan de totale variantie is. Vervolgens kan in een zogenaamde D-studie (D = decisie) geschat worden wat de relatieve invloed van de verschillende variabelen is op de betrouwbaarheid. In Tabel 1 worden de resultaten van een D-studie naar het facet beoordelaars weergegeven. Bij deze analyse is gebruik gemaakt van de gegevens van een tweetal eerdere studies waarin het facet beoordelaars is opgenomen.[16]

Tabel 1.
D-studie beoordelaars

	generaliseerbaarheids coëfficiënt	
aantal casus	**een beoordelaar**	**twee beoordelaars**
10	.85	.85
20	.91	.92
30	.94	.95
40	.96	.96
50	.96	.97

Het toevoegen van een tweede beoordelaar blijkt nauwelijks verschil te maken voor de generaliseerbaarheid van SIMP. Zeker als in aanmerking genomen wordt dat generaliseerbaarheidscoëfficiënten over het algemeen wat lager uitvallen dan klassieke betrouwbaarheidsmaten, is de betrouwbaarheid van de SIMP-test alleszins bevredigend te noemen.

De betrouwbaarheid van de test als geheel is nader onderzocht in een D-studie, waarbij de gegevens van een zevental verschillende onderzoeken met SIMP zijn gecombineerd, door het poolen van de overeenkomstige variantiecomponenten.[19] In Tabel 2 is de generaliseerbaarheid van SIMP uitgezet als een functie van testlengte, waarbij is uitgegaan van een gemiddelde van 10 casus per uur.

Het gecombineerde resultaat van de zeven onderzoeken, waarbij in totaal 148 personen van verschillende opleidingsniveaus betrokken waren, geeft aan dat een betrouwbaar toetsresultaat gerealiseerd kan worden in een toetstijd van ongeveer twee uur.

Tabel 2.
D-studie testlengte

test tijd	**generaliseerbaarheids coëfficiënt**
1 uur	.70
2 uur	.83
3 uur	.88
4 uur	.90
5 uur	.92

VALIDITEIT

De validiteit van een toets heeft te maken met de vraag in hoeverre de toets datgene meet waarvoor hij bedoeld is. Om uit te maken of een toets aan het doel beantwoordt, kunnen uiteenlopende criteria gehanteerd worden. Hierdoor ontstaat een groot aantal verschillende soorten validiteit. Van Berkel geeft een overzicht waarin niet minder dan 87 verschillende validiteiten worden onderscheiden.[20]

Wat betreft meting van medisch probleemoplossen is het belangrijkste probleem dat er geen universeel geldig criterium van adequaat medisch probleemoplossen bestaat. Het ontbreken van een dergelijke algemeen geaccepteerde "gouden standaard" maakt het onmogelijk eenduidige criteria te formuleren waar studenten aan moeten voldoen. Als ervaren artsen een vraag verschillend beantwoorden, kan voor studenten niet één enkel antwoord als juist gelden.

Het is dus niet mogelijk de validiteit van een instrument te bepalen door vergelijking met een ideale criteriummaat. In plaats daarvan moet ondersteuning voor de validiteit worden afgeleid uit samenhang met andere metingen, die zelf mogelijk tekort schieten, en voorspellingen ten opzichte van gebrekkig gedefinieerde criteria.

De validiteit van SIMP is onderzocht door na te gaan hoe de toets correleert met concurrerende metingen.[21] Steun voor de validiteit van SIMP werd gevonden in de significante correlatie met een globale beoordeling van een Simulatie Patiënt contact (SP). Verdere ondersteuning voor de validiteit van SIMP werd gevonden in onderzoeken, waarbij SIMP is toegepast als een van de meetinstrumenten. Zo werd de samenhang tussen SIMP-scores en beoordelingen van prestaties met een Simulatie Patiënt nader geanalyseerd door Crijnen et al.[22] Onderdelen van de SIMP bleken hoog te correleren met overeenkomstige elementen van een instrument voor het beoordelen van medische gespreksvaardigheid.

De overeenstemming tussen feitelijke prestaties in de praktijk en meting met SIMP is onderzocht door Rethans en Van Boven.[23] In grote lijnen stemden de resultaten van de verschillende metingen overeen. Bij nadere analyse bleken de prestaties van artsen in werkelijkheid echter beter te zijn dan gesuggereerd door het antwoord op de schriftelijke simulatie. De artsen bleken niet alles op te schrijven wat ze wisten. De schriftelijke toets geeft daardoor mogelijk een onderschatting van de werkelijke capaciteiten van ervaren artsen. Dit effect kan wellicht verklaard worden vanuit het gegeven dat ervaren artsen hun conclusies weergeven, zonder alle tussen-stappen expliciet te vermelden. Studenten die een toets maken, doen uiteraard hun best zoveel mogelijk op te schrijven.

Een laatste aspect betreft de face validiteit, ofwel de indruk van respondenten ten aanzien van de geschiktheid van het instrument voor het beoogde meetdoel. Wat dit betreft rapporteert Van Leeuwen een positieve beoordeling van SIMP door respondenten.[24] De studenten zien SIMP over het algemeen als een welkome aanvulling op de gangbare objectieve kennistoetsing.

SAMENVATTING EN CONCLUSIE

Bij de ontwikkeling van SIMP als methode voor het toetsen van vaardigheid in het oplossen van medische problemen is gekozen voor open vragen. De score-sleutels waarmee de antwoorden worden nagekeken vormen de neerslag van het gecombineerde oordeel van verschillende ervaren artsen. Het beoordelen van de antwoorden met deze score-sleutels is een relatief eenvoudige herkenningstaak.

Uit onderzoek naar de beoordelaarsbetrouwbaarheid kwam een hoge mate van overeenstemming tussen verschillende beoordelaars naar voren. Het resultaat van een generaliseerbaarheidsanalyse geeft aan dat met één beoordelaar kan worden volstaan. Opvallend was dat vooral verpleegkundigen de beoordelingstaak zeer betrouwbaar bleken te kunnen uitvoeren. De onderlinge overeenstemming tussen verpleegkundigen in de rol van beoordelaar was zelfs hoger dan die tussen artsen.

De tekortkomingen van bestaande instrumenten voor het meten van medisch probleemoplossen lijken vooral samen te hangen met het geringe aantal casus per toetsafname. Casus blijken zich te gedragen als items in een toets. Zowel wat betreft de meet-betrouwbaarheid als wat betreft de validiteit dient een toets derhalve uit een voldoende aantal verschillende casus te bestaan. Met SIMP wordt dit probleem ondervangen. Binnen één uur toetstijd kunnen ongeveer tien verschillende

casus beantwoord worden. Uit de generaliseerbaarheidsanalyse kwam naar voren dat een acceptabele betrouwbaarheid gerealiseerd kan worden met een toetstijd van ongeveer twee uur.

De validiteit van SIMP als operationalisatie van het construct medisch probleemoplossen wordt door een aantal onderzoeken ondersteund. Met name de samenhang met beoordelingen van Simulatie Patiënt contacten en met medisch handelen in de praktijk is bemoedigend. Verschillende aspecten van de validiteit moeten echter nog nader worden onderzocht. Zo dient de geldigheid van de aanname dat de meting beperkt kan blijven tot de eerste momenten van het arts-patiënt contact te worden getoetst. Ook verdere exploratie van de relatie tussen SIMP en medische kennis is van belang. Hierbij kan worden aangesloten op recent onderzoek, waaruit blijkt dat de verschillen in probleemoplossende prestaties tussen medische experts en novicen beter verklaard kunnen worden op grond van verschillen in de structuur van medische kennisbestanden dan op grond van verschillen in strategische benadering van het probleem.[25 26] Over het geheel genomen lijkt SIMP als toetsmethode een bijdrage te kunnen leveren aan het verschaffen van betrouwbare informatie ten aanzien van het vermogen van studenten tot het oplossen van problemen in de praktijk. In elk curriculum waar voorbereiding op de beroepspraktijk een belangrijke rol speelt, is nauwkeurige informatie op dit gebied van groot belang.

LITERATUUR

1. Wijnen WHFW, Van der Vleuten CPM. Toetsing: hordenloop of voortgangscontrole? Universiteit en Hogeschool 1985; 31: 6.
2. Van der Vleuten CPM, Van Luyk SJ. Evaluating undergraduate training in medical skills. Paper presented at the symposium on the evaluation of innovative curricula for the health siences. Ismaila, Egypt, 1985.
3. Wakeford R, Bashook P, Jolly B, Rothman A, eds. Directions in clinical assessment. Report of the first Cambridge Conference, Cambridge, University School of Clinical Medicine, 1985.
4. Norman GR, Feightner JW. A comparison of behavior on simulated patients and patient management problems. J Med Educ 1981; 15: 26-32.
5. McGuire C. Simulation technique in the teaching and testing of problem solving skills. Journal of Research in Science Teaching 1976; 13(2): 89.
6. Elstein AS, Schulman LS, Sprafka SA. Medical problem-solving: an analysis of clinical reasoning. Cambridge, Massachusetts: Harvard University Press, 1978.
7. Norcini JJ, Swanson DB, Gross LJ, Webster GD. A comparison of several methods for scoring patient management problems. Proceedings of the twenty second conference on research in medical education. Washington DC, 1983.
8. Norman GR. Objective measurement of clinical performance. J Med Educ 1985; 19: 43-7.
9. Newble DI, Hoare J, Baxter A. Patient management problems: issues of validity. J Med Educ 1982; 16: 137-42.
10. Knox JDE. The modified essay question. Booklet no. 5. Association for the Study of Medical Education: Dundee, 1975.
11. Feletti GI, Smith EMK. Modified essay questions: are they worth the effort? J Med Educ 1986; 20: 126-32.
12. Grant I, Marsden P. The structure of memorized knowledge in students and clinicians: an explanation for diagnostic expertise. J Med Educ 1987; 21: 92-8.
13. De Graaff E. Simulation of initial medical problem-solving: a test for the assessment of medical problem-solving. Medical Teacher 1988; 10 (1): 49-55.
14. Frederiksen N. The real test bias; influences of testing on teaching and learning. American Psychologist 1984; 3: 193-202.
15. Weed LL. Medical records, medical education and patient care. Year book. Chicago: Medical Publishers Inc., 1969.
16. De Graaff E. Simulation of initial medical problem-solving: studies on a new measure of medical problem-solving ability. Haarlem: Thesis, 1989. 109 blz. Proefschrift.
17. Cronbach LJ, Glaser GC, Nanda H, Rajaratnam N. The dependability of behavioral measurements. New York: Wiley, 1972.
18. Thorndike RL. Applied psychometrics. Boston: Houghton Mifflin Comp., 1982.
19. Brennan RL. Elements of generalizability theory. Iowa City, Iowa: American College Testing Program, 1983.
20. Van Berkel HJM. De diagnose van toetsvragen. Amsterdam: Centrum voor Onderzoek van het Wetenschappelijk Onderwijs, 1984. Proefschrift.
21. De Graaff E, Post GJ, Drop MJ. Validation of a new measure of clinical problem-solving. J Med Educ 1987; 21: 213-18.
22. Crijnen AAM, Post GJ, Kraan HF, Van der Vleuten C, Imbos T, Zuidweg J. Interviewing skills and medical competence. In: Kraan HF, Crijnen AAM. The Maastricht history-taking and advice checklist. Amsterdam, 1987. Proefschrift.
23. Rethans JJE, Van Boven CPA. Simulated patients in general practice: a different look at the consultation. Br Med J 1987; 294: 809-12.
24. Van Leeuwen Y. Toetsing medisch probleemoplossen: een oplosbaar probleem? Maastricht: Rijksuniversiteit Limburg, 1988.
25. Schmidt HG, Hobus PPM, Patel VL, Boshuizen HPA. Contextual factors in the activation of first hypotheses: expert-novice differences. Paper presented at the AERA-conference, Washington DC, 1987.
26. Boshuizen HPA, Schmidt HG, Coughlin LD. On the application of medical science knowledge in clinical reasoning: implications for structural knowledge differences between experts and novices. Paper presented at the 10th annual conference of the Cognitive Science Society. Montreal, Canada, 1988.