Simulation of initial medical problem-solving : studies on a new measure for the assessment of medical problem-solving ability

SUMMARY

Measurement of Medical Problem-solving

Medical problem-solving is often designated as an important general goal of medical education. Therefore, measurement of medical problem-solving abilities should be represented in the assessment of medical students. The general objective of medical education is to train physicians, who are capable of functioning in several different positions within our modern health care system.

At the Rijksuniversiteit Limburg a special instrument has been developed for the measurement of increase of medical knowledge: the Maastricht Progress Test. As extension of the Progress Test, a measure was needed for the assessment of medical problem-solving abilities.

Research with existing instruments, like the Patient Management Problem (PMP) and the Modified Essay Question (MEQ), indicated these measures suffered from several weaknesses. The results of PMPs, for instance, were found to be biassed by a cueing effect of the optional answers. Furthermore, the consistent finding of a low correlation among cases casted doubt on the validity of the measures. Therefore it was decided to initiate a project for the development of a new measure for the assessment of medical problem-solving.

Simulation of Initial Medical Problem-solving

Basic research into the nature of medical problem-solving has indicated that the ability to formulate initial hypotheses, directly in the beginning of the contact has crucial impact on the rest of the encounter. Therefore, it seemed sensible to focus instrument development on the initial moments of the patient-physician encounter. Furthermore, the open-ended question format seemed a logical means to avoid the cueing-effect of optional answers.

Based on these principles a measure was developed, called Simulation of Initial Medical Problem-solving. The instrument consists of short case histories, followed by one single open-ended question: "What would you do as a physician in this situation?" This question puts almost no constraints on the respondents. The responses are only limited by the information from the presented cases. Feed-back is given after the test is completed. The answering of one question takes five to ten minutes. Hence, a two hour test may contain ten to twenty different cases.

Objectivity of measurement

A well known disadvantage of open-ended questions, is the subjective influence of raters. In order to reduce the effects of rater bias, scoring-models have been developed. These scoring-models take the format of checklists, describing elements of the correct answer. To facilitate scoring, the scoring-models are organized according to the SOAP-system for medical audit (Weed, 1969). Respondents, however, remain free to chose their own answer structure.

With these scoring-models, raters are not expected to judge the value of an answer. All they have to do, is mark the similarities between a respondents answer and the scoring-model. This kind of judgement does not demand expert medical knowledge. It is sufficient, if a rater understand medical terminology, in order to recognize synonymous expressions. Since, nurses, by training and experience, fulfill these requirements, the reliability of the scoring system was investigated in an experiment, where six nurses scored 500 SIMP answers.

Interrater-reliability among these six nurses proved to be high. The correlation of one random selected nurse, with the mean of the population, was estimated .83 with Intra Class Correlation (ICC). Further analysis revealed some imperfections in the scoring-models, and suggested a lack of precision by one of the nurses. It was estimated that improvement of the scoring-models, and selection of raters on the characteristic of accuracy, could elevate the interrater reliability to .93. It was concluded that the scoring method may be regarded as objective.

Next the answers on four cases were also rated by two experienced physicians, with the same scoring-models. The overall agreement among all eight raters was again high. The inter-rater reliability was estimated .80. Only one case, with a defective scoring-model, showed a significant difference between the ratings by the nurses and by the physicians. Further analysis revealed, that the agreement among the nurses was stronger than that among the physicians. The relatively low agreement between the two physicians, may be explained by their expert knowledge. Their experience enables them to value the context of the whole answer, thereby the marking of items may be influenced. Differences in medical judgement than become visible in the scoring. Nurses, who do not have the expert knowledge to judge the context of an answer necessarily comply closer with the scoring-models. Despite some evidence that the nurses err in a few cases by lack of expert knowledge, it was concluded, that nurses meet the requirements of objective judgement better than physicians.

Reliability

Rating is only one of the facets that determine the reliability of a measure. Reliability is the degree to which a measure produces the same results, when it is repeated. In the repetition, facets like respondents, questions, and raters may be constant, or varied. Generalizability theory provides a framework that allows the analysis of all these facets at the same time.

The results of generalizability analyses on the data from the rater-reliability studies were disappointing. It was estimated that even a test of 30 cases would not reach satisfactory reliability.

However, in a next study a generalizability coefficient of .74 was found for a six-cases SIMP test, scored by four raters. This resulted in an acceptable reliability with a test consisting of 11 to 15 cases. The difference in result between the generalizability analyses can be explained partly by methodological problems. An alternative explanation is that the generalizability is domain specific.

Validity

The validity of SIMP was investigated by means of correlations with concurrent measures. Support for the validity of SIMP as a measure of medical problem-solving was found in a significant correlation with a global rating of performance with a simulated patient (SP). The insignificant correlation with assessment of medical knowledge by means of the Maastricht Progress Test was interpreted as circumstantial evidence of support. A high correlation would have been unfavorable, since that would mean that the problem-solving measure does not add information to the knowledge test.

The validity of SIMP was further supported by investigations of clinical competence, that applied SIMP as a measure. For instance, more detailed analyses of the relation between SIMP-scores and ratings with a simulated patient were reported by Crijnen et al (1987). Parts of SIMP were found to correlate highly with equivalent elements of their instrument for the assessment of medical interviewing skills.

In a study by Rethans and Van Boven (1987) scores on the written SIMP-test were compared with actual performance in medical practice. On the whole the results seemed equitable. Further analyses, however, revealed that the performance of physicians in practice was actually better than suggested by the written test.

The relation between SIMP-scores and scores on a knowledge test is further investigated by Van Leeuwen (1987). A distinct relationship could not be demonstrated. Van Leeuwen also reports a positive judgement on SIMP by the subjects in her investigation. The regard SIMP as a welcome extension to the objective knowledge tests.

With respect to further research on the validity of SIMP attention should focus on the question how to sample a representative test content (a test blueprint). Investigation of the relation of performance in actual practice could provide insight in the factors determining that performance. Also important is further exploration of the relation between SIMP-scores and factual medical knowledge. More insight in the structure of medical knowledge could provide clues toward the question whether

knowledge acts as a condition for problem-solving or if problem-solving ability is inherent in the structure of knowledge.

Conclusion

Application of SIMP in investigations on medical competence, as a concurrent or criterion measure seems justified.
On the whole, the research reported in this thesis supports the operationalization of the construct medical problem-solving with SIMP. It appears that reliable test can be constructed with 11 to 15 cases (a testing time of two and a half hours to three hours). There is, however, evidence suggesting that such a reliability can only be attained within a limited domain. Since that would imply that a much larger number of cases would be necessary for the assessment of "medical problem-solving", in general implementation of SIMP as an extension to the Progress Test seems inappropriate. As an alternative SIMP, or SIMP-like measures could be applied for assessment within the clinical clerkships. Analogous to the SIMP-test that was constructed for the family physician clerkship, versions could be adapted for almost any practical healthcare profession.
In such a way SIMP could contribute to the standardization of feed-back to students on their ability to handle practical problems in medical practice. Especially in a problem-based curriculum, aiming at preparation for practice, such feed-back is of utmost importance.

SAMENVATTING

Het meten van medisch probleemoplossen

Medisch probleemoplossen wordt vaak benadrukt als een belangrijke algemene doelstelling van medisch onderwijs. Een meting van vaardigheid in medisch probleemoplossen dient daarom deel uit te maken van de toetsing van medische studenten.
Aan de Rijksuniversiteit Limburg is voor het toetsen van medische kennis is een apart instrument ontwikkeld: de Maastrichtse Voortgangstoets. In aansluiting op deze Voortgangstoets bestond behoefte aan een instrument voor het meten van medisch probleemoplossen.
Bestaande instrumenten, zoals het Patient Management Problem (PMP) en de Modified Essay Question (MEQ) vertoonden echter tekortkomingen. Zo bleek een sturend effect van de antwoordopties de resultaten van het PMP te vertekenen. Verder gaf het regelmatig vinden van lage correlaties tussen casus onderling aanleiding tot twijfel ten aanzien van de validiteit van de metingen. Daarom werd besloten een nieuw instrument voor het meten van medisch probleem-oplossen te ontwikkelen.

Simulatie van Initieel Medisch Probleemoplossen (SIMP)

Onderzoek naar de aard van medisch probleemoplossen heeft uitgewezen, dat het vermogen direkt in het begin van het kontakt initiële hypothesen te formuleren van cruciaal belang is voor het verdere verloop van het kontakt. Daarom is bij de constructie van een nieuw instrument de nadruk gelegd op het begin van het arts-patiënt kontakt. Verder is gekozen voor open vragen, om het sturende effect van antwoordopties te vermijden.
Het op basis van deze uitgangspunten ontwikkelde instrument: Simulatie van Initieel Medisch Probleemoplossen (SIMP), bestaat uit korte beschrijvingen van casuïstiek, gevolgd door een enkele open vraag: "Wat zou u

doen, als u als arts in de praktijk met een dergelijk geval werd gekonfronteerd?" De respondenten zijn bij deze vraag vrij hun eigen formuleringen te kiezen. De antwoorden worden alleen beperkt tot de informatie die uit de gepresenteerde casus kan worden afgeleid. Feed-back wordt pas achteraf gegeven. De benodigde tijd per casus kan daarmee worden teruggebracht tot vijf à tien minuten. Binnen een toets kunnen daardoor tien tot twintig verschillende casus worden opgenomen.

Betrouwbaarheid van de beoordelingen

Een bezwaar van toetsing door middel van open vragen is de subjectieve invloed van de beoordelaars. Om dit effekt te ondervangen zijn voor de beoordeling antwoordsleutels ontwikkeld. Deze antwoordsleutels hebben de vorm van een checklist met omschrijvingen van elementen van een correct antwoord. Om het scoren te stroomlijnen zijn de items ingedeeld volgens het SOEP-schema van Weed (1969. De respondenten zijn echter vrij in het kiezen van een eigen struktuur voor hun antwoord. Van de beoordelaars wordt geen inhoudelijk waardeoordeel verwacht, maar alleen het markeren van overeenstemming tussen het antwoord en de antwoordsleutel. De beoordelaar hoeft dan ook geen medicus te zijn. Kennis van medische begrippen en terminologie is echter noodzakelijk om alternatieve formuleringen te kunnen herkennen. Aangezien verpleegkundigen op grond van hun opleiding en ervaring aan dit kriterium voldoen, is de betrouwbaarheid van het scoringssysteem onderzocht in een experiment, waarbij zes verpleegkundigen in totaal 500 antwoorden beoordeelden. De overeenstemming tussen deze beoordelaars bleek hoog te zijn. Met de Intraclass Correlatie Coefficient (ICC) werd de correlatie van beoordeling door één random gekozen verpleegkundige met het gemiddelde van de populatie verpleegkundige beoordelaars bepaald op .83. Nadere analyse bracht aan het licht, dat onvolkomenheden in enkele antwoordsleutels en slordigheid van een der beoordelaars de betrouwbaarheid negatief beïnvloedden. Geschat werd dat verbetering van de antwoordsleutels en selectie van beoordelaars de betrouwbaar-

heid verhoogd zou kunnen worden tot .93. Ten aanzien van de scoringsmethode werd gekonkludeerd, dat deze als objectief te beschouwen is. Vervolgens is onderzocht in hoeverre de beoordelingen van deze verpleegkundigen verschillen van beoordelingen door artsen, die beschouwd kunnen worden als inhoudelijke experts. Daarvoor zijn de antwoorden op vier casus opnieuw gescoord door twee artsen, met dezelfde antwoordsleutels. Over het geheel genomen was de overeenstemming tussen alle beoordelaars hoog. De beoordelaarsbetrouwbaarheid werd geschat op .80. Alleen bij een casus, waar al eerder manco's in de antwoordsleutel aan het licht gekomen waren, werd een significant verschil gevonden tussen de beoordelingen door de verpleegkundigen en die door de artsen. Nadere analyse bracht echter aan het licht dat de onderlinge overeenstemming tussen de twee artsen aanmerkelijk lager was dan die tussen de verpleegkundigen. De relatief lage overeenstemming tussen de twee artsen kan verklaard worden vanuit hun inhoudelijke deskundigheid. Doordat zij op grond van hun eigen ervaring een oordeel hebben over de kwaliteit van het gehele antwoord, kan dit oordeel mee van invloed zijn op het scoren van items. Verschillen in opvatting ten aanzien van medisch handelen kunnen dan tot uiting komen in de scoring. Verpleegkundigen, die niet een dergelijk eigen oordeel over de kwaliteit van het antwoord hebben, houden zich strikter aan de scoringsinstruktie. Ondanks aanwijzingen dat de verpleegkundigen, door gebrek aan inhoudelijke deskundigheid, in enkele gevallen tot een onjuist oordeel komen, kon daarom gekonkludeerd worden dat verpleegkundigen in dit aspekt beter voldoen als objektieve beoordelaars dan artsen.

Betrouwbaarheid

De beoordelaars vormen slechts een van de facetten, die van invloed zijn op de betrouwbaarheid van een meetinstrument. Betrouwbaarheid kan worden opgevat als de mate waarin een meting bij herhaling dezelfde resultaten oplevert. Facetten als respondenten, vragen en beoordelaars kunnen daarbij worden gevarieerd. Generaliseerbaarheidstheorie biedt een kader, waarin al deze facetten gelijktijdig geanalyseerd kunnen

worden. De resultaten van een generaliseerbaarheidsanalyse op basis van het materiaal van het beoordelaarsonderzoek waren teleurstellend. Een toets bestaande uit 30 casus was niet voldoende voor het realiseren van een acceptabele betrouwbaarheid. In een volgende studie werd echter voor een SIMP-toets bestaande uit zes casus, met vier beoordelaars een generaliseerbaarheidscoëfficiënt van .74 gevonden, resulterend in een acceptabele betrouwbaarheid bij 11 tot 15 casus. Voor een deel kan het verschil in uitkomst tussen de generaliseerbaarheidsstudies op methodologische gronden verklaard worden. Een alternatieve verklaring wordt gevormd door de mogelijkheid dat de generaliseerbaarheid domein specifiek is.

Validiteit

De validiteit van SIMP is onderzocht, door na te gaan hoe de toets correleert met concurrerende metingen. Steun voor deze concurrente validiteit van SIMP werd gevonden in de significante correlatie met een globale beoordeling van een Simulatie Patiënt kontakt (SP). De niet significante correlatie met de meting van medische kennis met behulp van de voortgangstoets, werd als indirekte ondersteuning geïnterpreteerd. Een hoge correlatie zou ongunstig zijn, aangezien dat zou betekenen dat de probleemoplostoets geen informatie aan de kennistoets toevoegt.

Verdere ondersteuning voor de validiteit van SIMP werd gevonden in onderzoeken van medische competentie, waarbij SIMP is toegepast als een van de meetinstrumenten. Zo werd de samenhang tussen SIMP-scores en beoordelingen van prestaties met een Simulatie Patiënt is nader geanalyseerd door Crijnen et al (1987). 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 (1987). In grote lijnen stemden de metingen overeen. Bij nadere analyse bleken de prestaties van artsen in werkelijkheid beter te zijn dan gesuggereerd

door het antwoord op de schriftelijke simulatie. De ervaren artsen in dit onderzoek bleken niet alles op te schrijven wat ze wisten.
Verder is door Van Leeuwen (1987) onderzoek uitgevoerd naar de samenhang tussen resultaten van een meting met SIMP en een kennistoets. In dit onderzoek kon geen duidelijke relatie tussen kennis en probleemoplossend vermogen worden aangetoond. De lage betrouwbaarheid van de kennistoets kan echter een storende invloed hebben gehad. De betrouwbaarheid van de in dit onderzoek gebruikte SIMP toets was opnieuw bevredigend te noemen. Verder rapporteert Van Leeuwen over beoordeling van SIMP door de respondenten in haar onderzoek. Deze zijn over het algemeen zeer positief en zien SIMP als een welkome aanvulling op de gangbare objectieve kennistoetsing.
Wat betreft verder onderzoek naar de validiteit van SIMP verdient met name de vraag hoe een inhoudelijk representatieve toets moet worden samengesteld (een toets blauwdruk) nadere aandacht. Onderzoek naar de relatie met prestaties in de praktijk zou meer inzicht kunnen verschaffen, in de factoren die deze prestatie bepalen. Van belang is ook verdere exploratie van de relatie tussen SIMP en medische kennis. In de eerste plaats kan daarbij gedacht worden aan onderzoek naar de relatie tussen prestaties op SIMP en medische kennis. Daarnaast is echter ook onderzoek naar de struktuur van die medische kennis van belang. Hiermee kan worden nagegaan in hoeverre kennis fungeert als voorwaarde voor het kunnen oplossen van medische problemen, dan wel dat de probleemoplosvaardigheid verankerd is in de kennisstruktuur.

Conclusie

In het verlengde van eerdere onderzoeken waarbij SIMP gebruikt is als meting van medisch probleemoplossen, zijn er mogelijkheden voor toepassing van SIMP als concurrerende, of als criteriummeting.
Over het geheel genomen wordt de geldigheid van de operationalisatie van het construct medisch probleemoplossen met SIMP door het tot nu toe uitgevoerde onderzoek ondersteund. Gebleken is dat met deze methode betrouwbare toetsen kunnen worden samengesteld van 11 tot

15 casus (een testtijd van twee en een half a drie uur). Er zijn echter aanwijzingen dat een dergelijke betrouwbaarheid alleen gerealiseerd kan worden binnen een beperkt inhoudelijk domein. Aangezien voor meting van een algemene trek "medisch probleemoplossen" een veel groter aantal casus nodig zou zijn (met aanmerkelijk langere testtijd) ligt invoering van SIMP in het kader van de voortgangstoets niet voor de hand. Een goed alternatief is de toepassing van SIMP of SIMP-achtige meetinstrumenten bij de beoordelingen in de klinische stages. Naar analogie van de SIMP-toets voor het PMOH op het gebied van de huisartsengeneeskunde kunnen voor nagenoeg elke professie binnen de gezondheidszorg aangepaste versies worden ontwikkeld.
Op die wijze zou SIMP kunnen bijdragen aan standaardisering van de feed-back aan studenten ten aanzien van hun vermogen tot het oplossen van problemen in de medische praktijk. Het geven van dergelijke hoogwaardige feed-back is van groot belang, met name in een probleem-gestuurd curriculum waar voorbereiding op de praktijk sterk wordt benadrukt.