

Auswertungsmodelle für Fallsimulationen

Ute Fuchs, Tübingen

Zusammenfassung

Das von McGuire eingeführte Auswertungsmodell wird dargestellt, und ein neues allgemeines Auswertungsmodell abgeleitet, das auf die Auswertungskategorien: Datenerhebung und Management (Auswertungsmodell 2) und auf die Auswertungskategorien: Diagnosefindung, Therapieplanung und Ganzheitsbetrachtung (Auswertungsmodell 3) angewandt wird. Exemplarisch werden die Ergebnisse der Fallsimulation Anastasia dargestellt. Eine Validierung der Auswertungsmodelle erfolgt anhand der Klausurleistung.

Abstract

A new model for the evaluation of formalized simulations of reasoning has been developed. It is characterized by individual evaluation categories, for which the sum of the weighted information is set to the maximal value of one. The clinical competence is defined by the information content. In comparison to McGuire's model the resulting scores are not compensating. An example of its application is given.

Einleitung

Seit 1961 werden Patient-Management-Problems (PMP) oder Fallsimulationen, in den schriftlichen Prüfungen des National Board of Medical Examiners und vieler anderer medizinischer Fakultäten als Prüfungsinstrument zum Messen klinischen Kompetenz anstelle der Prüfungen am Kranken-

bett eingesetzt. Die Reliabilität der linear konstruierten PMPs liegt zwischen .80 und .85, die Inhaltsvalidität ist unbestritten (Feightner, 1985).

Methode

An der Universitäts-Frauenklinik Tübingen wurden im Wintersemester 1986/1987 und im Sommersemester 1987 5 verschiedene, schriftliche, im latenten Druckverfahren hergestellte Fallsimulationen, die in Zusammenarbeit mit dem Institut für Didaktik der Medizin an der Universität Bonn entwickelt wurden, unter induktiven und deduktiven Lernbedingungen in die klinische Ausbildung im Fach Gynäkologie und Geburtshilfe integriert. Deduktive Lernbedingung bedeutet: von der allgemeinen Regel zum Fall. Das heißt, die Studenten haben einen Tag vor Bearbeitung der Fallsimulation das in der Fallsimulation behandelte Krankheitsbild im Kleingruppenunterricht durchgesprochen. Induktive Lernbedingung heißt, vom konkreten Fall zur Regel. Die Studenten bearbeiteten ohne eine spezielle Vorbereitung eine Fallsimulation. Die Klausur am Semesterende diente der Wissensprüfung und als Außenkriterium zur Validierung der Fallsimulationsleistungen.

Fragestellung

Wie läßt sich die bei der Bearbeitung einer Fallsimulation (PMP) erbrachte Leistung objektivieren? Was wird in einer Fallsimulation gelernt?

Auswertungsmodelle für Fallsimulationen

Auswertungsmodell 1 (McGuire)

Das in der Literatur verwendete Auswertungsmodell wurde von McGuire (1976) entwickelt. Leistung wird in den 4 klassischen Kategorien klinischer Kompetenz: Anamneseerhebung, klinische Untersuchung, Labor und Management beurteilt. Das Scoring System geht von +2 bis -2 für die einzelnen Items. Dabei bedeutet: +2 : dieses Item ist notwendig für die Fallbearbeitung, +1 : dieses Item ist hilfreich, 0: Routine-Item, -1: inadäquates Item und -2: schädliches Item.

Als Kennwerte werden berechnet und in Prozent angegeben:

Effizienz (E) entspricht dem Anteil nützlicher Wahlen an der Gesamtzahl getroffener Wahlen.

Leistung (L) ist definiert als die prozentuale Übereinstimmung mit einer Expertengruppe.

Auslaßfehler (A) sind die nicht gewählten notwendigen, und wichtigen Items.

Entscheidungsfehler (F) sind die Wahlen schädlicher oder nutzloser und inadäquater Items.

Allgemeine Kompetenz (K) ist das gewichtete Verhältnis von Effizienz und Leistung. $K = (L * E + L) : 2$. Nachteile dieses Auswertungsmodells sind die globale Beschreibung und die Möglichkeit der Kompensationen bei den Parametern Leistung und allgemeine Kompetenz.

Allgemeines Auswertungsmodell

Im folgenden soll ein neuentwickeltes Auswertungsmodell vorgestellt werden, bei dem die Bewertungsklassen nicht kompensierend sind, die Gewichte je Auswertungskategorie auf 1 normiert und die Kompetenz über den Informationsgehalt, das heißt Wissen, definiert werden.

Dabei wird der Begriff der "Nützlichkeit" einer vom Studenten getroffenen Entscheidung unter dem Zielaspekt der richtigen Diagnostizierung und adäquaten Therapie eingeführt. Ein repräsen-

tatives Maß für die Nützlichkeit der vom Studenten in freier Form getroffenen Wahl ist die dadurch gewonnene Information. Unter dem Gesichtspunkt der verschiedenen auszubildenden ärztlichen Fähigkeiten (Kategorien) wird die gesamte, je Kategorie K in n_K Frage-Antwort-Kombinationen enthaltene Gesamtinformation GI^K zunächst auf 1 normiert.

$GI^K \equiv 1$, $n_K =$ Anzahl der innerhalb einer Kategorie relevanten und informationstragenden Fragen-Antwort-Kombinationen.

Im allgemeinen ist eine beliebige Verteilung von GI^K auf n_K Fragen mit unterschiedlichen Gewichten G_{jK} , $j = 1, \dots, n_K$ möglich. Die Berechnung der die Qualifikation des Studenten kennzeichnenden Größen wie Kompetenz und Effizienz erfordern allerdings einen höheren Aufwand bei der Berechnung, da dann kein einfacher Zusammenhang zwischen der Anzahl der gestellten und relevanten Fragen und der aus ihnen gewonnenen Informationsmenge mehr besteht. Darüber hinaus ist eine unterschiedliche Bewertung der Wichtigkeit der relevanten Fragen auch im Verhältnis zueinander, nicht ohne weiteres medizinisch zu untermauern.

Im folgenden sei daher vorausgesetzt, daß sich die medizinisch bedeutsame Information je Kategorie, GI^K , gleichmäßig auf die n_K relevanten Fragen verteilt. Dann gilt:

$$GI^K = \sum_{j=1}^{n_K} I_j^K, \quad I_i^K = I_j^K = I^K \neq 0; \quad i \neq j.$$

Daraus folgt:

$$GI^K = \sum_{j=1}^{n_K} I_j^K = \sum_{j=1}^{n_K} I^K = n_K * I^K \quad \text{und}$$

$$I^K = \frac{GI^K}{n_K} = \frac{1}{n_K}, \quad GI^K = 1.$$

Jede für die Kategorie K relevante Frage trägt damit den Beitrag I^K zur Gesamtinformation GI^K bei.

Damit läßt sich der Indikator "Kompetenz für die Kategorie K", K^K , definieren als der vom Studenten durch Fragen gewonnene Anteil der Gesamtinformation:

$$\text{Kompetenz } K = \frac{\text{Gewonnene Informationsmenge}}{\text{Gesamtinformationsmenge}}$$

Die vom Studenten gestellte Anzahl N^K von Fragen zur Kategorie K setzt sich jedoch zusammen aus der Anzahl N_{1K} in K relevanter Fragen mit dem Informationsgehalt I^K und der Anzahl N_0^K Fragen mit Informationsgehalt Null:

$$N^{K*I^K} = N_0^{K*I^K} + N_1^{K*I^K} = N_1^{K*I^K},$$

$$N^K = N_0^K + N_1^K, \quad N_1^K \leq N^K, \\ N_0^{K*I^K} = 0 \text{ per definitionem.}$$

Damit ergibt sich für die Kompetenz

$$K^K = \frac{N_1^{K*I^K}}{N^{K*I^K}} = \frac{N_1^K}{N^K}; \quad N_1^K \leq n_K,$$

$$N_1^K \leq N^K, \quad 0 \leq K^K \leq 1.$$

Werden alle notwendigen, informationstragenden

Fragen gestellt, so gilt:

$$N_1^K = n_K \text{ und } K^K = \frac{N_1^K}{N^K} = \frac{n_K}{n^K} = 1$$

d.h., das Maximum der Kompetenz wird dann erreicht. Werden hingegen nur irrelevante Fragen gestellt, gilt entsprechend

$$N^K = N_0^K, \quad N_1^K = 0$$

und daher

$$K^K = N_0^{K*I^K} = 0$$

Gestattet also die vom Studenten wie oben definierte, erreichte Kompetenz einen Rückschluß auf Fähigkeiten, wie Sorgfalt und Umsicht im ärzt-

lichen Handeln und Denken über die zu verwendenden Heuristiken zur Eingrenzung des Problemraumes, so ist damit über die Ökonomie der gewählten Vorgehensweise noch nichts gesagt. Grundsätzlich wäre ja der Fall denkbar, daß die Kompetenz erst nach sehr vielen Fragen einen bestimmten Wert W , $0 \leq W \leq 1$, erreicht, sofern hinreichend viele irrelevante Fragen zur Auswahl stünden, was unter dem Gesichtspunkt der Vermittlung realistischer, der klinischen Praxis entsprechender Lerninhalte nicht befriedigen kann. Ein Maß für die Ökonomie der gewählten Vorgehensweise wäre danach eine Kosten-Nutzen Relation, bezogen auf die dabei gewonnene Informationsmenge. Unter Beibehaltung der bisher festgelegten Nomenklatur läßt sich der Indikator "Effizienz" als Verhältnis der Anzahl gestellter relevanter Fragen zur Gesamtzahl gestellter, relevanter und irrelevanter, Fragen definieren:

$$E^K = \frac{N_1^K}{N_K} = \frac{N_1^K}{N_0^K + N_1^K}.$$

Dieser allgemeine Auswertungsansatz läßt sich auf 2 Kategorien klinischer Kompetenz: Datenerhebung und Management anwenden. Hierbei ist ein Scoring aller Items dahingehend nötig, ob sie relevant oder irrelevant für die Lösung des Falles sind (Auswertungsmodell 2).

Eine Anwendung auf weitere 3 - 5 Denkopoperationen, die in der Behandlungssituation relevant sind, wie Diagnosefindung, Therapieplanung und Ganzheitsbetrachtung ist möglich. Hierbei ist ein Scoring der Items nach Nützlichkeit für die einzelnen Kategorien nötig (Auswertungsmodell 3).

Ergebnisse:

Tabelle 1 zeigt exemplarisch eine Auswertung der Fallsimulation Anastasia nach dem Auswertungsmodell 1 von McGuire.

Tab. 1: Auswertung der Fallsimulation Anastasia entsprechend dem Auswertungsmodell 1 von McGuire.

| Auswertungsmodell 1 | Lernbedingung | | | | | |
|---------------------|-----------------|-----------|------|-----------------|-----------|------|
| | deduktiv (n=24) | | | induktiv (n=27) | | |
| | Median | \bar{x} | SD | Median | \bar{x} | SD |
| Effizienz | 0.52 | 0.52 | 0.07 | 0.58* | 0.58 | 0.09 |
| Leistung | 0.50 | 0.48 | 0.11 | 0.44 | 0.44 | 0.15 |
| Auslaßfehler | 0.34 | 0.35 | 0.09 | 0.41 | 0.41 | 0.15 |
| Entscheidungsfehler | 0.16 | 0.16 | 0.07 | 0.16 | 0.16 | 0.05 |
| Allg. Kompetenz | 0.37 | 0.37 | 0.10 | 0.36 | 0.34 | 0.13 |

*Wilcoxon-Test, p=.045.

Tab. 2: Auswertung der Fallsimulation Anastasia nach dem Auswertungsmodell 2

| | Lernbedingung | | | | | |
|---------------|-----------------|-----------|------|-----------------|-----------|-------|
| | deduktiv (n=24) | | | induktiv (n=27) | | |
| | Median | \bar{x} | SD | Median | \bar{x} | SD |
| Datenerhebung | | | | | | |
| Effizienz | 0.84 | 0.84 | 0.06 | 0.84 | 0.85 | 0.07 |
| Kompetenz | 0.57 | 0.57 | 0.09 | 0.49 | 0.45 | 0.15* |
| Management | | | | | | |
| Effizienz | 0.67 | 0.69 | 0.19 | 0.60 | 0.61 | 0.18 |
| Kompetenz | 0.60 | 0.57 | 0.15 | 0.60 | 0.56 | 0.12 |

Wilcoxon-Test, * p=.0073 (α *=.0175)

Die studentische Leistung gemessen an der Effizienz liegt bei 52% unter deduktiven und bei 58% unter induktiven Lernbedingungen. Unter der deduktiven Lernbedingung erzielten die Studenten eine Leistung von 48% und unter der induktiven von 44%. Die Auslaßfehler liegen bei 35 bzw. 41%. Entscheidungsfehler werden jeweils in 16% gefunden, die allgemeine Kompetenz lag bei 37% unter der deduktiven und bei 34% unter der induktiven Lernbedingung.

Tabelle 2 zeigt für die Fallsimulation Anastasia

die Auswertung nach dem Auswertungsmodell 2: Die Studenten erreichten eine signifikant höhere Kompetenz in Datenerhebung unter der deduktiven als unter der induktiven Lernbedingung. Die Effizienz in Datenerhebung betrug .84 unter der deduktiven und .85 unter der induktiven Lernbedingung. Im Management erzielten die deduktiv lernenden Studenten eine Kompetenz von .57 und eine Effizienz von .69 und die induktiv lernenden Studenten eine Kompetenz von .56 und eine Effizienz von .61.

Tab. 3: Auswertung der Fallsimulation Anastasia entsprechend dem Auswertungsmodell 3

| FS Anastasia Kompetenz in | Lernbedingung | | | | | |
|------------------------------|--------------------|-----------|------|--------------------|-----------|--------|
| | deduktiv (n=24) | | | induktiv (n=27) | | |
| | Median | \bar{x} | SD | Median | \bar{x} | SD |
| Diagnosefindung | 0.65 | 0.64 | 0.14 | 0.57 | 0.59 | 0.17 |
| Therapieplanung | 1.00 | 0.94 | 0.10 | 0.71 | 0.76 | 0.22** |
| Ganzheitsbetrachtung | 0.55 | 0.55 | 0.11 | 0.45 | 0.43 | 0.19* |

Wilcoxon-Test, *p=.004 ($\alpha^*=.025$). **p=.009 ($\alpha^{**}=.05$).

Eine weitere Anwendung des neuen Auswertungsmodells zeigt Tabelle 3.

Inbezug auf Diagnosefindung erzielten die Studenten unter der deduktiven Lernbedingungen eine Kompetenz von .64 und unter der induktiven von .59. Die deduktiv lernenden Studenten erreichten eine signifikant höhere Kompetenz in Therapieplanung und Ganzheitsbetrachtung als die induktiv lernenden Studenten. Ein Experte würde jeweils ein Kompetenz von 1 erzielen.

3. Validierung der Auswertungsparameter an der Klausurleistung

Zur Überprüfung der Auswertungsmodelle anhand eines Außenkriteriums wurden die Studenten, die eine Fallsimulation bearbeitet und die Klausur geschrieben haben, aufgrund ihres Klausurergebnisses in "gute", d.h. überdurchschnittliche, und "schlechte" Studenten eingeteilt. Schlecht heißt ein Klausurergebnis, das dem Mittelwert entspricht oder darunter liegt. Dann wurden Diskriminanzfunktionen mit den Arcus-Sinus transformierten Parametern der jeweiligen

Auswertungsmodelle berechnet. Durch die Arcus-Sinus Transformation der Prozentwerte wurde eine Approximation an die Normalverteilung erreicht. Die Diskriminanzfunktionen wurden bei Homogenität der Innerhalbgruppen-Kovarianzmatrizen aufgrund der gepoolten Kovarianzmatrix berechnet, bei fehlender Homogenität aufgrund der Innerhalbgruppen-Kovarianzmatrix. Das aufgrund der berechneten Diskriminanzfunktion ermittelte D^2 diente als Klassifikationskriterium zur Einteilung der Studenten in "gute" und "schlechte" aufgrund der Fallsimulationsleistung in Abhängigkeit vom Auswertungsmodell. Die so gewonnenen Klassifikationsverteilungen sind in Form von Vierfeldertafeln in Tabelle 4 aufgeführt.

Die aufgrund der Parameter des 1. Auswertungsmodells vorhergesagten Klassifikationen sind in Spalte 1, die aufgrund des 2. Auswertungsmodells vorhergesagten Klassifikationen in Spalte 2 und die aufgrund des 3. Auswertungsmodells in der Spalte 3 der Tabelle 5 eingetragen.

Zur Überprüfung, ob ein Modell mehr mit der

Tab. 4: Klassifikationen in gute und schlechte Studenten aufgrund der Klausur- und Fallsimulationsleistungen in den 3 Auswertungsmodellen der verschiedenen Fallsimulationen

| Fallsimulation | | Auswertungsmodell | | | | | |
|----------------|------|-------------------|-----------|-----|-----------|-----|-----------|
| | | gut | 1 schl | gut | 2 schl | gut | 3 schl |
| Margot | gut | 12 | 7 | 11 | 8 | 13 | 6 |
| | schl | 1 | 17 | 6 | 12 | 5 | 13 |
| Christine | gut | 13 | 9 | 13 | 9 | 13 | 8 |
| | schl | 9 | 24 | 15 | 18 | 10 | 22 |
| Liebchen | gut | 10 | 7 | 13 | 4 | 10 | 6 |
| | schl | 2 | 14 | 10 | 6 | 6 | 10 |
| Karoline | gut | 12 | 8 | 12 | 8 | 15 | 5 |
| | schl | 13 | 14 | 10 | 7 | 11 | 16 |
| Anastasia | gut | 10 | 8 | 12 | 6 | 12 | 6 |
| | schl | 6 | 18 | 7 | 17 | 6 | 18 |

Tab. 5: Überprüfung der Güte der 3 Auswertungsmodelle durch Chi²-Test über die Prozentsätze übereinstimmender Klassifikationen durch Klausur- und Fallsimulationsleistung

| Fallsimulation | Auswertungsmodell | | | Chi ² | FG |
|----------------|-------------------|-----|-----|------------------|----|
| | 1 | 2 | 3 | | |
| Margot | .78 | .62 | .70 | 2.37 | 2 |
| Christine | .67 | .56 | .66 | 1.68 | 2 |
| Liebchen | .73 | .58 | .62 | 1.72 | 2 |
| Karoline | .55 | .62 | .66 | 1.13 | 2 |
| Anastasia | .67 | .69 | .75 | 0.22 | 2 |
| gepoolt | .67 | .59 | .67 | 3.84 | 2 |

Klausurleistung übereinstimmende Klassifikationen erbringt als ein anderes, wurde der Prozentsatz übereinstimmender Klassifikationen für jedes Auswertungsmodell berechnet und mit Hilfe des Chi²-Tests auf Unterschiede geprüft. Es ergaben

sich keine signifikanten Unterschiede, wie Tabelle 5 zeigt.

Alle 3 Auswertungsmodelle lieferten annähernd gleich viele übereinstimmende Klassifikationen in bezug auf das Außenkriterium Klausurleistung.

Diskussion

Der Lernerfolg bei der Ausbildung zum Arzt ist erst nachweisbar, wenn der Arzt selbstverantwortlich sein Können bei der Betreuung von Patienten unter Beweis stellt (Renschler, 1987).

Eine Theorie der klinischen Kompetenz kann als Meßmodell definiert werden, das eine oder mehrere operational definierte Konstrukte enthält, die strukturell untereinander bezogen sind, und die Vorhersagen später zu erwartender klinischer Leistung maximieren.

Maatsch et al. (1986) testeten mit dem Programm LISREL, eine bestätigende Faktorenanalyse, die die Spezifikationen von Konstrukt-Komponenten und Annahmen über ihre Beziehungen voraussetzt, 4 Theorien klinischer Kompetenz an Feldtestdaten von 94 Personen.

Die Testbatterie umfaßte Multiple Choice Aufgaben, PMCQ (Pictorial Multiple Choice Items), PMPS (Patient Management Problems), SPEs (Simulated Patient Encounter) und SSEs (Simulated Situation Encounter).

Getestet wurde die Ein-Faktoretheorie von Ebel: alle Tests erfassen Wissen, die Zwei-Faktoretheorie von Senior 1976: Wissen ist schwach mit dem Leistungsfaktor Performance korreliert (Kompetenz = das, was ein Arzt tun kann, Performance = das, was ein Arzt aktuell in der praktischen Situation tut), die Drei-Faktoretheorie von Maatsch (1986): Komponenten der klinischen Kompetenz sind: Fähigkeiten, Fertigkeiten, Wissen- und das Fünf-Faktoren-Modell (Levine 1978, McGuire et al. 1976): Klinische Kompetenz in Anamnese, klinischer Untersuchung, Labordiagnostik, Diagnose und Therapie.

Nur das Drei-Faktoren-Modell konnte bestätigt werden. In seinem Modell messen PMPs Wissen, allerdings weit weniger effektiv als die MCQs und PMCQs (Betas: .55, .92, .91 respektive), außerdem haben sie die niedrigste Reliabilität.

Es ist sicher nicht sinnvoll, von der Bearbeitung einer Fallsimulation, deren Wesen es gerade ist, die Anpassung des Faktenwissens auf individuelle Gegebenheiten zu verlangen, auf das Können, das ein Arzt später einmal zeigen wird, zu schließen. Die Lösung einer Fallsimulation ist für eine Aussage über clinical judgment unzureichend (Bashook, 1976).

Datenerhebung und Management als Bewertungskriterien des ärztlichen Handelns zu betrachten basiert auf folgenden Untersuchungen. Juul et al. (1979) haben den Parameter Leistung in der Definition von McGuire für die Auswertung von 24 PMPs, die Studenten im junior year, und 26 PMPs, die von Studenten im senior year bearbeitet wurden, herangezogen und für folgende Kategorien berechnet: Anamnese, klinische Untersuchung, diagnostisches Vorgehen (Labor, Zusatzuntersuchungen), Lösungsweg, Therapie und Diagnose. Jedes Item der PMPs wurde auf einer Skala von +8 bis -8 von einem Expertengremium hinsichtlich seiner klinischer Wertigkeit beurteilt. Bei der faktorenanalytischen Auswertung wurden 2 Faktoren, die die PMP-Leistung charakterisieren, gefunden: Datenerhebung (Anamnese, klinische Untersuchung und diagnostisches Vorgehen) und Management. Diagnostisches Vorgehen lädt auf beiden Faktoren. Donnelly et al. (1974) haben ebenfalls eine faktorenanalytische Auswertung vorgenommen und 2 Faktoren gefunden, die sie Datenerhebung und Entscheidungstreffen nannten. Datenerhebung war hoch reliabel, während der 2. Faktor eine niedrige Reliabilität zeigte und als fallspezifisch angesehen wurde.

Datenerhebung und Management beschreiben 2 Aspekte ärztlichen Handelns bzw. Denkens. Andere Fähigkeiten, die im Rahmen einer Behandlungssituation von einem Arzt erwartet wer-

den, sind: Diagnosefindung, Therapieplanung, Ganzheitsbetrachtung, eigenes Lernen und eventuell Stadieneinteilung. Diese Fähigkeiten lassen sich in der Fallsimulation üben und auch messen, wie Auswertungsmodell 3 gezeigt hat. Auswertungsmodell 2 und 3 betrachten unterschiedliche Denkprozesse, unterschiedliche Kategorien klinischer Kompetenz, klinischer Urteilsfindung oder Problemlöseverhaltens, je nachdem, in welchen hypothetischen Konstrukten man ärztliches Können beschreiben will. Auswertungsmodell 2 und 3 sind keine rivalisierenden Auswertungsmodelle sondern verschiedene Betrachtungsweisen.

Die Güte der Auswertungsmodelle wurde über die %-Sätze übereinstimmender Klassifikationen aufgrund der Klausur und Diskriminanzfunktionen mit den Parametern der 3 Auswertungsmodelle getestet. Hierbei wurden übereinstimmende Klassifikationen zwischen .56 und .78 erreicht. Es ließen sich keine signifikanten Unterschiede in Abhängigkeit vom Auswertungsmodell feststellen. Gemeinsam messen die Klausur und die Auswertungsmodelle nur das Faktenwissen. Die fehlende Übereinstimmung ist wohl dadurch zu erklären, daß die Auswertungsmodelle verschiedene Kategorien ärztlichen Handelns erfassen.

Das in der Literatur verwendete Auswertungsmodell von McGuire zeigt Nachteile. Es ist global und läßt Kompensationen in der Leistung und Kompetenz zu. Das vorgestellte Modell hat Bewertungsklassen, die nicht kompensierend sind, die Gewichte der Auswertungskategorien sind auf 1 normiert und die Kompetenz über den Informationsgehalt definiert.

Literatur

Bashook, P.G.: A conceptual framework for measuring clinical problem-solving, in: *J. Med. Educ.* 52 (1976), 109-114.

Donnelly, M.B., Gallagher, R.E., Hess, J.W., Hogan, M.J.: The dimensionality of measures derived from complex clinical situations. *Annu. Conf. Res. Med. Educ.* 13 (1974), 14-19.

Feightner, J.W.: Patient Management Problems, 183-200, in: Neufeld, V.R., Norman, G.R. (eds.): *Assessing clinical competence*, New York 1985.

Juul, D.H., Noe, M.J., Nerenberg, R.L.: A factor analytic study of branching patient management problems, in: *Med. Educ.* 13 (1979), 199-203.

Levine, H.G.: Selecting evaluation instruments, in: Morgan M., D. Irby (Hrsg.): *Evaluating Clinical Competence in the Health Professions*: C.V. Mosby, St. Louis, Missouri 1978.

Maatsch, J.L., Huang, R.: An evaluation of the construct validity of four alternative theories of clinical competence. *Annu. Conf. Res. Med. Educ.* 25 (1986), 69-74.

McGuire, C.H. Solomon, L.M., Bashook, P.G.: Construction and use of written simulations. The Psychological corporation, New York 1976.

Renschler, H.E.: *Die Praxisphase im Medizinstudium*. Berlin 1987.

Senior, J.R.: *Toward the measurement of competence in medicine*. National Board of Medical Examiners, Philadelphia 1976.

Dr. med. Dipl.-Psych. Ute Fuchs
Universitäts-Frauenklinik Tübingen
Schleichstr. 4
7400 Tübingen