Multiple Choice als Numerus clausus (3)

Den schriftlichen Ärztlichen Prüfungen fehlt die Gültigkeit

Horst Kuni und Peter Becker (Marburg)

Die Feststellung, dass das System der Multiple-Choice-Prüfungen in eine "grundgesetzwidrige Handhabung abgeglitten" sei, weil mit seiner Hilfe die Numerus-clausus-Politik "mit anderen Mitteln" fortgesetzt werde (Heft 4/80, Seite 194ff.), teilten die Autoren im zweiten Beitrag dieser Artikelserie (Heft 5/80, Seite 292ff.) als Ergebnis ihrer Untersuchungen mit, dass zu viele Medizinstudenten am Messfehler des Prüfungsinstruments scheitern.

Die Veröffentlichung ihrer Überlegungen und Argumente erfolgt in der Monatsschrift des Marburger Bundes, weil der Verband an der Reform des Medizinstudiums nicht unwesentlich mitgewirkt hat und weil das neue Prüfungssystem bei der Einführung 1970 auch von ihm vor allem wegen seiner "konkurrenzlosen Objektivität" begrüßt wurde.

In unserem vorangegangenen Beitrag [14] haben wir gezeigt, dass viele Medizinstudenten am Messfehler der schriftlichen Prüfungen scheitern, da diese Tests keine absolute Zuverlässigkeit besitzen (können) und die Reliabilität fälschlich von der Bestehensregel nicht berücksichtigt wird. Mit Ausnahme des Dritten Abschnitts der Ärztlichen Prüfung war die Zuverlässigkeit aber grundsätzlich für die Zwecke einer individuellen Prüfung nach den Maßstäben der "klassischen" Testtheorie ausreichend. (Siehe Tabelle 1, S. 7 des Abschnittes (2)).

Nun ist die Zuverlässigkeit eines Tests zwar eine notwendige, aber keine hinreichende Voraussetzung. Der Test muss vor allem auch gültig (valide) sein. Fragen wir zunächst nach der Kriterienvalidität. Sie drückt aus, inwieweit das Testergebnis es erlaubt, auf ein gesetztes Kriterium zu schließen, wobei die beste Validierung an einem so genannten Außenkriterium erfolgt [1, 15]. Hier wäre das die spätere "Bewährung im ärztlichen Beruf". Es braucht sicher nicht weiter ausgeführt zu werden, dass dieses Außenkriterium kaum allgemeingültig zu definieren und zu messen ist [7, 16].

Korrelative Vergleiche

Eine Abschätzung der Kriterienvalidität erhält man durch so genannte innere Validierung [15]. Hierzu kann man zum Beispiel aufeinander folgende Prüfungsabschnitte miteinander vergleichen. Die Validität wird dabei als Korrelationskoeffizient r ausgedrückt, der hier zwischen 0 und 1 zu erwarten ist. $r = 0$ bedeutet, dass das Ergebnis der ersten Prüfung keinerlei Zusammenhang mit dem der

folgenden aufweist, r = 1 besagt, dass aus der ersten Prüfung das Ergebnis der zweiten Prüfung mit Sicherheit vorhergesagt werden kann. Zwischenwerte veranschaulicht man sich am besten durch das so genannte Bestimmtheitsmaß (r^2), das ausdrückt, zu welchem Anteil das eine Ergebnis durch das andere erklärt wird.

Das IMPP hat solche korrelativen Vergleiche angestellt (Tabelle 1) [8]. Ein r von ca. 0,7 mag für statistische Zwecke ein befriedigendes Ergebnis darstellen, für individuelle Aussagen ist es unzulänglich. Das zeigt schon das Bestimmtheitsmaß, nach dem sich die Ergebnisse einer Prüfung generell nur zur Hälfte aus den Fähigkeiten erklären lassen, die die vorangegangene Prüfung gemessen haben will.

Tab. 1: Korrelationskoeffizient ® und Bestimmtheitsmaß der Korrelation in Prozent (r^2 * 100) beim Vergleich der Prüfungsergebnisse untereinander von jeweils N Kandidaten (aus [8]).

		1 Abschnitt	2 Abschnitt	3 Abschnitt
		der Ärztlichen Prüfung		
Ärztliche Vorpüfung	r r^2 N	0,77 45,00 v. H. 3645	0,70 48,30 v. H. 1263	Mündl. Prüfung n. d. Übergangs- bestimmungen
	r r^2 N		0,75 55,95 v. H. 2897	0,72 51,72 v. H. 811
	r r^2 N			0,80 64,32 v. H. 811

Nun sagt ein statistisches Durchschnittsmaß recht wenig über individuelle Einzelfälle aus. Wir müssen deshalb die Verteilung der einzelnen Kandidaten prüfen, bevor wir das Bestimmtheitsmaß auf den Einzelfall übertragen.

Ein Blick auf Tabelle 2 zeigt, dass in der Regel mehr als die Hälfte der Kandidaten, die an einer (damals hypothetischen) 60 v. H.-Bestehensregel gescheitert wären, die Folgeprüfung auch mit der strengeren Bestehensregel bestanden hätten! Bis zu 20 v. H. haben sogar ein Ergebnis erzielt, das gleich oder besser als der Durchschnitt war!

Tab. 2: Weiteres Schicksal derjenigen Kandidaten, die zwar nach altem Recht eine Prüfung bestanden haben, aber an der 60 v. H.-Bestehensgrenze gescheitert wären (Auswertung aller bis dato vom IMPP als Abbildung publizierten Korrelationen [4-7])
Die Terminangaben in beziehen sich auf die Folgeprüfung, weil die verglichenen Kandidaten die vorhergehende Prüfung zum Teil erst nach einem größeren als dem in der AOÄ vorgeschriebenen zeitlichen Mindestabstand absolviert haben.
N: Zahl der verglichenen Kandidaten.
r: Vom IMPP berechneter Korrelationskoeffizient.
r^2: Bestimmtheitsmaß in Prozent.
>50 v. H. <60 v. H.: Zahl der Kandidaten, die bei hypothetischer Anwendung der heutigen Bestehensregel in der vorangegangenen Prüfung gescheitert wären.
>60 v. H.: Absolutzahl und relativer Anteil an der Zahl in Spalte 5 (in v. H.) der Kandidaten, die aber in der Folgeprüfung auch der 60 v. H.-Bestehensregel gewachsen gewesen wären.
> Durchschnitt: wie zuvor, jedoch Ergebnis der Folgeprüfung gleich oder besser als der Durchschnitt.
Die zahlen sind durch Auszählen graphischer Darstellungen des IMPP ermittelt worden, wobei sich durch das relativ grobe Raster des Maßstabs geringe Rundungsfehler ergeben können. Die Vorlagen des IMPP des Folgetermins März 1977 haben als Maßstab keine Rohwerte, sondern mit Mittelwert und Standardabweichung umgerechnete Standardwerte. Da 2828 (= 83,2 v. H.) der 3399 Kandidaten aus dem Ersten Abschnitt der Ärztlichen Prüfung die Ärztliche Vorprüfung im März 1976 abgelegt hatten und 1833 (= 86,8 v. H.) der Kandidaten aus der Zweiten Ärztlichen Prüfung den Ersten Abschnitt im März 1975 absolviert hatten, haben wir den 60 v. H.-Wert mit den Mittelwerten und Standardabweichungen dieser Termine in Standardwerte umgerechnet, um den Cutoff-Punkt zu erhalten.

Ärztliche Vorprüfung gegen Ersten Abschnitt der Ärztlichen Prüfung								
Termin der Folge-prüfung	N	r	r^2	> 50 v. H. < 60 v. H.	davon in der Folgeprüfung			
					< 60 v. H.		≥ Durchschnitt	
Aug. 1975	1551	0,78	60,8 v. H.	231	117	50,7 v. H.	17	7,4 v. H.
März 1976	2714	0,70	49,0 v. H.	553	274	49,6 v. H.	119	21,5 v. H.
Aug. 1976	3034	0,74	54,8 v. H.	670	436	65,1 v. H.	110	16,4 v. H.
März 1977	3399	0,70	49,0 v. H.	863	440	51,0 v. H.	137	15,9 v. H.
Erster Abschnitt gegen Zweiten Abschnitt der Ärztlichen Prüfung								
Termin der Folge-prüfung	N	r	r^2	> 50 v. H. < 60 v. H.	davon in der Folgeprüfung			
					< 60 v. H.		≥ Durchschnitt	
Aug. 1976	876	0,75	56,3 v. H.	72	41	56,9 v. H.	3	4,1 v. H.
März 1977	2111	0,76	57,8 v. H.	278	222	79,9 v. H.	15	5,4 v. H.

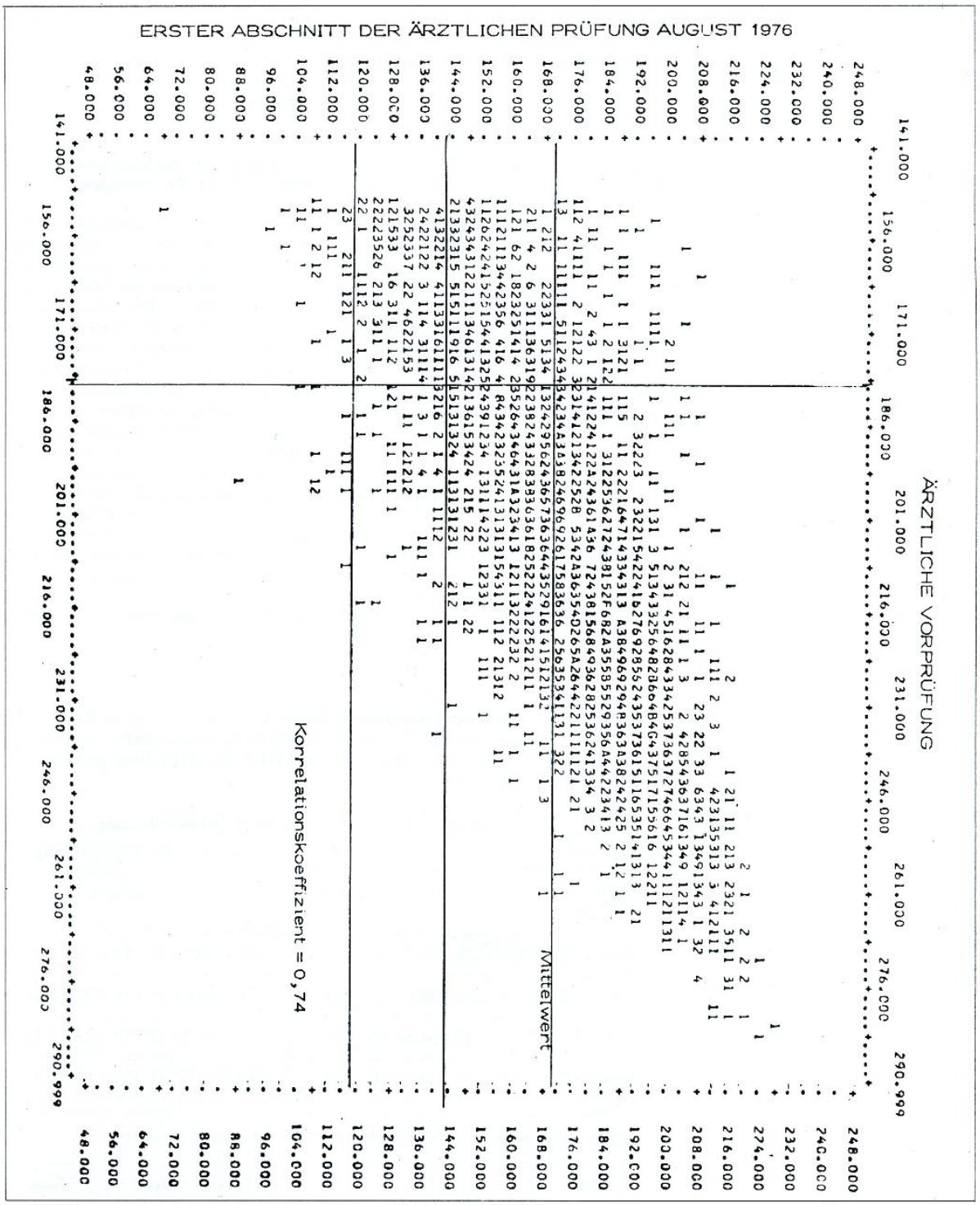

Abb. 1: Wiedergabe der Original-Korrelation des IMPP aus [6] zwischen den Prüfungsergebnissen des Ersten Abschnittes der Ärztlichen Prüfung im August 1976 und der Ärztlichen Vorprüfung von 3034 Kandidaten. Die Symbole geben die Zahl der Kandidaten mit gleichen Koordinaten an, wobei Ziffern von 10 und aufwärts durch die folgenden Buchstaben des Alphabets repräsentiert werden.
Waagrechte Linie: 60 v. H.-Bestehensgrenze der Ärztlichen Vorprüfung
Senkrechte Linien: Unter dem Mittelwert: 60 v. H.-Bestehensgrenze,
darunter 50 v. H.-Bestehensgrenze

Der Leser könnte vielleicht meinen, die Tabelle 2 enthalte Druckfehler oder den Autoren sei ein Auswertefehler unterlaufen. Wir haben deshalb eine repräsentative Original-Korrelation als Abbildung 1 beigefügt. Daran erkennt man deutlich: Die Korrelation wird fast ausschließlich durch die Spitzengruppe erzeugt, im Durchschnittsbereich und darunter geht sie praktisch völlig verloren! Gerade in diesem Bereich, wo durch Nichtbestehen der Prüfung über Einzelschicksale entschieden wird, ist aber Validität erforderlich und offenkundig nicht vorhanden.

"Niveaupflege" zu Lasten der Tauglichen?

Hier kann das Instrumentarium seine Geschichte nicht verleugnen. Seine wesentlichen Entwicklungsimpulse hat es in den USA während des Zweiten Weltkriegs zum Beispiel zur Selektion von Anwärtern für eine Pilotenausbildung erhalten [2, 3]. Wendet man den Test zur Selektion an, ist natürlich jede noch so geringe Validität ein Gewinn: Nehmen wir an, unter 10 000 Bewerbern um eine Aufgabe befänden sich 2 v. H. Untaugliche, also 200. 7000 möglichst geeignete sollen für die vorgesehene Tätigkeit ausgesucht werden. Ein Zufallszugriff, wie es auch ein Test ohne Validität darstellt, liefert unter den 7000 Selektierten 2 v. H. Untaugliche, also 140. Mit einem Test von einer Validität r = 0,7 erzielt man, dass nur noch 56 Untaugliche, also nur 0,8 v. H. der Selektierten, passieren. Die Mehrzahl der Untauglichen ist ausgesondert worden. allerdings befinden sich unter den Zurückgewiesenen 2856 Taugliche, also 29 v. H. aller Tauglichen.

Je schärfer die Selektion betrieben wird, um so mehr verschiebt sich unter den Ausgelesenen das Verhältnis zugunsten der Tauglichen, um so mehr Taugliche werden aber auch zu Unrecht ausgeschieden (weitere Beispiele dazu Tabelle 19, S. 293 in [2]).

Bei den Ärztlichen Prüfungen ist es natürlich ein verfassungswidriges Unterfangen, die mangelhafte Validität dadurch auszugleichen, dass eine "Niveaupflege" in den angedeuteten Dimensionen zu Lasten der Tauglichen betrieben wird.

Dieses erschreckende und deprimierende Resultat lässt die Frage nach der Konstruktvalidität aufkommen. Die Auseinandersetzung des IMPP mit diesem Begriff erschöpft sich in einer Mitteilung der Definition: Konstruktvalidität hat ein Test, wenn er Ergebnisse liefert, die in Richtung der Hypothesen liegen, die auf Grund des theoretischen Konzepts (Konstrukts) gebildet werden.

Der Vorprüfung fehlt die Konstruktvalidität

Das IMPP stellt zur Diskussion der Kriterienvalidität wörtlich fest:"…dass die schriftlichen medizinischen (gemeint sind die Ärztlichen; die Autoren) Prüfungen Lernleistungstests sind, die prüfen sollen, ob die erforderlichen Kenntnisse nach Abschluss eines Studienabschnitts vorhanden sind. Sie haben nicht die Aufgabe, die Berufsunfähigkeit der Medizinstudenten festzustellen"(!!). Auf diese Aussage werden wir noch zurückkommen.

In diesem Zusammenhang sei festgestellt, dass der Nachweis von Kenntnissen in einer Vorprüfung als Voraussetzung für die Erlaubnis, das Studium fortzusetzen, nur durch das theoretische Konzept legitimiert sein kann, dass bei Nichtbestehen die für eine erfolgreiche Fortsetzung des Studiums erforderlichen Grundkenntnisse nicht vorhanden sind. Denn nach dem Konstrukt einer Vorprüfung dürfte es einem von ihr als unfähig eingestuften Kandidaten nicht gelingen, die folgende Prüfung zu bestehen [12].

Die Ergebnisse aus Tabelle 2 belegen eindeutig, dass der Ärztlichen Vorprüfung die Konstruktvalidität fehlt, denn die Mehrzahl der nach der neuen Bestehensregel in der Ersten Prüfung ausgesonderten Kandidaten hätte die folgende Prüfung bestanden. Sollte eine nähere inhaltliche Überprüfung ergeben, dass der Erste Abschnitt der Ärztlichen Prüfung entgegen seiner Deklaration als Teil der Hauptprüfung inhaltlich und de facto (da von seinem Bestehen die Zulassung zum nächsten Prüfungsabschnitt abhängt) eine weitere Vor-Prüfung ist, muss man auch ihm die Konstruktvalidität absprechen.

Grenzzone zwischen Messtechnik und Didaktik

Wie man sieht, bewegt sich die klassische Testtheorie mit den Validitätsbegriffen an einer Grenzzone, wo sich Begriffe der testtheoretischen Messtechnik mit Fragen der Didaktik berühren [1]. Deshalb zieht man sich angesichts der dürftigen und auch nur aus innerer Validierung geschöpften Kriterienvalidität auf die inhaltliche Gültigkeit, die Kontentvalidität, zurück [7]. Stellen wir an dieser Stelle eine kritische Betrachtung der Prüfungsinhalte selbst zurück und setzen eine Menge von Prüfungsgegenständen voraus, von denen inhaltlich valide geprüft werden soll und kann, ob sie der Kandidat beherrscht.

Zur besseren Veranschaulichung dieses Zusammenhangs greifen wir hier auf das Binomialmodell zurück [13]. Es ist anwendbar, wenn eine Teilmenge von Fragen einem Prüfungsgegenstand gewidmet ist (Kontentvalidität), alle diese Prüfungsfragen eine vergleichbare Schwierigkeit besitzen und ihre Beantwortung voneinander unabhängig ist. Dann kann nämlich nach den Gesetzen der Wahrscheinlichkeitsrechnung präzise ermittelt werden, wie viele solcher Fragen vorgegeben werden müssen, damit man das Antwortverhalten mit der geforderten Sicherheit vom Rateverhalten dann unterscheiden kann, wenn der Kandidat das Lehrziel im geforderten Umfang erreicht hat.

Wir entnehmen den Tabellen von Klauer [13] folgendes Beispiel: Ratewahrscheinlichkeit bei einer aus fünf Auswahl-Antwortaufgaben: 0,2, mit Sicherheitszuschlag für schlechte Frageformulierung (offenkundig falsche Alternativen): 0,33. Dann muss man mindestens zwölf kontentvalide Fragen zu einem Lehrziel stellen, um mit 99 v. H. Sicherheit feststellen zu können, dass der Kandidat nicht geraten hat. Wenn er dabei neun (=75 v. H.) der Fragen richtig beantwortet hat, folgt daraus außerdem, dass das Lehrziel mit der gleichen Sicherheit zu 95 v. H. (!) erreicht wurde.

Erfüllen die Fragen nicht die Voraussetzungen des Binomialmodells, ändert sich das skizzierte Prinzip nicht wesentlich. Wegen der größeren Unsicherheit in der Vorabkalkulation wird man die Zahl der Fragen zu einem Gegenstand erhöhen müssen.

Exemplarische Stichproben ohne verbindliche Aussagen

Das Medical Board der USA, Vorbild des bundesdeutschen Prüfungssystems, stellt zunächst eine Auswahl von Gegenständen zusammen (das ist dann der eigentliche Gegenstandkatalog), die wesentlich erscheinen und zu denen eine ausreichende Zahl von Fragen zur Verfügung steht [3, 37]. Das IMPP stellt aber die Aufgaben definitiv so zusammen [4], dass des den größten Teil (zum Beispiel zum August 1975 80 v. H.) nach einem Zufallsprinzip(!) aus dem Fragenpool auswählt, nachdem es vorab die Zahl der Aufgaben pro Fachgebiet (!, die AOÄ kennt nur Stoffgebiete) festgelegt hat. Bei einer anschließenden Überprüfung werden Fragen ausgeschieden, die sich sachlich überschneiden.

Die restlichen Fragen werden gezielt auf der Basis der früher gewonnenen statistischen Kennwerte (zum Beispiel Schwierigkeit und Trennschärfe) unter Abrundung fachlicher(!) Gesichtspunkte übernommen. Dabei sind die Zahlen so verteilt, dass es ganze Fächer gibt, die selbst dann nicht kontentvalide geprüft werden können, wenn alle Fragen einem Lehrziel gewidmet wären (so zum Beispiel das sicher wichtige Gebiet der "Akut lebensbedrohlichen Zustände" mit 10 Fragen (August 1977 und März 1979: 15) [4-11]).

Um es nochmals anders zu formulieren: Jede Prüfung kann nur eine exemplarische Stichprobe aus dem geforderten gesamten Lehrziel abfragen. Diese Stichprobe muss aber so beschaffen sein, dass man nach der Prüfung mit der erforderlichen Sicherheit aussagen kann, dass der Kandidat die exemplarisch ausgewählten Lehrziele auch im geforderten Umfang erreicht hat. Dem IMPP ist eine solche Aussage bei der derzeitigen Art der Prüfungskonstruktion und -bewertung nicht möglich.

Schon von einer ins Einzelne gehenden inhaltlichen Überprüfung müssen wir also global feststellen, dass die Mangelhaftigkeit der Kriterienvalidität und das Fehlen der Konstruktvalidität durch ein Ungenügen der Kontentvalidität komplettiert werden.

Literatur

1. Fischer, G. H.: Einführung in die Theorie psychologischer Tests. Grundlagen und Anwendung, Verlag Hans Huber, Bern, Stuttgart, Wien 1974

2. Hofstätter, P. R. (Hrsg.): Das Fischer-Lexikon, Psychologie, Fischer-Verlag, Frankfurt 1957

3. Hubbard, J. P.: Erfolgsmessung der medizinischen Ausbildung. Verlag Hans Huber, Bern, Stuttgart, Wien 1974

4. IMPP: Ergebnisbericht über die schriftlichen Prüfungen nach der Approbationsordnung für Ärzte, August 1975

5. IMPP: Ergebnisbericht über die schriftlichen Prüfungen nach der Approbationsordnung für Ärzte, März 1976

6. IMPP: Ergebnisbericht über die schriftlichen Prüfungen nach der Approbationsordnung für Ärzte, August 1976

7. IMPP: Ergebnisbericht über die schriftlichen Prüfungen nach der Approbationsordnung für Ärzte, März 1977

8. IMPP: Ergebnisbericht über die schriftlichen Prüfungen nach der Approbationsordnung für Ärzte, August 1977

9. IMPP: Ergebnisbericht über die schriftlichen Prüfungen nach der Approbationsordnung für Ärzte, Vorbericht, Frühjahr 1978

10. IMPP: Ergebnisbericht über die schriftlichen Prüfungen nach der Approbationsordnung für Ärzte, Vorbericht, Herbst 1978

11. IMPP: Ergebnisbericht über die schriftlichen Prüfungen nach der Approbationsordnung für Ärzte, Vorbericht, Frühjahr 1979

12. Klauer, K. J.: Einführung in die Theorie lehrzielorientierter Tests. In: Klauer, K. J. et al. Lehrzielorientierte Tests, Schwann Verlag, Düsseldorf, 1975

13. Klauer, K. J.: Zur Theorie und Praxis des binomialen Modells lehrzielorientierter Tests. In: siehe [12]

14. Kuni. H., Becker, P.: Multiple Choice als Numerus clausus (2) Zu viele Medizinstudenten scheitern am Messfehler "der arzt im krankenhaus" (1980) 292-293

15. Lienert, G. A.: Testaufbau und Testanalyse. Verlag J. Beltz, Weinheim, 1961

16. Schumacher, Ch. F.: Auswertung und Analyse der Prüfung. In: Hubbard, J. P.: Erfolgsmessung der medizinischen Ausbildung, Verlag Hans Huber, Bern, Stuttgart, Wien 1974

Anschrift der Verfasser

Prof. Dr. Horst Kuni, Auf dem Wüsten 5, 35043 Marburg, horst@kuni.org

Rechtsanwalt Dr. Peter Becker, Gisonenweg 9, 35037 Marburg