

DER SCHWIERIGKEITSGRAD DER SCHRIFTLICHEN PRÜFUNGEN IN DER MEDIZIN (FRAGEN ZUM ZWEITEN ABSCHNITT DER ÄRZTLICHEN PRÜFUNGEN)

S. Drolshagen, Frankfurt a. M.

Im Dezember 1986 trat die Fünfte Verordnung zur Änderung der Approbationsordnung in Kraft. Mit der Einführung zusätzlicher mündlicher Prüfungen in der Ärztlichen Vorprüfung und im zweiten Abschnitt der Ärztlichen Prüfung wurde einem allgemeinen und seit Jahren geäußerten Unbehagen daran Rechtskraft verliehen, daß allein schriftliche Prüfungen nach dem Antwort-Auswahl-Verfahren über die Qualifikationen der Studenten zum weiteren Studium entschieden. Beseitigen die mündlichen Prüfungen aber auch die Probleme der schriftlichen?

Das mit der Durchführung der schriftlichen Prüfungen betraute IMPP hat sich seit seiner Einrichtung die unterschiedlichste Kritik am Verfahren und den Inhalten seiner Arbeit gefallen lassen müssen.

Bei der Durchführung wurden ihm mangelnde Sorgfalt (Lohölter 1985a, Sachverständigenkommission 1981), bei der testtheoretischen Absicherung des Verfahrens Verstöße gegen wesentliche Punkte dieser Testtheorie vorgeworfen (Hilke 1980, Kuni & Becker 1981, Harth 1984). Zentraler Kritikpunkt aber war und ist wohl der Inhalt der Fragen:

- 1.) Sie seien zu den jeweiligen Terminen nicht gleichwertig, also gleich schwierig (Sachverständigenkommission 1981, Fuchs 1984, Kern 1986),
- 2.) sie seien inhaltlich oft fehlerhaft (Buhl 1981, Rühle 1986),
- 3.) sie hätten mit der zukünftigen ärztlichen Praxis, aber auch mit dem an der Hochschule Vermittelten oft nur recht wenig zu tun (Buhl 1981, Barnikol 1982, Golenhofen 1985, Schenkel 1985),
- 4.) sie führten zu einer "Kanonisierung des ärztlichen Wissens" (Lötterle & Wuermeling 1983, Harth 1984) und
- 5.) sie erzeugten ein nur in Ansätzen bekanntes Lernverhalten, von dem man aber annimmt, daß es im Hinblick auf die

Qualifikationen eines Arztes unerwünscht erscheint (Kienle & Kreysch 1978, Wehner & Kruse 1979, Kuni 1985).

Im folgenden soll uns nur der Schwierigkeitsgrad der Examina beschäftigen.

1. "Schwere" und "leichte" Prüfungen:

Zur Beantwortung dieser Frage sollen die vorläufigen Ergebnisberichte sowie die Item-Analysen des IMPP herangezogen werden, in denen sich Kennwerte über den Schwierigkeitsgrad jeder einzelnen Frage ausmachen lassen.

1.1 Durchfallquote:

Wie in Tabelle 1 dargestellt, sind "Katastrophenmeldungen" wie bei der Ärztlichen Vorprüfung nicht festzustellen. Aber auch hier gibt es "Ausreißer".

Termin	Teilnehmer	Nichtbesteher	N.-b. in %	Mittelwert	Bestehensgrenze
M 77	2111	11	0,52	74,01	250 =50,0%
A 77	2897	39	1,35	69,59	250 =50,0%
M 78	3500	51	1,46	70,18	250 =50,0%
A 78	3431	88	2,56	69,67	250 =50,0%
M 79	4080	89	2,18	70,10	250 =50,0%
A 79	3061	584	19,08	67,59	348 =60,0%
M 80	4019	393	9,78	71,35	348 =60,0%
A 80	4151	560	13,49	69,77	348 =60,0%
M 81	4501	666	14,80	68,59	348 =60,0%
A 81	4469	325	7,27	66,72	318 =54,8%
M 82	5162	264	5,11	73,24	348 =60,0%
A 82	4310	287	6,66	73,27	348 =60,0%
M 83	5163	241	4,67	72,02	343 =59,1%
A 83	5179	301	5,81	71,10	339 =58,4%
M 84	5226	285	5,45	73,22	348 =60,0%
A 84	4866	282	5,80	72,72	346 =59,7%
M 85	5562	323	5,81	72,34	345 =59,5%
A 85	5619	255	4,54	73,82	348 =60,0%
M 86	6042	448	7,41	66,64	317 =54,7%
A 86	4503	300	6,67	70,77	337 =58,1%

Tab.1 Teilnehmer, Nichtbesteherquoten und Bestehensgrenzen im 2. Abschnitt der Ärztlichen Prüfung (Quelle:IMPP-Vorberichte)

Tatsächlich lag der Prozentsatz der Nichtbesteher der Prüfung vom Frühjahr 1986 (=M=März) beispielsweise mit 7,4% erkennbar höher, als bei den letzten 9 Prüfungsterminen. Die durchschnittliche Prüfungsleistung sank auf 386,5 (66,6%) richtig beantwortete Fragen, die Bestehensgrenze auf 317 Fragen (54,65%) und lag damit so niedrig wie noch nie, seit es die Gleitklausel gibt!

1.2 Die Prüflinge:

In den Vorberichten des IMPP werden die Prüflinge vornehmlich anhand von Merkmalen beschrieben, von denen angenommen wird, daß sie Einfluß auf die Prüfungsergebnisse haben (Tab.2): So schneiden Ausländer (Spalte 4) und Prüfungswiederholer (Spalten 2+3) aus verschiedenen Gründen (s.IMPP 1981, Schweitzer 1983, Lohölter 1985b) schlechter ab als Deutsche und Prüflinge, die erstmalig und zum frühestmöglichen Zeitpunkt (also mit 10 Fachsemestern; s. Spalte 5) zur Prüfung antreten. Tab. 2 weist für das Frühjahr 1986 deutlich mehr Prüfungsteilnehmer als zum Frühjahr zuvor aus, was auf die damals unmittelbar bevorstehende Verlängerung der Ausbildung um eineinhalb Jahre "Arzt im Praktikum" zurückführbar sein dürfte. Der Schluß liegt nahe, daß der "Ausreißer" auf eine überdurchschnittliche Zahl verhältnismäßig unvorbereiteter Prüflinge zurückzuführen ist, die diesen Examenstermin ohne den o.a. Außendruck nicht wahrgenommen hätte. Dem entspricht, daß der Anteil der Erstteilnehmer mit 10 Fachsemestern (Tab. 2, Spalte 5) ebenfalls deutlich anstieg.

Termin	Teiln.	1.Wh.	2.Wh.	Ausl.	11n.m.10FS
M 1977	2.111	0,3%	---	3,7%	20,8%
M 1978	3.500	1,0%	0,1%	5,2%	28,9%
M 1979	4.080	2,0%	0,2%	7,6%	35,5%
M 1980	4.019	11,8%	0,5%	7,4%	34,4%
M 1981	4.501	9,1%	1,7%	7,4%	35,5%
M 1982	5.162	5,2%	1,8%	6,4%	34,5%
M 1983	5.163	3,7%	1,2%	5,0%	32,6%
M 1984	5.226	4,8%	0,8%	5,6%	32,8%
M 1985	5.562	4,0%	0,8%	4,3%	32,9%
M 1986	6.042	3,5%	0,6%	3,9%	39,9%

Tab.2. Prüfungswiederholer, Prüflinge aus dem Ausland und Erstteilnehmer mit 10 Fachsemestern im 2. Abschnitt der Ärztlichen Prüfung.

Einer solchen Erklärung des relativ schlechten Frühjahrsergebnisses 1986 widerspricht aber, daß es 1. in der Vergangenheit mehrfach einen noch größeren Anstieg der Teilnehmerzahl gegeben hat, ohne, daß es zu einem schlechteren Prüfungsergebnis gekommen wäre und daß 2. die Verteilung der Antworten (Abb.1) lediglich eine Verschiebung des Kreuzungspunktes nach links bei sonst gleich verlaufender Kurve aufweist. Das bedeutet, daß das schlechte Ergebnis nicht durch eine Gruppe besonders schlecht vorbereiteter Studenten verursacht worden ist.

Immerhin handelt es sich bei den Prüflingen um mit dem Verfahren und seinen "Tücken" vertraute Studenten, die zudem nicht weit vor dem Abschluß ihrer Ausbildung stehen.

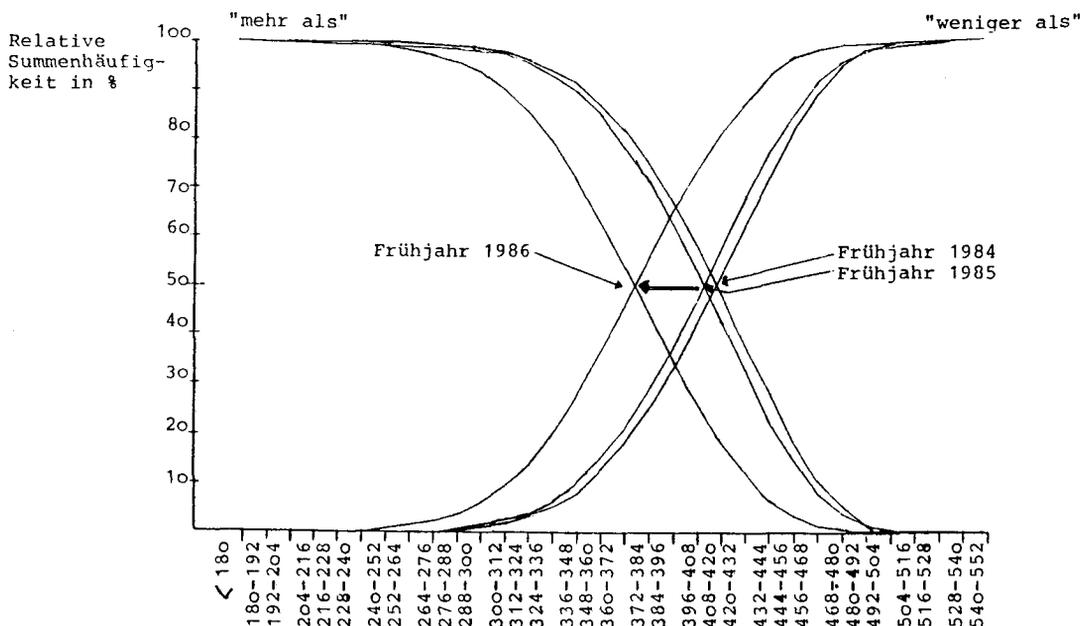
1.3 Item-Analysen:

Diese vom IMPP veröffentlichten Analysen der Prüfungsfragen lassen eine Beurteilung jeder Einzelfrage anhand von zwei Merkmalen zu, die, jedes für sich, lediglich hinweisende Funktion haben, beide zusammen aber Aufschlüsse über die Qualität einer Frage geben.

Der Schwierigkeitsindex (P) entspricht dem Prozentsatz der Prüflinge, die die betreffende Frage richtig beantworteten. Je höher dieser Prozentsatz, je größer also der Schwierigkeitsindex, desto leichter die Frage - ein etwas verwirrendes Wortspiel.

Sieht man sich die Item-Analysen der letzten sechs Prüfungstermine im Hin-

Abb. 1 Anzahl der richtig beantworteten Fragen im 2. Abschnitt der ärztlichen Prüfung im Frühjahr 1984, 1985 und 1986 - Summenpolygone der Rohwerte



blick auf den Schwierigkeitsindex an, so fällt auf, daß die Zahl der schweren Fragen zwar anstieg, selbst wenn man als schwere Fragen nur die ansieht, die von weniger als 20% der Prüflinge richtig beantwortet wurden. Tab. 3. zeigt die Verteilung der Fragen mit 2 Schwierigkeitsgraden: Eine leichte Verschiebung zugunsten der schweren Fragen ist festzustellen, aber insgesamt ist ihre Zahl vergleichsweise gering.

Tab. 3:
Diesen als schwer einzustufenden Fragen steht ein Mehrfaches an Fragen gegenüber, die von mehr als 80% der Prüflinge richtig beantwortet wurden (s.Tab.3), wobei im Frühjahr 1986 ein deutlicher Einbruch festzustellen ist. Zu diesem Termin lag die Zahl der von mehr als 80% der Prüflinge richtig beantworteten Fragen mit 251 etwa 27% unter dem Durch-

schnitt der übrigen Termine. Bei einer Reihe von Fragen vereinigte außerdem eine falsche Antwortalternative die große Mehrheit der Kandidaten (im Einzelfall 68%) auf sich. Außerdem lag das Gesamtergebnis der Prüflinge, die eine falsche Antwortalternative wählten, deutlich höher als das der Prüflinge, die richtig antworteten (Tab. 4,N1). Zieht man zusätzlich noch den Trennschärfeindex nur dieser Fragen heran, so bleibt immer noch eine respektable Anzahl von Fragen übrig (Tab.4,N2), auf die dieses Phänomen zutrifft.¹

Bei den in der zweiten Spalte der Tab. 4 aufgeführten Fragen lag der Trennschärfeindex der falschen Antwortalternative zusätzlich höher als der der richtigen Lösung.

	< 20%	> 80%
Frühjahr 83	10	261
Herbst 83	16	243
Frühjahr 84	9	260
Herbst 84	9	261
Frühjahr 85	10	258
Herbst 85	14	241
Frühjahr 86	19	183
Herbst 86	13	235

Tab. 3 Anzahl der Fragen, die von weniger als 20% bzw. mehr als 80% der Prüflinge richtig beantwortet wurden (Quelle: Item-Analysen).

	N 1	N 2
Frühjahr 83	35	20
Herbst 83	34	22
Frühjahr 84	36	24
Herbst 84	36	23
Frühjahr 85	36	21
Herbst 85	39	18
Frühjahr 86	49	30
Herbst 86	36	25

Tab. 4: Anzahl der Fragen, bei denen eine falsche Antwortalternative von Studenten mit besserem Gesamtergebnis gewählt wurde, als die richtige (N 1). N 2 bezeichnet die Zahl der Fragen mit zusätzlich höherem Trennschärfeindex der Falschantwort (Quelle: Item-Analysen).

Ein Faktor allein kann demnach nicht für den unterschiedlichen Schwierigkeitsgrad der Prüfungen verantwortlich gemacht werden - dafür sind die jeweils einem kritisierbaren Index zuzuordnenden Fragen bei einer Gesamtzahl von 580 kaum von ausschlaggebender Bedeutung.

Kommen jedoch mehrere Faktoren bei einer Prüfung zusammen, wie im Frühjahr 1986, kann von einer gleich schweren Prüfung kaum mehr die Rede sein. Die "kollektiven Fehlreaktionen" der Prüflinge bei einer offensichtlich schwierigeren Prüfung zeigen zweierlei:

1. Die Formulierung angemessener "Ablenker" von der richtigen Lösung ist ein immer noch unterschätztes (und offenbar ungelöstes) Problem. 2. Die Prüflinge, die die schweren Fragen richtig beantworten, gehören nicht unbedingt zur besonders leistungsstarken Spitzengruppe, die ja mit diesen Aufgaben von den übrigen Leistungsgruppen unterscheidbar werden sollte.

2. "Dechiffrierarbeit"

Wie kompliziert der Vorgang der Fragegenerierung wirklich ist, zeigt sich erst, wenn man sich vor Augen führt, welche Überlegungen einerseits zur Formulierung einer bestimmten Frage geführt haben und welche Überlegungen andererseits ein Prüfling angesichts derselben Frage anstellt, um zu seiner Lösung zu gelangen.

Man darf sich den Vorgang wie einen Verschlüsselungsprozeß vorstellen, bei dem der Prüfer seine Absicht (die Kenntnis

des Gegenstands X oder die "richtige" Ableitung von Y aus J, K + L z.B. zu prüfen) als MC-Frage verschlüsselt und der Prüfling diese Absicht zunächst wieder entschlüsseln muß, um zur richtigen Lösung des Problems zu gelangen.

Vom Prüfling entschlüsselt werden muß v.a. zunächst das Komplexitätsniveau der Frage. Die Anhebung des Komplexitätsniveaus macht eine Frage mit Sicherheit schwerer. Der Kandidat bekommt dafür aber ebenso nur einen einzigen Punkt, wie für die leichtere Formulierung. Zum anderen unterstellt der Autor einer solchen Frage, daß der Prüfling seinen Gedankengang nachvollzieht - bei entsprechendem Wissen einfach nachvollziehen muß. Der Kandidat geht aber den umgekehrten Weg, ohne allerdings das Ziel dieses Weges zu kennen. Die geistige Leistung des Prüflings, passives Wissen auf Kommando reproduzieren zu können, ist zu ergänzen um die Entschlüsselung der Absicht des Fragenautors. Für die Studenten stellt sich das Problem ganz konkret bei jeder Prüfung: Was will das IMPP eigentlich wissen - denken "die" nun direkt oder "wieder um drei Ecken"? Das entscheiden zu können, ist für die Prüflinge oft von ebenso großer Bedeutung wie das Wissen um die vom IMPP vertretene Lehrmeinung. Hierzu ein paar Beispiele:

2.1 Komplexitätsniveau:

Beispiel 1

Der gut vorbereitete Student hat gelernt, daß für einen solchen Rat allein das Stadium und der Allgemeinzu-

Beispiel 1.

2. Tag Frage 60 (Gynäkologie):

Einer Patientin mit einem Endometriumkarzinom muß man zur Operation raten,

weil

die Strahlentherapie eines Endometriumkarzinoms weit weniger effektiv und mit schlechteren Heilungsaussichten verbunden ist als die operative Therapie. (Richtige Lösung "A", d.h.: beide Aussagen und die Verknüpfung sind richtig)

Kennwerte:¹

	N	P(%)	RWM	R
A	2743	45	382	-0.10
B	66	1	371	-0.03
C	676	11	382	-0.03
D	414	7	389	0.02
E	2141	35	394	0.11

stand der Patientin maßgeblich sind (Stegner, 1981, S. 384 f.). Der in der Frage formulierte Imperativ wurde entsprechend von immerhin 35% der Prüflinge verneint und keine der angebotenen Alternativen als richtig bewertet. Diese Studenten haben die Reichweite der Frage falsch interpretiert und es waren leistungsmäßig nicht die schlechtesten (durchschnittlich 394 Fragen im Gesamtergebnis richtig).

Nicht zuletzt deshalb werden bei der

Prüfungsvorbereitung so extensiv Altfragen "gepaukt"! Der Kandidat muß sich auf das sich "in der Regel" darstellende Komplexitätsniveau einstellen, wie auf die vom IMPP zugrundegelegten Lehrbücher. Zu nahezu jeder MC-Frage muß sich also der Prüfling die Fragen "dahinter" stellen: Auf welchem Komplexitätsniveau liegt die richtige Antwort und was genau ist gefragt?

2.2 Mißverständnisse:

Beispiel 2:

Beispiel 2:

1. Tag Frage 34 (Pathologie):

Welche Aussage trifft nicht zu?

Mit einer Nebennierenrindenhyperplasie kann im kausalen Zusammenhang stehen ein(e)

- (A) Morbus Cushing
- (B) adrenogenitales Syndrom
- X (C) Hypernephrom
- (D) kleinzelliges Bronchialkarzinom
- (E) essentielle Hypertonie

Kennwerte:

	N	P(%)	RWM	R
A	20	0	351	-0.04
B	86	1	345	-0.11
C	1634	27	383	-0.06
D	197	3	330	-0.22
E	4103	68	392	0.16

Die richtige Lösung (C) ist zwar im einschlägigen Lehrbuch (Eder/Gedigk) - dort aber im wahrsten Sinne des Wortes im Kleingedruckten zu finden. Gehen die Pathologiefragen aus oder steht in dem genannten Lehrbuch zuwenig "Wissenswertes", um es abzufragen?

Die hohe Zahl derer, die sich für eine, wenn auch falsche, Antwort entschieden (s.Kennwerte) läßt vermuten, daß durchaus nicht geraten, sondern Wissen angeeignet wurde - aber das "falsche", wobei die unklare Ätiologie der essentiellen Hypertonie vielleicht erklärt, warum 68% diese Alternative wählten.

Beispiel 3:

1. Tag, Frage 145 (Innere Medizin):

Welche der folgenden Zuordnungen von Blutgerinnungsstörung und Laborbefunden treffen zu?

- (1) von Willebrand - Syndrom:
 - Blutungszeit: verlängert
 - Thromboplastinzeit (Quick): normal
 - part. Thromboplastinzeit (PTT): verlängert
 - Thrombozytenzahl: normal

- (2) isolierte Thrombozytenfunktionsstörung:
 - Blutungszeit: normal
 - Thromboplastinzeit (Quick): normal
 - part. Thromboplastinzeit (PTT): normal
 - Thrombozytenzahl: erniedrigt

- (3) Faktor XIII-Mangel:
 - Blutungszeit: normal
 - Thromboplastinzeit (Quick): normal
 - part. Thromboplastinzeit (PTT): normal
 - Thrombozytenzahl: normal

- (A) Keine der Zuordnungen (1) bis (3) trifft zu
- (B) nur 1 und 2 sind richtig
- X (C) nur 1 und 3 sind richtig
- (D) nur 2 und 3 sind richtig
- (E) 1-3= alle sind richtig

Kennwerte:

	N	P(%)	RWM	R
A	3481	58	382	-0.10
B	761	13	381	-0.05
C	1500	25	403	0.19
D	124	2	371	-0.05
E	168	3	370	-0.06

2.3 Lehrbuch:

Beispiel 3:

Die große Mehrheit der Prüflinge zog sich mit der Alternative "A" aus dem Dilemma, aus den Vorgaben eine herauszusuchen, obwohl eine korrekte Antwort anhand der angebotenen Alternativen nicht möglich ist. Die Zuordnungen unter (1) sind zutreffend, alle anderen jedoch nicht. Anders als bei der IMPP-Lösung, nach der auch die Zuordnungen unter (3) richtig sein sollen, weiß der informierte Prüfling nach einschlägiger Lektüre,

daß der PTT bei Faktor XIII-Mangel nicht normal ist (Schettler 1984, Bd.2, S. 186; Vosschulte 1979, S. 228). Er muß daher schließen, daß in Ermangelung anderer Alternativen eben keine der Zuordnungen stimmt. Der Antwort im Lösungsschlüssel des IMPP unterliegt sicher eine Aussage in irgendeinem Fachbuch, in dem eine andere Auffassung vertreten wird; eine solche Frage über einen offenbar unentschiedenen Tatbestand zum

Beispiel 4:

2. Tag Frage 46 (Gynäkologie):

Welche Aussage trifft nicht zu?

Bei starkem (evtl. auch nur zeitweilig auftretendem) Fluor genitalis muß man u.a. an folgende Ursachen denken:

- (A) gonorrhöische Infektion
- X (B) Salpingitis isthmica nodosa
- (C) Trichomonadeninfektion
- (D) psychische Genese
- (E) Hydrops tubae profluens

Kennwerte:

	N	P(%)	RWM	R
A	131	2	379	-0.02
B	1956	32	395	0.11
C	24	0	339	-0.06
D	349	6	360	-0.14
E	3577	59	385	-0.04

Gegenstand einer Prüfung zu machen, die keinerlei Rückfragen oder Korrekturen erlaubt, ist schlicht unfair dem Prüfling gegenüber. Dazu gleich ein weiteres Beispiel:

2.4 Verbindlichkeit:²

Beispiel 4:

Die richtige Lösung wird mit "B" angegeben, in der Logik der Frage soll Salpingitis isthmica nodosa also falsch sein. Das klassische Wörterbuch, der "Pschyrembel" (1986) ist da anderer Ansicht (S. 1484):... Ausfluß (Fluor) ist ein Hauptsymptom dieser Krankheit. Die Lösung "E" stellt zwar eine recht unwahrscheinliche, aber keine auszuschließende Ursache dar und wurde (vielleicht deshalb?) von weit über der Hälfte der Prüflinge gewählt.

Die im Gespräch mit Studenten darüber hinaus kritisierten Fragen wurden z.T. mehrheitlich richtig beantwortet. Auch der Schwierigkeitsindex gibt anscheinend nur bedingt Auskunft über schwere und leichte Fragen bzw. Prüfungen.

3. Schluß:

Die Beantwortung unserer Frage, welche Prüfungen schwer und welche leicht sind, hängt also vom jeweiligen Standort ab - Prüflinge sehen offensichtlich Schwierigkeiten an Punkten, die die Fragenkonstrukteure nicht sehen, vielleicht nicht sehen konnten. Zwischen beiden gibt es keine Verbindung und keine Verbindlichkeit.

Daß keine Vorbereitungsstrategie wirklich sicherstellen kann, daß man diese Prüfung besteht, demonstrierten schon die wenigen Beispiele, die weiter oben

dargestellt wurden. Es zeigt sich darüber hinaus die Hilflosigkeit von Prüfern und Geprüften, die jeweils andere Seite auch nur annähernd einzuschätzen. So bleibt nur zu fordern, daß MC-Fragen sich auf das Abfragen des Faktenwissens beschränken, das in allen wesentlichen Fachbüchern gleichermaßen und als verbindlicher Lehrinhalt kenntlich und zugänglich ist - und dies auf einem für die Prüflinge berechenbaren Komplexitätsniveau. Wie umfangreich auch immer dieses Wissen definiert ist - dem Prüfling kommt es darauf an, zu wissen, was er tun muß, um den Anforderungen der Prüfung gerecht zu werden. Dies scheint mir z.Zt.nicht gegeben.

Wenn schon die schriftlichen Prüfungen nicht jedem Examensjahrgang Gleichbehandlung durch gleich schwere Prüfungen garantieren können und Dechiffrierprobleme wie oben aufwerfen, so trifft dies in sehr viel größerem Maß auf mündliche Prüfungen zu. Mündliche Examen in der Medizin sind nicht neu. Die Fakultäten wären gut beraten, wenn sie die mit dieser Prüfungsform begangenen Fehler der 50er und 60er Jahre vermieden.

Literatur:

Barnikol, W., Ist das Multiple-Choice-Prüfungssystem für die medizinische Ausbildung geeignet? Teile 1+2, in: der arzt im Krankenhaus 2/1982, S. 110-115 + 3/1982, S. 159-166.
Buhl, R., Fehlerhafte Multiple-choice-Fragen im Kreuzfeuer der Kritik, Teile I. bis III. in: DÄBl. Heft 6, S.239-243; H.7, S.302-307 H.8, S.356-363, 1981
Fuchs, R.K., Die Absurdität der Bestehensregel für die Prüfungen in der ärztlichen Ausbildung, in: Medizinische Ausbildung Nr.2, 1985, S. 19-26.

Golenhofen, K., Physikum 1985: Bocksprünge auf Mainzer Art. Das IMPP hat aus den Fehlern der Vergangenheit nichts gelernt.

In: Deutsche Universitätszeitung 21, 1985, Heft 14, S. 15-18

Guilbert, J.-J., Ausbildung in den Gesundheitsberufen - Pädagogischer Leitfaden, Bern u.a. 1979

Harth, O., Triumph des Details, in: Deutsche Universitätszeitung 20, 1984, S. 15-17.

Hilke, R., Grundlagen normorientierter und kriteriumorientierter Tests - eine kritische Auseinandersetzung mit der klassischen Testtheorie und den logistischen Testmodellen, Bern u.a. 1980

Hinrichsen, K., Die wirklich ungeeigneten Fragen wurden gar nicht eliminiert. Zur Nachbesserung der ärztlichen Vorprüfung 1985,

in: DÄBl. 1985, 82, S.2534-2537

Hubbard, J.P., Erfolgsmessung der medizinischen Ausbildung, Bern u.a.1974

Hubbard, J.P., Measuring Medical Education. The Tests and the Experience of the National Board of Medical Examiners, Philadelphia 1978²

IMPP - Kraemer, H.-J., Möglichkeiten und Grenzen schriftlicher Prüfungen in der Medizin,

in: Hess. Ärztebl. 4/1981, S. 303-306

Kienle, G. & W. Kreysch, Multiple-Choice-Prüfung und ärztliche Berufsfähigkeit, in: DÄBl. Heft 14, 1978, S.838-841.

Kuni, H. & P. Becker, Multiple-Choice als Numerus Clausus, Teile 1-9+Nachtrag, in: der arzt im krankenhaus, Nr.4/1980 bis 2/1981.

Kuni, H., Die fehlende Validität der Ärztlichen Vorprüfung. In: der arzt im krankenhaus 14, 1985, S.1013-1015

Lötterle, J. & H.-B. Wuermeling, Medizinische Multiple-Choice-Fragen - Forensische Bedeutung, in:Münch.med.Wschr. 125, Nr.13,1983, S. 259.

Lohölter, R., Das Prüfungssystem hat erneut versagt - Anmerkungen zur ärztlichen Vorprüfung vom März 1985, in: Medizinische Ausbildung 2,Heft 1, 1985a, S. 27-31

Lohölter, R., Die Hürden werden immer höher. Die Schwierigkeiten, die ausländischen Studenten drohen, in: Deutsche Universitätszeitung 11/1985b, S. 22-23

Rühle, W., Opfer der Testtheoretiker, in: DÄBl. 83, Heft 45, 1986,S.3090-3091

Sachverständigen-Kommission zur Prüfung der Gründe für das Ergebnis der Ärztlichen Vorprüfung im März 1981: Arbeitsbericht.

In: Deutsche Universitätszeitung 1982 a, 30, Heft 3, S. 25-29

Sachverständigen-Kommission zur Prüfung der Gründe für das Ergebnis der Ärztlichen Vorprüfung im März 1981: Wachsende Fernsteuerung in der Medizin. Die Sachverständigenkommission bezieht erneut Stellung zur Prüfungsgestaltung am IMPP. In: Deutsche Universitätszeitung 1982 b, 30, H.9, S. 14-16

Scheer, J. & H. Zenz, Studenten in der Prüfung, Frankfurt/M. 1973

Schenkel, H.,Werdegang einer Multiple-choice-Prüfung.

In: Schweizerische Ärztezeitung Nr. 26, 1985, S. 1194-1202

Schenkel, H., Die Gültigkeit der Multiple-Choice-Prüfungen in Innerer Medizin gemessen am Außenkriterium der Relevanz für die Allgemeinpraxis, Diss. Bern, Juli 1985

Schweizer, D., Die staatlichen Prüfungen im Medizinstudium am Beispiel des zweiten Abschnittes der Ärztlichen Prüfung. Möglichkeiten und Grenzen. Diss. Marburg 1983

Verwaltungsrat des Instituts für medizinische und pharmazeutische Prüfungsfragen (IMPP): Stellungnahme zum Schlußbericht der unabhängigen Sachverständigen-Arbeitsgruppe zur Prüfung der Gründe für das Ergebnis der Ärztlichen Vorprüfung im März 1981.

In: Deutsche Universitätszeitung 1982, 30, H. 4, S. 22-23

Wehner, Th. & P. Kruse, Analyse des Multiple-choice-Verfahrens - Prüfung als Denkresultatkontrolle oder als denkverlaufspezifische Rückmeldung.

In: DÄBl. 1979, Heft 43, S. 2831 - 2837

Wehner, Th. & P. Kruse, Multiple-choice-Prüfung des Medizinstudenten: Ritual mit Konsequenzen.

In: Klinikarzt mit Medizinstudent 9, 1980, Nr. 3, S. 226-237

Lehrbücher:

Eder, M. & P. Gedigk:Lehrbuch der Allgem. Pathologie u.d.Pathologischen Anatomie, 31.Aufl.1984

Schettler, F.G., (Hrsg.): Innere Medizin, 6.neubearb.Aufl.1984, Stuttgart

Stegner, H.E., Gynäkologie und Geburtshilfe, Stuttgart 1981⁶

Vosschulte, K., Innere Medizin und Chirurgie, F. Stuttgart u.a.2.durchges. Aufl.1981

Stefan Drolshagen, Didaktik der Medizin am Fachbereich Humanmedizin, Theodor-Stern-Kai 7, D-6000 Frankfurt/Main

Anmerkungen:

1. Der Trennschärfeindex belegt jede der 5 Antwortalternativen mit einem Zahlenwert, der Aussagen darüber erlaubt, inwieweit die betreffende (also auch die falsche) Antwortalternative mehrheitlich von "leistungsstarken" oder "leistungsschwachen" Studenten, gemessen am Gesamtergebnis, gewählt wurde, also die Eigenschaft einer Frage, gute von schlechten Kandidaten zu trennen. Der Index kann Werte zwischen -1 und +1 annehmen, wobei der beschriebene Zusammenhang beim Wert +1 vollständig, beim Wert 0 nicht gegeben ist. In der Literatur (Hubbard 1974, S.66-67; Guilbert 1979, S.234-240) werden Indices über .25 als ein Merkmal für eine gute Frage, Werte über .50 bei MC-Fragen medizinischer Prüfungen für fast nicht erreichbar (Hubbard) gehalten. Werte zwischen .15 und .24 weisen auf Mängel bei der Formulierung der Frage hin, Werte unter .15 verweisen auf erhebliche Mängel der Frage hin, die wahrscheinlich zu ihrer Eliminierung aus dem Fragenpool führen.
2. Daß zwei von vier Beispielen der Gynäkologie entstammen, ist keine Absicht - die Beispiele wurden nach Plausibilitätsgesichtspunkten ausgewählt