DIE SCHRIFTLICHE PRÜFUNG NACH DER ÄAPPO

PROBLEME DER LEISTUNGSMESSUNG UND LEISTUNGSBEWERTUNG

Gerfried Gebert, Mainz

Das Bestehen der Ärztlichen Prüfung gilt nach der Bundesärzteordung als Nachweis dafür, daß der Kandidat die fachlichen Voraussetzungen für die Erteilung der Approbation als Arzt erfüllt. Die Ärztliche Vorprüfung dient analog der Beurteilung der Frage, ob der Kandidat über die für die Zulassung zum klinischen Ausbildungsabschnitt für erforderlich gehaltenen Kenntnisse verfügt.
Die Prüfungen sind Leistungsprüfungen, die verhindern sollen, daß unzureichend fähige Kandidaten den Beruf des Arztes ausüben. Ihr Ziel ist nicht, einen bestimmten Anteil der Kandidaten bestehen oder nicht bestehen zu lassen.

Prüfungsstoff der schriftlichen Examina

Der globale Rahmen für den Inhalt des in den einzelnen Stoffgebieten zu prüfenden Wissens ist die ÄAppO mit ihren Prüfungsstoffkatalogen. Inhaltliche Leitlinie für deren Umsetzung in zu messende Kenntnisse ist die Vorschrift des § 14(2) ÄAppO: Die Prüfungsfragen sind auf die für den Arzt allgemein erforderlichen Kenntnisse abzustellen.
Es wäre nicht sachgerecht, diese Vorgabe im Sinne von "für den Allgemeinarzt erforderlichen " Kenntnissen auszulegen. Dies ergibt sich bereits daraus, daß in der Bundesrepublik Deutschland nur etwa ein Viertel der berufstätigen Ärzte als niedergelassene Allgemeinärzte bzw. Praktische Ärzte arbeiten. Es geht vielmehr um die Kenntnisse, die die allgemein als Grundlage der ärztlichen Tätigkeit angesehen werden und die die Basis für die durch Weiterbildung erfolgende Qualifikation zur selbständigen Tätigkeit in einem der zahlreichen Sektoren des Berufsfeldes "Arzt" liefern.

Dieser Ansatz entspricht auch der Berufswirklichkeit, denn selbst die Tätigkeit als Praktischer Arzt, bei der die selbständige Berufsausübung nicht an eine vorhergehende Weiterbildung gebunden ist, wird nach der Statistik der Kassenärztlichen Bundesvereinigung überwiegend erst nach vierjähriger und längerer Assistenzarztzeit aufgenommen, d.h. nach nur formell nicht dokumentierter Weiterbildung.

Problematischer ist die Anwendung des in § 14(2) formulierten Prinzips auf die Ärztliche Vorprüfung. Dies gilt insbesondere für die Fächer Physik, Chemie und Biologie, in denen weniger "für den Arzt allgemein erforderliche Kenntnisse" als naturwissenschaftliche Grundlagen für den Erwerb dieser Kenntnisse zu prüfen sind.

Messung von Prüfungsstoffkenntnis mit m.c.-Fragen

Bei Antwortauswahlaufgaben hat der Kandidat keine Möglichkeit, durch Rückfragen Interpretationshilfen in bezug auf die Fragestellung zu erlangen oder problematische Sachverhalte zu diskutieren. Als Prüfungsstoff kommen daher nur eindeutig formulierbare und ebenso eindeutig als richtig oder

falsch einstufbare Fakten oder Zusammenhänge infrage. Hypothesen sollten nur geprüft werden, wenn sie allgemein anerkannt sind, und auch dann nur unter Kennzeichnung ihres Charakters (z.B. Symptome ersten Ranges der Schizophrenie nach K. SCHNEIDER).

Durch Antwortauswahlfragen wird der Prüfungsstoff in Form von Prüfungsgegenständen parzelliert. Bei Fragen, in denen eine richtige Aussage (Lösung) zusammen mit mehreren falschen (Distraktoren) angeboten wird, ist als Prüfungsgegenstand der Prüfungsstoff anzusehen, dessen Kenntnis zur Bewertung der Lösung erforderlich ist.

Die Anforderung durch eine Frage kann vom einfachen Wiedererkennen eines ausformulierten Sachverhalts bis zur Verwertung des Wissens in Form von Analyse oder Synthese reichen (taxonometrische Fragenschwierigkeit).

Die Zahl der Prüfungsgegenstände, die den Prüfungsstoff eines Faches bilden, läßt sich grob abschätzen, wenn man die gegenüber den Prüfungsstoffkatalogen der ÄAppO differenziertere Auflistung der Prüfungsthemen in den Gegenstandskatalogen des IMPP zuhilfe nimmt. Im Fach Anatomie beispielsweise enthält der GK 1 in 21 Kapiteln insgesamt 362 Themen vom Abstraktionsniveau "Rektum", "Schädelbasis" usw. Zu jedem dieser Themen lassen sich im Durchschnitt mindestens 15 Prüfungsgegenstände formulieren, wenn man die Spannweite der in der Ausbildung vermittelten Kenntnisse berücksichtigt (Form, Struktur, Funktion, Nachbarschaftsbeziehungen, Gefäße, Nerven usw.). Danach umfaßt der Prüfungsstoff im Fach Anatomie mindestens 5000 Prüfungsgegenstände. Bei einer anderen Schätzmethode kann man von den in den bisherigen Prüfungsterminen eingesetzten Fragen ausgehen. Im Fach Physiologische Chemie ergab eine Analyse der Prüfungen bis einschließlich Herbst 1984 (eigene, noch unveröffentlichte Ergebnisse) über 1500 lösungsrelevante Prüfungsgegenstände. Die Zahl der pro Termin neu geprüften Sachverhalte zeigte noch keine abnehmende Tendenz, so daß der Umfang des gesamten Prüfungsstoffs des Faches erheblich größer sein muß als der des bisher geprüften Stoffes.

Prüfungsleistung in der Antwortauswahlprüfung in Relation zur Prüfungsstoffkenntnis des Kandidaten

Als Maß der schriftlichen Prüfungsleistung dient die Zahl der richtig beantworteten Fragen im Verhältnis zur Zahl der gestellten Fragen. Die von der Beantwortung der einzelnen Frage gegebene dichotome (ja/nein) Information über die Kenntnis des geprüften Gegenstands wird durch Mittelung über die Fragen eines Termins zu einer quantitativen Aussage umgeformt. Ein solches Verfahren erscheint nur gerechtfertigt, wenn die Beantwortung der einzelnen Fragen eine zumindest formal vergleichbare Information über die Prüfungsstoffkenntnis liefert.

Bei Fragen vom Typ positive Einfachauswahl (eine richtige von fünf angebotenen Alternativen) ist zur korrekten Beantwortung in der Regel die Kenntnis eines Prüfungsgegenstands notwendig. Fragen vom sogenannten Verknüpfungstyp, bei denen die Richtigkeit von zwei Aussagen und (im Fall, daß beide zutreffend sind) das Vorhandensein einer kausalen Beziehung beurteilt werden muß, verlangen die Kenntnis von mindestens zwei Prüfungsgegenständen. Bei Fragen vom Kombinationstyp (nur Aussagen 1,2 und 5 sind richtig) können bis zu vier Prüfungsgegenstände lösungsrelevant geprüft werden.
Der gegenüber Einfachauswahlaufgaben erhöhte Anspruch von Verknüpfungs- und Kombinationsfragen an die Prüfungsstoffkenntnis kommt in der geringeren Häufigkeit richtiger Antworten bei diesen Fragentypen zum Ausdruck. Von den den Studenten noch nicht bekannten Fragen der Vorprüfungen von Herbst 1979 bis Frühjahr 1984 wurden im Fach Anatomie 53,4 % der Fragen vom Typ positive Einfachauswahl, aber nur 44,6 % der Kombinationsfragen richtig beantwortet. Im Fach Physiologische Chemie betrug die entsprechende Zahl 58,6 % bzw. 46,8 % (eigene, noch unveröffentlichte Untersuchungen).

Bei Verwendung von mehreren Fragentypen mit unterschiedlichem Anspruch an die Kenntnis des Prüfungsstoffs hat die in einem Termin erbrachte Prüfungsleistung eines Kandidaten den Charakter einer höchstens im Rahmen einer Rangordnung in-

terpretierbaren Zahl ohne quantifizierbaren Aussagewert für die Prüfungsstoffkenntnis.
Auch wenn nur Fragen mit vergleichbarem Anspruch an die Prüfungsstoffkenntnis eingesetzt werden, muß bedacht werden, daß mit den Fragen eines Termins nur ein kleiner Bruchteil des Prüfungsstoffs erfaßt werden kann. Im Fach Anatomie z.B. kann an einem Termin (ca. 80 Fragen) nur die Kenntnis von 1-2 % der Prüfungsgegenstände ermittelt werden, wenn Einfachauswahlaufgaben gestellt werden. Die Prüfungen haben somit den Charakter einer Stichprobe, und die Genauigkeit von Hochrechnungen aus dem Prüfungsergebnis auf die Prüfungsstoffkenntnis ist an die Repräsentativität und die Validität der Stichprobenzusammenstellung gebunden.

Festlegung der zu fordernden Prüfungsleistung

Es ist nicht möglich, die Qualität ärztlicher Leistungen allgemeinverbindlich zu quantifizieren und aus der späteren Berufstätigkeit von erfolgreichen Examenskandidaten retrospektiv einen Maßstab für die Bewertung der Examensleistung abzuleiten. Die Entscheidung über die Bestehensgrenze ist daher eine Ermessensentscheidung, die sich am Prinzip der Verantwortung gegenüber dem potentiellen Patienten und der Fairneß gegenüber dem Kandidaten zu orientieren hat.
Dabei ist möglich :

1. eine apodiktische Festlegung eines geforderten Mindestanteils richtiger Antworten unter Außerachtlassung des Aussagewerts der Aufgaben über die Prüfungsstoffkenntnis des Kandidaten
2. die Fixierung einer unteren Grenze für die Prüfungsstoffkenntnis, wobei der Mindestanteil richtiger Antworten als Folgegröße unter Berücksichtigung der Aussagefähigkeit und -genauigkeit des Prüfungsverfahrens bestimmt wird.

Die normorientierte Bestehensregel (Gleitklausel), bei der die Leistung eines Kandidaten an der aller Kandidaten eines Prüfungstermins gemessen wird, ist sachlich und möglicherweise auch rechtlich unbefriedigend. Mit einer solchen Bewertung wird von vornherein festgelegt, daß eine (kleine) Zahl von Kandidaten ungeachtet ihrer wahren Prüfungsleistung nicht bestehen darf und daß der (weit überwiegende) Rest der Kandidaten bestehen muß. Derzeit "beweist" nach der Gleitklauselregelung ein Kandidat seine Qualifikation für den Übergang ins Praktische Jahr, wenn er mit seiner Prüfungsleistung hinter nicht mehr als 94 % seiner Kommilitonen zurückbleibt. Für die Bewältigung dieser Aufgabe hat er zudem drei Anläufe zur Verfügung.

Wenn man den oben unter 2. beschriebenen Ansatz einer Festlegung der geforderten Mindestkenntnis des Prüfungsstoffs wählt, könnte die Bewertung "bestanden" z.B. an den Nachweis gebunden werden, daß der Kandidat mindestens die Hälfte des Stoffes beherrscht. Eine derartige Anforderung darf nicht als niedrig eingeschätzt oder gar mit dem Begriff "Halbwissen" abqualifiziert werden. Bei dem enormen Umfang des in der Ausbildung zum Arzt zu erwerbenden Wissens (allein in der Ärztlichen Vorprüfung umfaßt der Prüfungsstoff mindestens 20 000 Gegenstände) stellt ihre Erfüllung eine respektable Leistung dar.

Die in der Prüfung mit Antwortauswahlfragen (im Gegensatz zur mündlichen Prüfung) wegfallende Gewichtung der Prüfungsgegenstände nach ihrer Bedeutung sollte nicht zum Anlaß genommen werden, den Prüfungsstoff in sogenanntes Kern- oder Basiswissen und in sonstiges Wissen einzuteilen und der Beantwortung von Fragen zum "Basiswissen" größere Bedeutung zuzumessen. Abgesehen davon, daß es kaum möglich ist, einen Kon-

sens über die Zuordnung der Prüfungsgegenstände zum Basiswissen zu erreichen, wird der unterschiedlichen Bedeutung der einzelnen Prüfungsthemen bereits dadurch Rechnung getragen, daß zu wichtigeren Bereichen mehr Prüfungsgegenstände formuliert und entsprechend auch mehr Fragen gestellt werden.

Fehlerquellen der Kenntnisschätzung über die Prüfungsleistung

Systematische Fehler bei dem Rückschluß von dem Anteil richtiger Antworten bei den Fragen eines Termins auf die Prüfungsstoffkenntnis eines Kandidaten sind zu erwarten, wenn

1. Fragen zum Einsatz kommen, deren Beantwortung auch ohne das in der Ausbildung zum Arzt vermittelte Wissen möglich ist
2. verschiedene Fragentypen mit konstruktionsbedingt unterschiedlichen Anforderungen an die Prüfungsstoffkenntnis verwendet werden
3. die geprüften Gegenstände nicht ausreichend repräsentativ für den gesamten Prüfungsstoff sind
4. den Kandidaten bereits bekannte Fragestellungen wiederholt eingesetzt werden.

Bei Häufung wiederholter, aus Fragensammlungen bekannter Fragen bezieht sich die Prüfung statt auf den gesamten Prüfungsstoff hauptsächlich auf den bereits geprüften Stoff, der bisher nur einen geringen Teil des gesamten ausmacht. Studenten, die sich schwerpunktmäßig nach Fragensammlungen vorbereiten, erzielen bei hohem Anteil wiederholter Fragen ein besseres Ergebnis als es ihrer eigentlichen Fachkenntnis entspricht.

Einfluß der Fragenschwierigkeit auf die Prüfungsleistung

Als eigentliche Fragenschwierigkeit ist die Anforderung zu definieren, die die Fragestellung an die Umsetzung der Kenntnis des Prüfungsgegenstands in die richtige Beantwortung stellt (taxonometrisches Niveau). Für die Lösung entscheidend wird ein höherer Fragenschwierigkeitsgrad überhaupt erst, wenn der Kandidat über eine gewisse Kenntnis des geprüften Sachverhalts verfügt.

In den Antwortauswahlprüfungen in der Ausbildung zum Arzt werden (nicht nur in der Bundesrepublik Deutschland) überwiegend Fragen auf dem unteren Niveau des Wiedererkennens lehrbuchgemäß formulierter Sachverhalte gestellt. Der Anteil richtiger Antworten, den ein Kandidat erzielt, hängt deshalb hauptsächlich von seiner bloßen Kenntnis der geprüften Gegenstände ab. Gleiches gilt für den Anteil der Kandidaten, der eine Frage richtig beantwortet hat. Ungeachtet dessen wird der Prozentsatz der Kandidaten mit richtiger Lösung einer Frage als Fragenschwierigkeitsindex (FSI) bezeichnet, obwohl es sich eher um einen Index der Kenntnis vom geprüften Gegenstand handelt.

Der FSI stellt eine von der Leistungsfähigkeit der geprüften Population abhängige Größe dar. Die Häufigkeitsverteilung des FSI aller denkbaren Fragen kann somit nur hypothetisch auf der Basis einer Bezugspopulation definiert werden. Im Idealfall wäre von einer Normalverteilung auszugehen.

Für die Fragen der einzelnen Prüfungstermine ist auch wegen der zu verlangenden Chancengleichheit für die Kandidaten eine vergleichbare (an dem FSI einer Bezugspopulation orientierte) FSI-Verteilung zu fordern. Dies kann als erreicht angesehen werden, wenn (nach Aus-

schluß von Populationsunterschieden) kein statistisch signifikanter Unterschied zwischen den FSI-Verteilungen des einzelnen Termins und der Gesamtzahl der denkbaren Fragen (Fragengrundgesamtheit) besteht.

Die FSI-Häufigkeitsverteilung in der Fragengrundgesamtheit ist einer direkten Bestimmung nicht zugänglich, weil die für die Zusammenstellung der Prüfungen nach einem bestimmten Verfahren erarbeiteten Fragen nur einen kleinen Teil der zum Prüfungsstoff erarbeitbaren Fragen darstellen. Außerdem steht keine konstante Bezugspopulation, sondern nur die von Prüfungstermin zu Prüfungstermin verschiedene jeweilige Kandidatenpopulation zur Ermittlung der FSI-Verteilung zur Verfügung. Wenn die Schwankung der Populationsleistung jedoch gering gegenüber der Streubreite des FSI ist, kann durch die Zusammenfassung der Ergebnisse einer Reihe von Prüfungsterminen geeignete Information über die wahrscheinliche FSI-Verteilung in der Fragengrundgesamtheit erhalten werden.

Die Sicherstellung einer von Termin zu Termin vergleichbaren FSI-Verteilung ist bei den Prüfungen nach der ÄAppO nicht über eine gezielte Zusammenstellung der Prüfungsfragenstichproben realisierbar. Die Verwendung von durch Einsatz in früheren Terminen vorgetesteten Fragen verbietet sich, weil dadurch das Prüfungsergebnis systematisch verfälscht wird (s.o.). Eine prospektive Schätzung des FSI durch Sachverständige ist weder zuverlässig noch genau genug möglich. Der einzig sinnvolle Weg für die Zusammenstellung von nach der FSI-Häufigkeitsverteilung vergleichbaren Stichproben geht daher über eine Zufallsauswahl.

<u>Sind die bei den bisherigen Prüfungen beobachteten Unterschiede in der FSI-Verteilung der nicht systematisch verschiedenen Prüfungsfragen mit der Annahme einer Zufallsschwankung vereinbar?</u>

Die Untersuchung dieser Frage ist auch von prospektivem Interesse, denn sie ermöglicht Aussagen über die künftig zu erwartende Schwankungsbreite der Prüfungsschwierigkeit auf der Basis der bisherigen Erfahrungen.

Als Material dienten die Fragen der Fächer Anatomie, Physiologie und Physiologische Chemie der Ärztlichen Vorprüfungen vom Herbst 1979 bis zum Frühjahr 1984 (10 Termine). Zur Gewährleistung der Vergleichbarkeit wurden nur Fragen des Typs positive Einfachauswahl herangezogen und von diesen auch nur solche, deren lösungsentscheidender Inhalt den Kandidaten noch nicht aus früheren Fragen bekannt sein konnte.

Als Beispiel wird die FSI-Häufigkeitsverteilung der Fragen des Faches Anatomie dargestellt:

FSI (%)	n	Fragenzahl graphisch
6-15	3	ooo
16-25	2	oo
26-35	18	oooooooooooooooooo
36-45	13	ooooooooooooo
46-55	20	oooooooooooooooooooo
56-65	23	ooooooooooooooooooooooo
66-75	19	ooooooooooooooooooo
76-85	10	oooooooooo
86-95	2	oo

Die Verteilungsparameter (Mittelwert und Standardabweichung) für die zusammengefaßten Fragen aller Termine waren für das Fach

Anatomie	53,4 $\pm$ 18,0 %
Physiologie	54,6 $\pm$ 20,3 %
Physiol. Chemie	58,6 $\pm$ 18,5 % .

In allen drei Fächern fielen Mittelwert und Median der FSI-Verteilung praktisch zusammen, d.h. die Verteilungen waren weitgehend symmetrisch. Zur Prüfung der Zufallsannahme wurde untersucht, ob die FSI-Mittelwerte der einzelnen Prüfungen innerhalb des 95 % -Vertrauensbereichs blieben, der sich aus der FSI-Häufigkeitsverteilung der einzelnen Fragen des Faches und dem Umfang der Stichprobe des Termins errechnen läßt, wenn man von dem Prinzip normalverteilter FSI-Werte ausgeht.

Die von Herbst 1979 bis Frühjahr 1984 aufgetretenen Schwankungen im FSI-Mittel der Termine waren eher geringer als es bei einer Zufallsauswahl aus einer Fragengrundgesamtheit unter Annahme einer konstanten Populationsleistung zu erwarten war. Der 95 %- Vertrauensbereich für den FSI - Mittelwert wurde in keinem Fall überschritten. Innerhalb des 68 % -Vertrauensbereichs, in dem nach Zufallsgesichtspunkten nur 6 - 7 der 10 FSI-Mittelwerte pro Fach zu erwarten waren, blieben im Fach Anatomie 9, im Fach Physiologie 8 und im Fach Physiologische Chemie 6 der 10 Terminmittelwerte.

In der Diskussion der Zufallsbedingtheit der Schwankungen läßt sich das obige Ergebnis allerdings nur im Sinne des Fehlens eines Gegenbeweises interpretieren.

Ein positiver Hinweis für den Zufallscharakter der Unterschiede in den FSI-Mittelwerten (bezogen auf eine konstante Populationsleistung) läßt sich auf einem anderen Weg gewinnen. Die Leistungsfähigkeit der Kandidaten ist in den Herbstterminen im Mittel besser als in den Frühjahrsterminen, denn im Herbst ist der Anteil von Erstteilnehmern nach Mindeststudiendauer höher, und diese Kandidaten sind nach der IMPP-Statistik leistungsstärker als die übrige Population. Wenn sich die FSI-Verteilung der Frühjahrs- und der Herbsttermine nur zufällig unterscheidet, muß der FSI-Mittelwert im Herbst durchschnittlich höher liegen als im Frühjahr.

Bei Zusammenfassung der Fragen aus den jeweils 5 Terminen erhält man als FSI - Mittel für

	Herbst	Frühjahr
Anatomie	55,7 %	50,4 %
Physiologie	54,3 %	54,8 %
Physiol. Chemie	61,7 %	54,2 %

Die nach dem Unterschied in den Kandidatenpopulationen zu erwartende Differenz tritt deutlich zutage, wenn man vom Fach Physiologie absieht. In diesem Fach wird der Unterschied durch das mehr als 14 Prozentpunkte über dem sonstigen Mittel liegende Ergebnis der Prüfung vom Frühjahr 1984 verwischt. Die Fragen dieses Termins waren (bedingt durch das Ausscheiden der Sachverständigen der Physiologie) nicht mehr von einer regulären Fachkommission erarbeitet worden.

Wie hoch ist die Zufallsschwankung der Prüfungsschwierigkeit bei Prüfungen mit noch unbekannten Fragen vom Typ positive Einfachauswahl?

Als Maß für die zu erwartende Zufallsschwankung der Prüfungsschwierigkeit kann der Vertrauensbereich für den FSI-Mittelwert des Prüfungstermins dienen, der sich unter Annahme einer konstanten Leistungsfähigkeit der geprüften Population abschätzen läßt. Unter Zugrundelegung der Normalverteilungsannahme für den FSI ergibt sich die Standardabweichung des FSI-Terminmittelwerts vom wahren Mittelwert der Fragengrundgesamtheit aus der Standardabweichung der Einzelfragen-FSI geteilt durch die Wurzel aus der Zahl der Fragen pro Termin.

Bei den drei untersuchten Fächern, die etwa zwei

Drittel der Fragen der Ärztlichen Vorprüfung stellen, liegt die (auf eine konstant leistungsfähige Population bezogene) FSI-Standardabweichung der einzelnen Fragen höchstens bei 18 Prozentpunkten, denn in dem beobachteten Wert von 18 - 20 Prozentpunkten ist zusätzlich u.a. der Frühjahrs/Herbst - Unterschied in der Populationsleistung enthalten.
Wenn pro Termin 80 Fragen eines Faches gestellt würden, wäre von einer Standardabweichung des FSI-Mittelwerts von 18/Wurzel aus 80 = ca. 2 Prozentpunkten auszugehen. Bei Hochrechnung auf die 320 Fragen der Ärztlichen Vorprüfung wäre (unter Annahme einer vergleichbaren Einzelfragenstreuung in den anderen Fächern) die Standardabweichung des FSI-Mittelwerts auf nur einen Prozentpunkt zu schätzen.

Modellrechnung für eine Bestehensgrenze bei einer Prüfung aus unveröffentlichten Einfachauswahlaufgaben bei einer geforderten Prüfungsstoffkenntnis von 50 %

Bei Angebot einer richtigen zusammen mit vier falschen Alternativen wird die Lösung auch ohne Kenntnis des geprüften Gegenstands in 20 % der Fälle zufällig angekreuzt. Wenn eine zur statistischen Realisierung der Zufallschance ausreichende Fragenzahl angeboten wird, ist von einem Kandidaten, der die Hälfte des Prüfungsstoffs beherrscht, eine Prüfungsleistung von 60 % richtig beantworteter Fragen zu erwarten.
Bei einer derartigen Ermittlung der Bestehensgrenze sollte zugunsten des Kandidaten berücksichtigt werden, daß die Prüfung an seinem Termin zufällig schwieriger sein kann als an den sonstigen Terminen. Für die Ärztliche Vorprüfung mit 320 Fragen ist der 99 % - Vertrauensbereich des FSI-Mittelwerts als dreifache Standardabweichung, also etwa 3 Prozentpunkte, zu schätzen. Da die Bestehensgrenze in der Nähe des FSI-Mittelwerts liegt, könnten diese 3 Prozent auf die eigentlich zu fordernden 60 % richtiger Antworten angerechnet werden. Zusätzlich könnte als Vorhalt für nachträglich als ungeeignet erkennbare (z.B. nicht eindeutig lösbare) Fragen ein Vorwegabzug von zusätzlich zwei Prozentpunkten vorgesehen werden.
Bei einer so ermittelten Bestehensgrenze von 55 % richtiger Antworten würde (unter den Bedingungen der Modellrechnung) sichergestellt, daß praktisch alle Kandidaten, die die Lösung von 50 % der gestellten Fragen aufgrund ihrer Prüfungsstoffkenntnis finden, auch bestehen. Zugunsten der Zweifelsfälle würde inkauf genommen, daß die Prüfung auch von einigen bestanden wird, die nicht über die verlangte Prüfungsstoffkenntnis verfügen.

Resumé

Die hier vorgelegte, unter Verwendung von bisher eingesetzten Fragen der Ärztlichen Vorprüfung durchgeführte Analyse zeigt, daß bei Vermeidung systematischer Fehler eine quantitative, präzise und zuverlässige Messung der Prüfungsstoffkenntnisse mit Antwortauswahlaufgaben möglich ist. Eine nicht standardisierte m.c.-Prüfung gestattet dagegen nur Aussagen über die Rangordnung der Kandidaten nach ihrer Leistung an einem Termin.

Prof. Dr. G. Gebert,
Friedrich-Schneider-Str. 5, 6500 Mainz 1