Multiple Choice als Numerus clausus (4)

Prüfung nicht bestanden: die Kandidaten oder die Fragen?

Horst Kuni und Peter Becker (Marburg)

Zum Titel dieser Artikelserie haben die Autoren bisher konstatiert, 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); dass zu viele Medizinstudenten am Messfehler des Prüfungsintruments scheitern (Heft 5/80, Seite 292ff); dass es den schriftlichen Ärztlichen Prüfungen an Zuverlässigkeit und Gültigkeit fehle (Heft 6/80, Seite 345ff).

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 zwei vorangegangenen Beiträgen [11, 12] haben wir die Bedeutung der Mängel in der Zuverlässigkeit (Reliabilität) und insbesondere der Gültigkeit (Validität) der schriftlichen Prüfungen nach der Approbationsordnung für Ärzte dargelegt. Diese Mängel des Instruments an sich müssen ihre Ursachen in den einzelnen Prüfungsfragen haben.

In der Tat galt bislang das Hauptaugenmerk der Prüfungsanfechtungen Mängeln einzelner Fragen. Im Regelfall versuchte man nachzuweisen, dass die vom IMPP verlangte Antwort unrichtig und eine andere dagegen zutreffend sei, die Frage sich überhaupt nicht beantworten lasse oder mehrere Antworten als richtig anzuerkennen seien.

Diese Verfahren sind durch Gutachten und Gegengutachten ebenso aufwendig wie langwierig und führen selten zum Erfolg, weil dem Prüfungsamt nach der Rechtsprechung ein Beurteilungsspielraum für die Festlegung der Frageninhalte zugebilligt wird. Nur offensichtliche Konstruktionsfehler der Frage führten bisher zur Aufhebung. Hierzu verzichten wir auf konkrete Beispiele, da diese Art der Fragenkritik bereits Publizität erlangt hat [14].

Kombinationsaufgaben können die Misserfolgsquote steuern

Vielmehr wollen wir zunächst zeigen, wie *an sich richtige Fragen in Kombination unzulässig* sein können (aus [9], S. 70ff.):

"Die folgenden Angaben beziehen sich auf die Aufgaben 174, 175 und 176: Etwa zwei Wochen nach einem Infekt treten die in der Abbildung Nr. 3 (s. Beilage) dargestellten Hauterscheinungen auf". (Auf die Abbildung haben wir verzichtet, da dies für das hier demonstrierte Prinzip unwesentlich ist; die Autoren).

"174 Welche Erkrankung liegt hier am wahrscheinlichsten vor," etc…

"175 Welches der folgenden Symptome paßt zur Diagnose?" etc…

"176 Welche Angaben zur Behandlung sind richtig?" etc…

Es ist offenkundig: Wer Aufgabe 174 nicht beantworten kann, dem bleiben 175 und 176 unzugänglich. Ein Item hat angesichts der absoluten Bestehensgrenze und des vorgeschriebenen Bewertungssystems unlegitimiert eine Gewichtung mit dem Faktor 3 erhalten.

Um es am Extrem zu verdeutlichen: Hingen alle Fragen der Prüfung so von der ersten ab, bedeutete die Nichtbeantwortung dieser ersten Frage das Nichtbestehen der ganzen Prüfung. Durch Zahl und Umfang dieser nach dem jetzigen Bewertungssystem rechtswidrigen Kombinationsaufgaben kann also zusätzlich die Misserfolgsquote gesteuert werden.

Vermutlich haben Gerichtsurteile den Verfahrensablauf im IMPP in einer Weise beeinflusst, die bislang in keinem der Ergebnisberichte offiziell, sondern lediglich inoffiziell erwähnt wurde [10]:

Nach Eingang der Antwortbögen wird zunächst in einem Vorlauf eine Item-Analyse durchgeführt. Zu ihrem Verständnis seien vorab einige Begriffe der "klassischen" Testtheorie erläutert.

Schwierigkeits-Index und Trennschärfen-Index

Die relative Häufigkeit, mit der eine Frage die vom IMPP als richtig anerkannte Antwort erhalten hat, bezeichnet man als Schwierigkeits-Index p (an sich müsste er Leichtigkeits-Index genannt werden, denn eine hohe Antwortquote auf eine leichte Frage führt paradoxerweise zu einem hohen Schwierigkeits-Index).

Das Vermögen einer Frage, erfolgreichere von weniger erfolgreichen Kandidaten zu unterscheiden, misst man mit dem Trennschärfe-Index. Er drückt als Korrelationskoeffizient r den Zusammenhang der als richtig bewerteten Beantwortung der Testfrage mit dem gesamten Testergebnis aus.

Das IMPP berechnet den Trennschärfe-Index aus der punktbiserialen Korrelation, da eine Variable alternativ, die andere quantitativ verteilt ist [3-7]. Der Trennschärfe-Index ist wesentlich vom Schwierigkeits-Index abhängig. Es leuchtet unmittelbar ein, dass eine Aufgabe, die kein Kandidat gelöst hat (p=0), ebenso wenig wie eine Aufgabe, die jeder gelöst hat (p=1), erfolgreichere von weniger erfolgreichen unterscheiden kann.

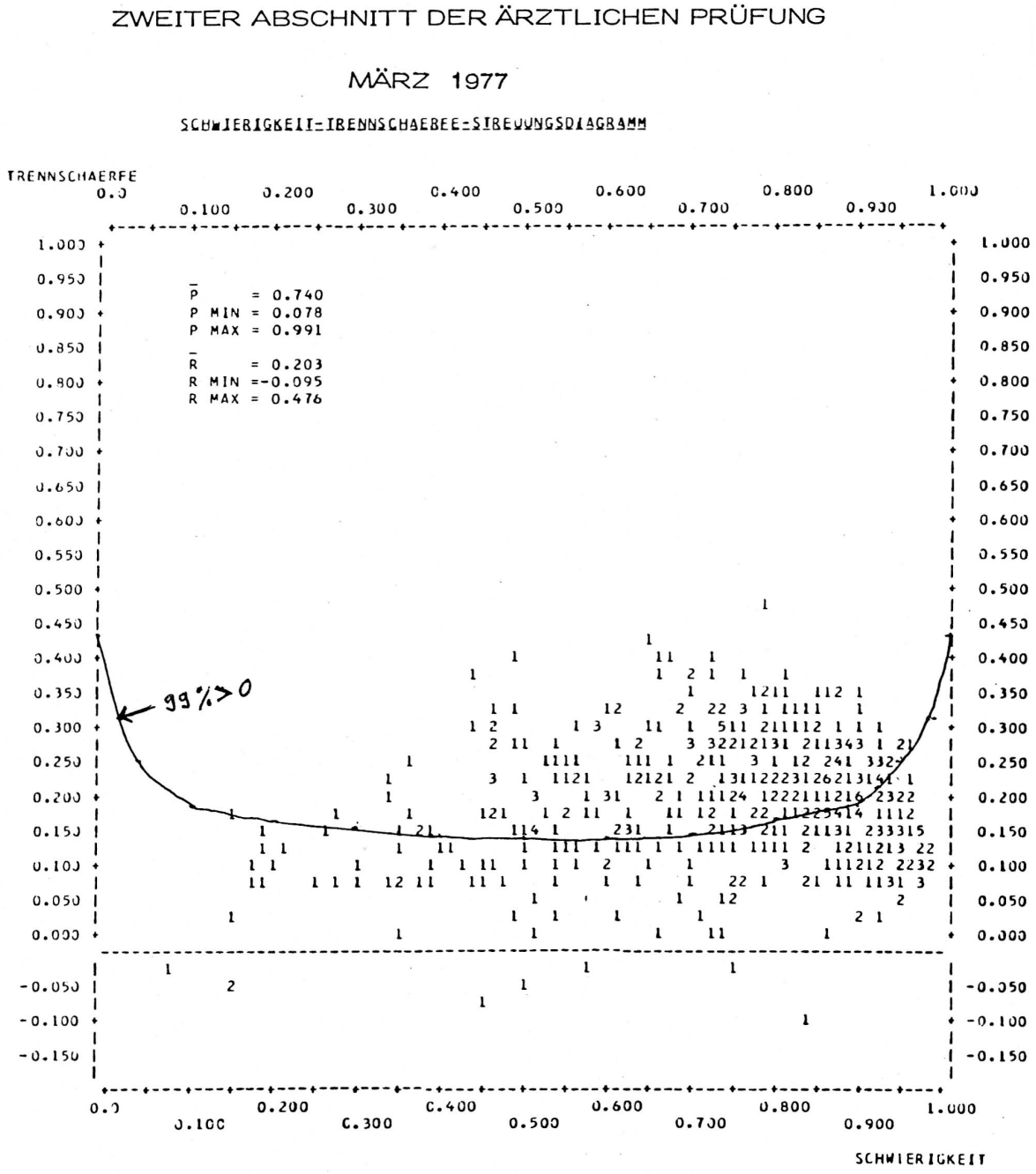

Abb. 1: Beispiel einer Korrelation von Trennschärfe-Index r und Schwierigkeits-Index p. In die Originalabbildung aus dem Ergebnisbericht des IMPP wurde zusätzlich die Signifikanzgrenze für 99 v. H.-Sicherheit gegen Null eingezeichnet. Ziffern über 1 bedeuten, dass die entsprechende Zahl von Aufgaben auf einen Koordinatenpunkt fällt. Eine negative Trennschärfe bedeutet, dass die Aufgabe(n) von den an sich erfolgreichen Kandidaten häufiger falsch und von den weniger erfolgreicheren Kandidaten häufiger als richtig beantwortet worden ist [7].

Bei den vom IMPP ausnahmslos gestellten 1 aus 5 Auswahl-Antwort-Aufgaben bedeutet schon ein p=0,2 im Durchschnitt Rateverhalten und damit in der Regel Fehlen der Trennschärfe. Bei p=0,5 kann die Trennschärfe ihr Maximum erreichen, von dem aus sie mit zu- und abnehmendem p parabelförmig

abfällt. Die in den Ergebnisberichten des IMPP dargestellten Korrelationen lassen diesen Sachverhalt erkennen (Abbildung 1).

Bei der Aufgaben-Konstruktion wird darauf geachtet, dass die vier als falsch gewerteten Antwort-Alternativen, die so genannten Ablenker oder Distraktoren, möglichst von gleicher Attraktivität sind. Denn zu offenkundig falsche Antworten führen im Endeffekt zu einer 1:4 (usw.)-Aufgabe und damit zu einer Erhöhung der Ratewahrscheinlichkeit. Um die Trennschärfe der Aufgabe zu optimieren, wird man für jeden Distraktor eine Antwortquote von etwa 10 v. H. anstreben.

Nicht nur die Kandidaten, auch die Fragen sind zu prüfen

Folgende Ergebnisse der Item-Analyse veranlassen das IMPP zu einer Aufgabenrevision [9]: Eine Trennschärfe <ca. 0,08-0,1, ein p<0,2, die Wahl eines Distraktors zu mehr als 20 bis 25 v. H., positive Trennschärfe eines Distraktors. Kommt man bei der Überprüfung zu dem Entschluss, eine andere Antwort als richtig oder als auch richtig zu bewerten, werden daraus nun bei der endgültigen Bewertung der Ergebnisse Konsequenzen gezogen.

So begrüßenswert es ist, dass trotz fehlender eindeutiger Vorgabe von Revisionskriterien durch die Approbationsordnung für Ärzte nicht nur die Kandidaten, sondern auch die Fragen geprüft werden, muss angesichts der absoluten Bestehensgrenze das Verfahren der Aufgabenrevision grundlegend erweitert werden. Die Trennschärfe kann nämlich durch untaugliche Formulierung des Items und der Distraktoren vermindert werden, ohne dass in den Augen der Experten formale oder inhaltliche Mängel offenkundig sind, aber trotzdem die erfolgreicheren Kandidaten häufiger zu falsch bewerteten Antwort-Alternativen verführt werden. Dies vermag das bisherige Verfahren der am Ende subjektiven Aufgabenrevision nicht zu erfassen.

Schließlich stellt das Antwortverhalten mehrerer tausend Studenten, die - gemessen am Abiturnoten-Durchschnitt - zur geistigen Elite der Nation gerechnet werden, für das Urteil über die sprachliche Verständlichkeit der Frage einen relevanteren Bezugspunkt dar als die Meinung eines kleinen Zirkels von in ihrem Fach hoch spezialisierten Experten, die zudem durch ihre Beteiligung an der Fragen-Herstellung in ihrem Urteil befangen sind.

Zu einer ins Grundrecht der freien Berufswahl eingreifenden Entscheidung darf aber nur eine Prüfungsfrage beitragen, deren Parameter ausgewiesen haben, dass sie mit der erforderlichen Sicherheit bei der Trennung von erfolgreicheren und weniger erfolgreichen Kandidaten mitwirkt. Das heißt: Ihr Trennschärfen-Index muss signifikant (wegen der Bedeutung der Entscheidung mit 99prozentiger Sicherheit, also bei dem hier vorliegenden Prüfungsfragenumfang der 2,59 bis 2,6fachen Standardabweichung) über 0 liegen.

Tab. 1: Anzahl der Prüfungsfragen, deren Trennschärfe-Index sich nicht mit 99 v. H. Sicherheit von Null unterschieden hat. Signifikanzberechnung nach [13]. Auszählung aus den vom IMPP publizierten Korrelationen (Beispiel Abbildung 1). Seit 3/77 wurden solche Korrelationen vom IMPP nicht mehr publiziert [3-7].

		Fragen mit zu niedrigem Trennschärfe-Index	
	Termin	Absolut	Relativ
Ärztliche Vorprüfung Gesamt: 300 Fragen	8/75	41	13,7 v. H.
	3/76	51	17,0 v. H.
	8/76	66	22,0 v. H.
	3/77	84	28,0 v. H.
1. Abschnitt der Ärztlichen Prüfung	3/75	131	54,6 v. H.
	8/75	112	46,7 v. H.
Gesamt: 240 Fragen	3/76	123	51,3 v. H.
	8/76	106	44,2 v. H.
	3/77	116	48,3 v. H.
2. Abschnitt der Ärztlichen Prüfung	8/76	209	41,8 v. H.
Gesamt: 500 Fragen	3/77	195	39,0 v. H.

Aus einer absoluten wird eine relativ Bestehensgrenze

Wie Abbildung 1 und Tabelle 1 zeigen, genügt eine hohe Zahl von Fragen dieser mit Recht strengen Anforderung nicht - ein Befund, der einer allgemeinen Erfahrung entspricht und nur bei einem anderen Bewertungssystem tolerierbar wäre. Leider können wir nicht errechnen, wie viele Kandidaten bislang an solchen untauglichen Fragen scheiterten; nach dem in Tabelle 1 gezeigten Anteil waren es sicher die meisten.

Der Zusammenhang von Trennschärfe-Index und Schwierigkeits-Index enthält aber noch ein weiteres Problem: Je besser die Ausbildung ist, je besser die Kandidaten sich vorbereitet haben, je mehr sie die Lehrziele erreichen, um so mehr wird p ansteigen und sich gegen 1,0 nähern. Dies reduziert zwangsläufig den Trennschärfe-Index der Aufgaben [2].

Das IMPP wird sich gezwungen fühlen, durch sophistischere Formulierungen der Ablenker bei der Konstruktion neuer Fragen den davoneilenden Schwierigkeits-Index in den Bereich von 0,5 bis 0,6 zu justieren, um den Trennschärfe-Index hoch zu halten. Damit wird aber eine absolute Bestehensgrenze im Endeffekt zu einer relativen pervertiert: Noch so gutes Lernen garantiert auf die Dauer keinen Prüfungserfolg.

Anhand einer kurzen Beobachtungszeit lässt sich dieser Effekt bereits beweisen, da mit dem neuen Prüfungssystem auch eine neue Ausbildungsordnung eingeführt wurde. Wird die Misserfolgsquote an der anfänglich völlig unzulänglichen Lehrsituation geeicht, zwingt die Verbesserung des Unterrichts durch die angestrebte Studienreform zur ständigen Adaption des Prüfungsinstrumentes. Wir zeigen diesen Effekt an Aufgaben, die das IMPP wiederholt gestellt hat.

Der Trennschärfe-Index ist populationsabhängig

Die "aktuellste" Aufstellung [8] zeigt, dass p zunimmt und der Trennschärfen-Index abnimmt. Nur die Aufgaben aus dem jeweils ersten Prüfungstermin einer Prüfungsart verhalten sich umgekehrt. Dies liegt an der Sonderpopulation von Studenten, die sich aufgrund der Übergangsbestimmungen als schmale Elite den jeweils ersten schriftlichen Prüfungen gestellt hat (wie das IMPP auch an anderen Parametern wie Studiendauer, Misserfolgsquote usw. gezeigt hat).

Scheidet man diese Werte aus der Berechnung aus, haben 65 Aufgaben der Ärztlichen Vorprüfung nach durchschnittlich 1,5 Jahren (!) ein um 14,8 v. H. höheres p und eine um 4,3 v. H. geringere Trennschärfe. Dagegen wurde mit Hilfe der übrigen Fragen für die gesamte Prüfung in diesem Zeitraum das p mit nur 0,75 v. H. Abweichung praktisch konstant gehalten und die Trennschärfe sogar um 3,3 v. H. gesteigert. Die Differenzen von p und r sind hochsignifikant.

Zusätzlich zeigt Tabelle 1 (neben der Auswirkung der Testlänge, das heißt der gesamten Fragenzahl einer Prüfung) auch die Populationsabhängigkeit des Trennschärfen-Index: Mit zunehmendem Studienfortgang und aufgrund vorangegangener Selektionen sind nicht mehr so große Unterschiede zwischen den Kandidaten zu verzeichnen. Die derzeitige Bewertungsregel lastet die Verschlechterung der Auftrennungsfähigkeit system- und rechtswidrig den besser gewordenen Kandidaten an.

Aufgabe zu schwierig - oder Kandidat nicht fähig?

Die falsche Anwendung der Parameter der "klassischen" Testtheorie maskiert also den Lernzuwachs der Studenten. Die bisherige Methodik ist systemimmanent nicht zur lehrzielorientierten Erfolgskontrolle geeignet [2]. Sie vermag aus der Beantwortung oder Nichtbeantwortung einer Frage nicht zu erkennen, ob die Aufgabe zu schwierig oder der Kandidat nicht fähig war.

Deshalb hat die moderne pädagogische und psychologische Forschung sich in den letzten zwanzig Jahren anderen Messmodellen zugewandt. Sie kehrt sich von der deterministischen Testtheorie ab

und berechnet durch propabilistische mehrdimensionale Schätzverfahren die Parameter für die Schwierigkeit der Aufgaben und Fähigkeit der Kandidaten separat [1].

Literatur

1. Fischer, G. H.: Einführung in die Theorie psychologischer Tests. Grundlagen und Anwendung, Verlag Hans Huber, Bern, Stuttgart, Wien 1974
2. Herbig, M.: Die Unzulänglichkeit der klassischen Testtheorie bei lehrzielorientierter Messung. In: Klauer, K. J. et al. Lehrzielorientierte Tests, Schwann Verlag, Düsseldorf, 1975
3. IMPP: Ergebnisbericht über die schriftlichen Prüfungen nach der Approbationsordnung für Ärzte, März 1975
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: Aufgabenheft, 2. Abschnitt der Ärztlichen Prüfung Herbst 1979, !. Tag, Auflage B
10. IMPP: Gespräch mit dem Marburger Bund am 24. April 1980
11. Kuni. H., Becker, P.: Multiple Choice als Numerus clausus (2) Zu viele Medizinstudenten scheitern am Messfehler "der arzt im krankenhaus" (1980) 292-293
12. Kuni. H., Becker, P.: Multiple Choice als Numerus clausus (3) Den schriftlichen Ärztlichen Prüfungen fehlt die Gültigkeit "der arzt im krankenhaus" (1980) 345-358
13. Lienert, G. A.: Testaufbau und Testanalyse. Verlag J. Beltz, Weinheim, 1961
14. Spiegel: Viel Freiheit, Nr. 14 (1980) 44

Anschrift der Verfasser

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

Rechtsanwalt Dr. Peter Becker, Gisonenweg 9, 35037 Marburg