Multiple Choice als Numerus clausus (2)

Zu viele Medizinstudenten scheitern am Messfehler des Prüfungsinstruments

Horst Kuni und Peter Becker (Marburg)

Intro der Redaktion: Im ersten Beitrag dieser Artikelserie (Heft 4/80, Seite 194ff) stellten die Autoren "als evident" fest: "Multiple Choice als Numerus clausus, das heißt Fortsetzung der Numerus-clausus-Politik mit anderen Mitteln. Das Prüfungsverfahren ist damit in eine grundgesetzwidrige Handhabung abgeglitten".

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 auch von ihm vor allem wegen seiner "konkurrenzlosen Objektivität" bei der Einführung 1970 begrüßt wurde.

Nachdem wir in unserem ersten Beitrag [7] evident gemacht haben, dass die schriftlichen ärztlichen Prüfungen seit Herbst 1979 als verfassungswidriges Selektionsinstrument missbraucht werden, könnte der Eindruck entstehen, dass die Prüfungen bis dahin einwandfrei durchgeführt worden wären. Wenn wir nun dieser Frage nachgehen, wollen wir uns nicht mit einzelnen Pannen aufhalten [11], sondern grundsätzliche Perspektiven betrachten.

Die Approbationsordnung für Ärzte fordert in Paragraph 14 Absatz 2 ausdrücklich, dass die Prüfungsfragen zuverlässige Prüfungsergebnisse ermöglichen müssen. Zuverlässigkeit (Reliabilität) der Prüfung ist ein wesentlicher Begriff der "klassischen" Testtheorie. Obwohl wir uns diese Betrachtungsweise keineswegs unkritisch zu Eigen machen können, wollen wir zunächst die ärztlichen Prüfungen in ihrem Licht untersuchen [8].

Der Verordnungsgeber selbst hat in der amtlichen Begründung gefordert, dass die Prüfungen die Kriterien der Objektivität, der Zuverlässigkeit und Gültigkeit mitbringen müssen, auf die Arbeit von Kapuste und Noack [4] verwiesen und für die Herstellung der Fragen Kenntnisse der Testtheorie vorausgesetzt. Schließlich wurden diese Maßstäbe auch vom Medical Board in den USA, dem großen Vorbild der deutschen Prüfung, angelegt [10]. Das IMPP hat bis heute ausschließlich seine Analysen auf dem Boden dieser Testtheorie durchgeführt [3].

Grundlagen der klassischen Testtheorie

Zum besseren Verständnis der folgenden Aussagen seien kurz die Grundlagen der klassischen Testtheorie skizziert.

Die Prüfung kann als spezieller Test aufgefasst werden. Es soll ein Persönlichkeitsmerkmal (zum Beispiel ein umrissenes Faktenwissen), das zunächst latent, also nicht ohne weiteres erkennbar, existiert und als "wahres" Merkmal T (true score) vorausgesetzt wird, messen. Die Messung führt zu einem beobachtbaren Ergebnis, der manifesten Variablen X. Da es keine Messung ohne Fehler gibt, wird aber das Messergebnis auch noch die Auswirkung des Messfehlers enthalten. Die Ursache des Messfehlers kann zum Beispiel in einer unterschiedlichen Tagesverfassung eines Prüflings liegen [9].

Bei wiederholter Testung (Übungs- und Ermüdungseffekte seien ausgeschlossen) würde die manifeste Variable um einen Mittelwert streuen, der mit wachsender Zahl der Messungen einer immer besseren Abschätzung des wahren Wertes entspricht. Die Größe dieser durch den Messfehler verursachten Streuung kann als Varianz ausgedrückt werden, das ist der Durchschnitt der quadratischen Abweichung der Einzelergebnisse vom Mittelwert (die Quadrierung erlaubt es, bei der Durchschnittsbildung positive und negative Werte gleich zu behandeln). Die klassische Testtheorie geht davon aus, dass der Messfehler vom "wahren" Wert des Merkmals selbst unabhängig ist.

Diese Annahmen überträgt die klassische Testtheorie nun auf die gesamte Kandidatenpopulation in einer Prüfung und fasst die individuellen Unterschiede des "wahren" Merkmals als Varianz dieses Merkmals σ_T^2 auf. Nach der schon von Gaus formulierten Gesetzmäßigkeit setzt sich dann die im Testergebnis gefundene Varianz σ_X^2 aus der des "wahren" Merkmals σ_T^2 und der des Messfehlers σ_E^2 zusammen.

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 \tag{1}$$

Die Zuverlässigkeit eines Tests drückt die klassische Testtheorie im Anteil aus, den die Varianz des "wahren" Wertes an der Varianz des Testergebnisses hat. Diese Verhältniszahl nennt man Reliabilitätskoeffizient r_{tt}.

$$r_{tt} = \frac{\sigma_T^2}{\sigma_X^2} \tag{2}$$

Eine Umformung der obigen Gleichung (1) ergibt

$$\sigma_E^2 = \sigma_X^2 - \sigma_T^2 \bullet \frac{1}{\sigma_X^2}$$

$$\frac{\sigma_E^2}{\sigma_X^2} = 1 - \frac{\sigma_T^2}{\sigma_X^2}$$

$$\sigma_E^2 = (1 - \frac{\sigma_T^2}{\sigma_X^2}) \bullet \sigma_X^2$$

Substituiert man nun durch die Formel (2), erhält man

$$\sigma_E^2 = (1 - r_{tt}) \bullet \sigma_X^2$$

das heißt aus der Varianz der Testergebnisse und dem Reliabilitätskoeffizienten lässt sich die Varianz des Messfehlers berechnen. Die Quadratwurzel der Varianz nennt man Standardabweichung s. Aus der Standardabweichung lässt sich bei Annahme einer sogenannten Normalverteilung abschätzen, mit welcher Wahrscheinlichkeit ein Messwert um ein bestimmtes Ausmaß vom "wahren" Wert abweicht, zum Beispiel kann ein Messwert in 99 v.H. der Fälle bis zum 2,5fachen der Standardabweichung vom "wahren" Wert (nach beiden Seiten) entfernt sein.

$$s_E = \sqrt{1 - r_{tt}} \bullet s_X$$

Wie zuverlässig ist das Prüfungsinstrument?

Um den Messfehler eines Tests zu bestimmen, muss man also seine Standardabweichung und den Reliabilitätskoeffizienten empirisch ermitteln.

Die theoretisch besten Methoden, um den Reliabilitätskoeffizienten als Maß der Übereinstimmung eines Tests mit sich selbst zu errechnen, die Paralleltest- oder Wiederholungstestmethode, sind bei einer Prüfung selbstverständlich nicht möglich. Man muss deshalb die Testteilungsmethode anwenden. Dabei teilt man den Test in zwei möglichst gleichwertige Hälften und stellt deren Übereinstimmung fest.

Die Gleichwertigkeit ist natürlich recht unsicher, und das Ausmaß der Übereinstimmung wird durch die nun geringere Zahl der Items kleiner ausfallen. Die vom IMPP angewandte Konsistenzanalyse treibt diesen Gedanken weiter: Der Test wird in so viele Teile geteilt, wie er Items hat, und das mittlere Ausmaß der Übereinstimmung aller Items errechnet.

Das Ergebnis war nach den Maßstäben der klassischen Testtheorie für statistische Zwecke sehr befriedigend und auch für eine individuelle Beurteilung prinzipiell ausreichend, wobei der Wert im Dritten Abschnitt der ärztlichen Prüfung allerdings knapp unter der Grenze liegt, die das Medical Board in den USA mit 0,9 als Minimalwert für diesen Zweck fordert.

Damit bestätigt sich auch hier, dass die schriftliche Prüfung nicht nur in ihrer Objektivität, sondern auch in der Zuverlässigkeit mit einem r_{tt} um 0,9 weit besser ist als mündliche Prüfungen, für die sich in einer Literaturübersicht Werte zwischen 0,3 und 0,6 für Einzelprüfer und 0,8 für Teams fanden [5].

Dieser Gewinn an Zuverlässigkeit darf jedoch nicht darüber hinwegtäuschen, dass auch das neue Prüfungsinstrument nicht absolut zuverlässig ist. Ein erfahrener Prüfer bezieht die selbstkritische Einschätzung seiner eingeschränkten Zuverlässigkeit in die Bewertung des Probanden mit ein; der Computer in Mainz kann es nicht, solange eine starre Bewertungsregel der AOÄ dies unterlässt.

Häufig trifft man das Vorurteil an, der Prüfling habe zu zeigen, was in ihm steckt, habe zu beweisen, dass er die geforderte Qualifikation hat. Dies hängt mit der Praxis eines guten Prüfers zusammen, dem Kandidaten Gelegenheit zu geben, seine Fähigkeiten zu einer Fragestellung frei zu entfalten und möglichst so darzustellen, dass ein individuell optimaler Eindruck entsteht.

Der schriftliche Test gibt dem Probanden keinerlei Chance zur freien Entfaltung. Diejenigen der latenten Kenntnisse, die der Test auf Grund mangelnder Zuverlässigkeit nicht manifestiert, bleiben verborgen. Der Tester hat die Pflicht, zu beweisen, mit welcher Wahrscheinlichkeit nicht gefundene Kenntnisse tatsächlich nicht vorhanden sind. Denn nach überwiegender Wahrscheinlichkeit (mindestens mit rund 97,5 v.H. der Fälle [8]) ist zu erwarten, dass ein Kandidat (zumal mit einer eher überdurchschnittlichen Abiturnote) nach Überwindung aller Schwierigkeiten bei der Zulassung zum Studium und während des Studiums auch mindestens ausreichende Kenntnisse erworben hat.

So wie es zum Beispiel anerkannter Grundsatz der Rechtsprechung ist, dass bei der Feststellung der Überschreitung eines Tempolimits der Messfehler des Radars zugunsten des angeschuldigten Autofahrers zu berücksichtigen ist, gehört es auch zu anerkannten Bewertungsregeln, dass eine Bestehensgrenze von dem zunächst für den Idealfall einer absoluten Zuverlässigkeit festgesetzten Wert um den jeweils empirisch gefundenen Messfehler zugunsten des Kandidaten zu verschieben ist [2, 6].

Elementarer Grundsatz vom Gesetzgeber nicht beachtet

Bei der individuellen Schicksalsentscheidung "Prüfung bestanden / nicht bestanden" muss gefordert werden, dass die Bestehensgrenze mit mindestens 99 v.H. Sicherheit unterschritten wurde, das heißt, es muss ein Irrtumsbereich vom 2,58fachen des Messfehlers berücksichtigt werden.

Eine große Zahl von Medizinstudenten hat nur deshalb die Prüfung nicht bestanden, weil der Verordnungsgeber diesen elementaren Grundsatz nicht beachtet hat. Dass der größere Teil der Wiederholer schließlich die Prüfung, wenn auch nach dem Verlust mindestens eines halben Studienjahres, bestanden hat, unterstreicht das Gewicht dieser Feststellung.

Man beachte, dass bei Anheben der absoluten Bestehensgrenze von 50 v.H. auf 60 v.H. die absolute Zahl der am Messfehler scheiternden Studenten zunimmt, der relative Anteil an der Misserfolgsquote aber kleiner wird, so dass auch von hier aus ein ungünstigerer Ausgang der Wiederholungsprüfungen zu erwarten ist.

Betrachten wir zum Abschluss zwei wichtige Einflussgrößen der Zuverlässigkeit: die Testlänge (Zahl der Prüfungsfragen, der sogenannten Items) sowie die Varianz des Merkmals in der geprüften Population.

Die Zweite Änderungsverordnung der Approbationsordnung vom 24. Februar 1978 hat die Zahl der Prüfungsfragen verändert. Eine Abschätzung der sich daraus ergebenen Änderung des Messfehlers lässt ausgerechnet bei der Prüfung mit der höchsten Misserfolgsquote die geringste Verbesserung und bei der letzten ärztlichen Prüfung sogar eine beachtliche Verschlechterung erwarten. Hier wird r_{tt} eindeutig unter die Mindestanforderung sinken. An dem Imperativ der Berücksichtigung des Messfehlers zugunsten des Kandidaten kann man mit diesen Auszieselierungen des Messinstrumentes nicht rütteln.

Negative Folgen der Populationsabhängigkeit

Ein Rückblick auf die oben angeführten Formeln lässt erkennen, dass mit abnehmender Varianz in der Kandidatenpopulation der Anteil der Varianz des Messfehlers an der Gesamtvarianz zunimmt und damit der Reliabilitätskoeffizient absinkt. Eine durch vorangegangene Vorprüfungen "homogenisierte" Population wird also in späteren Prüfungen zwangsläufig eine schlechtere Reliabilität auch bei gleicher Testlänge und gleicher Qualität der Fragen produzieren.

Je mehr dieser Effekt durch schärfere Selektionen verstärkt wird oder je mehr Studenten durch bessere Studienbedingungen das Lehrziel erreichen, um so mehr wird die "wahre" Varianz zurückgehen - und um so mehr wird die Varianz des Prüfungsergebnisses nur noch die Varianz des Instrumentes, seines Messfehlers beinhalten: Die Zuverlässigkeit, gemessen mit dem Reliabilitätskoeffizienten, wird gegen Null tendieren!

Orientiert man sich, wie bisher das IMPP, ausschließlich an den Axiomen der klassischen Testtheorie, wird man versuchen, durch immer weitere Verlängerung des Messinstrumentes der schwindenden Reliabilität nachzulaufen.

Falls nicht die überfällige Berücksichtigung von Testtheorie nach dem heutigen Stand der Wissenschaft [1] Platz greift, ist abzusehen, dass auf dem Rücken der Studenten bald das amerikanische Übermaß von 800 bis 1000 Fragen in Teil I des Medical Boards (inhaltlich etwa unserer Vorprüfung und dem Ersten Abschnitt der ärztlichen Prüfung entsprechend) abgeladen wird, dem wir uns seit Herbst 1979 mit einem Schritt von 540 auf 610 Fragen weiter genähert haben.

Diese Populationsabhängigkeit gebietet es als ein weiterer rechtlicher Grund, den Messfehler zu berücksichtigen. Es stellt einen massiven Verstoß gegen das Übermaßverbot und den Gleichheitsgrundsatz dar, wenn das Bestehen einer Prüfung deshalb von einem wachsenden Messfehler gefährdet wird, weil der Kandidat auf Grund zuvor bestandener selektierender Prüfungen und guter Lernleistungen einer Population mit homogeneren Leistungen angehört.

Weil die klassische Testtheorie die Varianzbetrachtung vom Individuum auf die Population überträgt, erhält der Messfehler also eine andere Dimension. Auch das Vermögen der Fragen, ein unterschiedlich ausgeprägtes Merkmal überhaupt anzusprechen, geht wesentlich ein.

Diesen mehr inhaltlichen Aspekt werden wir im nächsten Beitrag untersuchen, der auch eine Tabelle zur vorstehenden Untersuchung enthält.

(Anmerkung beim Nachdruck: Da wir den äußeren Beschränkungen der Monatszeitschrift beim Nachdruck nicht unterliegen, können wir die zu diesem Kapitel gehörende Tab. 1 (S. 7), die im Originaltext mit dem Kapitel 3 abgedruckt worden war, hier bringen.)

Tabelle 1 (S. 7) :
Maßzahlen für die (Un-)Zuverlässigkeit der schriftlichen ärztlichen Prüfungen am Beispiel der Ergebnisse im Herbst 1977 (letzter bisher vorliegender ausführlicher Ergebnisbericht des IMPP), die weitgehend repräsentativ sind.

Tabelle 1:

			Ärztliche Vorprüfung	Abschnitte der Ärztlichen Prüfung		
				1	2	3
1	Reliabilitätskoeffizient r_{tt}		0,964	0,926	0,950	0,889
2	Messfehler s_E [Rohwerte]		7,5	7,0	9,2	5,8
3	99 % Vertrauensbereich absolut		19	18	24	15
4	99 % Vertrauensbereich relativ		6,3	7,5	4,8	6,3
5	Misserfolg	< 50%	11,3	5,2	1,4	0
6	Misserfolg mit 99% Sicherheit	< 50%	5,1	1,4	0,4	0
7	Misserfolg	< 60%	29,7	21,1	12,9	2
8	Misserfolg mit 99% Sicherheit	< 60%	17,1	7,9	5,4	0
9	Testlänge alt		300	240	500	240
10	Testlänge neu		320	290	580	180
11	Reliabilitätskoeffizient r_{tt}		0,966	0,938	0,957	0,857
12	Messfehler s_E [Rohwerte]		7,3	6,4	8,5	6,6
13	99 % Vertrauensbereich absolut		19	16	22	17
14	99 % Vertrauensbereich relativ		5,8	5,6	3,8	9,4

Zeile **Erläuterung**

1 Reliabilitätskoeffizient nach Angaben des IMPP (berechnet mit der Kuder-Richardson-Formel 20)

2 Messfehler, berechnet aus r und der Standardabweichung nach Angaben des IMPP, Angabe in Rohwerten, das heißt Anzahl der Items.

3 99% Vertrauensbereich in Rohwerten (einseitig, d.h. Abstand einer Seitenschranke vom Mittelwert)

4 wie 3, jedoch in Prozent der gesamten Aufgabenzahl einer Prüfung (siehe Zeile 9)

5 Tatsächliche Misserfolgsquote auf Grund der alten 50 v. H. - Bewertungsregel in Prozent der Kandidaten.

6 Geschätzter Misserfolg in Prozent, wobei nur Kandidaten berücksichtigt wurden, die mit 99% Sicherheit unter der 50 v. H. - Schranke lagen.

7 Geschätzter Misserfolg in Prozent bei hypothetischer Anwendung der neuen 60 v. H. - Bewertungsregel

8 Geschätzter Misserfolg in Prozent, wie 6 jedoch für 99% Sicherheit unter der 60 v. H. - Schranke

9 Damalige Testlänge (Anzahl der Prüfungsfragen)

10 Seit Herbst 1979 angewandte Testlänge

11 -14 Dieselben Maßzahlen wie Zeilen 1-4, jedoch wurde r anhand der Formel

$$r_{ttneu} = \frac{r_{ttalt} \bullet K}{1 + r_{ttalt}(K-1)} K$$

aus dem bisherigen r_{tt} (Zeile 1) und dem Verlängerungsfaktor K abgeschätzt [1]

Literatur

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

2. Fricke, R.: Lehrzielorientierte Messung mithilfe stochastischer Meßmodelle. In: Klauer, K. J. et al. Lehrzielorientierte Tests, Schwann Verlag, Düsseldorf, 1975

3. IMPP: IMPP: Ergebnisberichte über die schriftlichen Prüfungen nach der Approbationsordnung für Ärzte, März 1975 bis August 1977

4. Kapuste, H., Noack,H.: Leistungstests in der Medizin – Möglichkeiten und Grenzen, Institut für Ausbildungsforschung, München

5. Kerschbaum, Th., Flörkemeier, T.: Die Zuverlässigkeit und Gültigkeit von Prüfungen. Objektivierte Leistungskontrollen in der medizinischen und zahnmedizinischen Ausbildung(II), Dtsch. Ärztebl. (1974) 71, 2133-2139

6. Klauer, K. J.: Zur Theorie und Praxis des binomialen Modells lehrzielorientierter Tests. In: siehe [2]

7. Kuni. H., Becker, P.: Multiple Choice als Numerus clausus (1). "der arzt im krankenhaus" (1980) 194-200

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

9. Jord, F. M., Novick, M. R. (Hrsg.): Statistical theories of mental test scores. Addison-Wesley, Reading / Mass. 1968

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

11. Stegmaier, A.: Angriffspunkte, Dtsch. Ärztebl. (1979) 2986

Anschrift der Verfasser

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

Rechtsanwalt Dr. Peter Becker, Gisonenweg 9, 35037 Marburg