Multiple Choice als Numerus clausus (8)

Absolute Leistungsmessung - Utopia?

Horst Kuni und Peter Becker (Marburg)

In den sieben bisher erschienen Artikeln dieser Serie stellten die Autoren fest: Das System der Multiple-Choice-Prüfungen setzt die Numerus-clausus-Politik "mit anderen Mitteln" fort (Heft 4/80, Seite 194ff.); zu viele Medizinstudenten scheitern am Messfehler des Prüfungsinstruments (Heft 5/80, Seite 292ff.); den schriftlichen Ärztlichen Prüfungen fehlt es an Zuverlässigkeit und Gültigkeit (Heft 6/80, Seite 345ff.); nicht nur die Kandidaten, sondern auch die Fragen (Heft 7/80, Seite 406ff.) und das Prüfungsinstitut selbst (Heft 8/80, Seite 475ff.) müssen der Prüfung unterliegen; zu prüfen ist nur exemplarisch unverzichtbares Basiswissen (Heft 9/80, Seite 522ff.); die Bestehensregel kann nicht bestehen (bleiben) (Heft 10/80, Seite 590ff.).

Die Autoren veröffentlichen ihre Überlegungen und Argumente in der Monatsschrift des Marburger Bundes, weil der Verband an der Reform des Medizinstudiums nicht unwesentlich beteiligt war und weil das neue Prüfungssystem bei der Einführung 1970 auch von ihm vor allem wegen seiner "konkurrenzlosen Objektivität" begrüßt wurde.

Im letzten Beitrag haben wir gezeigt, dass bei richtiger Anwendung die in den USA praktizierte normbezogene Leistungsmessung mit relativer Bewertungsschranke so abgewandelt werden kann, dass sie in pragmatischer Weise eine mit der deutschen Standard-Schulnotenskala vergleichbare und rechtlich vertretbare Bewertung der schriftlichen Ärztlichen Prüfung erlaubt [8].

Zwei Gründe zwingen uns allerdings zu einer ernsthaften Auseinandersetzung mit der kriterienbezogenen Leistungsmessung nach einer absoluten Bewertungsregel:

1. Die bereits dargelegten Einwände gegen die "klassische Testtheorie" schließen auch die normbezogene Leistungsmessung ein. Ihre Anwendung wäre deshalb nur als Übergangslösung zu tolerieren, bis nach Ablösung der jetzigen, rechtswidrigen Praxis die Voraussetzungen für eine korrekte kriterienbezogene Leistungsmessung erarbeitet sind.

2. Didaktische, psychologische und juristische Argumente sprechen dafür, dass eindeutige Wissenskriterien gesetzt werden, deren ausreichendes Vorhandensein als Voraussetzung zur Zulassung zum Arztberuf in einer staatlichen Prüfung gemessen wird.

Die in der Approbationsordnung vom 28. Oktober 1970 ursprünglich zur relativen Bewertungsregel hinzugefügte absolute Fünfzig-Prozent-Bestehensregel ist, wie schon aus der amtlichen Begründung hervorgeht, keineswegs im Sinne einer kriterienbezogenen Leistungsmessung gesetzt worden, wie

das IMPP unterstellt [5]. Vielmehr hat man diese "Auffangposition" empirisch so konstruiert, dass die bisherigen Misserfolgsquoten mündlicher Physikums-Prüfungen nicht überschritten werden sollten.

Aus den bereits geschilderten Gründen hatten sich sowohl in den M.C.-Prüfungen des Medical Board als auch in einer in Deutschland von *Koransky* promovierten Erprobung des Systems im Fachgebiet Pharmakologie Erfahrungen über die Durchschnittswerte und Verteilungen der Rohwerte herausgebildet, aus der heraus man die Fünfzig-Prozent-Grenze als die richtige Justierung ableitet [1, 7].

Der Fragentypus wurde nicht festgelegt

Wie befangen der Verordnungsgeber bei dieser empirischen Adaptation war, kann man auch daran erkennen, dass (rechtsfehlerhaft) versäumt wurde, mit der Fünfzig-Prozent-Grenze zugleich auch den Fragentypus festzulegen. Angenommen, das IMPP hätte ausschließlich eine aus zwei M.C.-Fragen gestellt, hätte bereits die Ratewahrscheinlichkeit das Bestehen der Prüfung ermöglicht. Die Fünfzig- und auch die Sechzig-Prozent-Grenze sind also schon deshalb nur eine scheinbar objektive Leistungsvorgabe, weil allein durch die Zahl der Antwort-Alternativen die Ratewahrscheinlichkeit und damit die Schwierigkeit der Prüfung und damit die Misserfolgsquote gesteuert werden können. Schon aus diesem Grund muss im jetzigen Bewertungssystem das IMPP am "Eine aus fünf M.C.-Fragentyp" starr festhalten, obwohl eine abweichende Zahl von Antwort-Alternativen sinnvoller wäre.

Es bringt immer wieder Verwirrung in die Diskussion, wenn die Verfechter einer kriterienbezogenen Leistungsmessung darauf verweisen, dass in den USA (auch) mit einem absoluten Standard unter Verwendung des schriftlichen Prüfungsinstruments gemessen werde und dabei die Anforderungen viel höher als in Deutschland seien.

So hat der Westdeutsche Medizinische Fakultätentag am 16./17. Juni 1970 in Mainz beschlossen, die Klassifizierung "Nicht bestanden", "Bestanden" und "Mit Auszeichnung bestanden" vorzuschlagen. Das bezog sich offenkundig auf die amerikanische Tradition, bei 75 v. H. richtigen Antworten das Bestehen und bei 88 v. H. richtigen Antworten das Bestehen mit Prädikat (honor) festzustellen.

Dabei hatte *Hubbard* bei dem Sachverständigen-Gespräch am 16. September 1969 im Bundesministerium für Jugend, Familie und Gesundheit nochmals erläutert, was bereits in seinem Buch nachzulesen war: Mit Hilfe der vom Medical Board festgesetzten Misserfolgsquote (zum Beispiel drei v. H.) und der Quote, wie viel Prozent der Kandidaten mit Auszeichnung bestehen sollen (hier fehlen konkrete Angaben), wurden mit Hilfe eines Systems aus zwei linearen Gleichungen mit zwei Unbekannten die Rohwerte der Prüfungsergebnisse so umgerechnet, dass den traditionellen Schranken derart angepasste Skalenwerte resultierten: drei v. H. der Kandidaten wurden eben als "nicht bestanden" bewertet, weil sie den Skalenwert 75 nicht erreichten, und die festgelegte Quote der Kandidaten wurde mit "bestanden mit Prädikat" bewertet, weil sie den Skalenwert 88 überschritten.

Auf diese Weise wurde also trotz der in den Prüfungsordnungen mancher Staaten der USA gesetzlich fixierten absoluten Schranken eine normbezogene relative Leistungsbeurteilung praktiziert [1, 10]!

Kein unmittelbarer Rückschluss auf das Wissen

Häufig trifft man in der Diskussion um die absolute Bestehensgrenze auch unter Hochschullehrern die laienhafte Ansicht an, der in der Prüfung erzielte Anteil richtiger Antworten würde einen unmittelbaren Rückschluss auf das Wissen des Kandidaten erlauben. Der Leser, der unseren Beiträgen bis hierher gefolgt ist, weiß natürlich, dass die Schwierigkeit der Aufgaben diesen Zusammenhang wesentlich modifizieren kann. Bei entscheidenden Prüfungen ist zu verlangen, dass der Kandidat nur besteht, wenn er nicht weniger als zu 95 v. H. das Lehrziels (hier das staatlich verlangte Mindestwissen) erreicht hat [6].

Ist eine Prüfung schwierig konstruiert, wie die des IMPP, können Kandidaten, die gerade noch 95 v. H. des Mindestwissens haben, schätzungsweise nur 50 v. H. der Aufgaben lösen. Sehr gute Kandidaten mit nahezu 100 v. H. des Mindestwissens erreichen nicht viel mehr als 80 bis 85 v. H. richtiger Lösungen.

Abbildung 1 (S. 4) zeigt, dass die daraus resultierende Kennlinie für den Zusammenhang zwischen der Zahl richtiger Antworten und dem Erreichen des Lehrziels sowohl im Bereich einer absoluten Bestehensgrenze als auch in ihrer Umgebung sehr empfindlich misst.

Eine solche Lage der Kennlinie hat aber den erheblichen physiologischen Nachteil, dass Kandidaten, Öffentlichkeit, Arbeitgeber, Professoren, Ausland usw. aus den niedrigen Erfolgsquoten der richtigen Antworten falsche Schlüsse ziehen. Es sei hier nur an das Schlagwort erinnert in Bezug auf die Fünfzig v. H.-Bestehensregel, die Studenten würden mit "Halbwissen" auf den Patienten losgelassen.

Wir haben zur Verdeutlichung eine Kennlinie aufgezeichnet, die durch eine zu leichte Prüfung tatsächlich 50 v. H. richtige Antworten bei 50 v. H. Erreichen des Lehrziels liefert. Man sieht, dass auch bei Annahme einer wesentlich schlechteren Trennschärfe der Prüfung (Abflachung der Kennlinie) schon Kandidaten mit noch völlig unzureichendem Wissen (etwa 70 v. H. Erreichung des Lehrziels) 100 v. H. richtige Antworten erzielen würden.

Der richtige Weg läge wohl darin, die Schwierigkeit der Prüfung so zu justieren, dass die Bestehensgrenze von 95 v. H. Erreichen des Lehrziels mit etwa 75 v. H. richtiger Antworten überschritten wird (wobei natürlich immer noch der Messfehler der Prüfung mit 3s zusätzlich zu berücksichtigen wäre).

Diese Kennlinie hätte folgende Vorteile:

1. Sie erlaubt noch eine ausreichend genaue Messung an der Bestehensgrenze.
2. Sie gestattet noch sowohl eine Differenzierung innerhalb der erfolgreichen Kandidaten als auch die Angabe, wieweit die Distanz eines Erfolglosen von der Bewertungsgrenze ist.
3. Sie entspricht einem traditionellen (amerikanischen) Standard von 75 v. H. richtiger Antworten.
4. Sie vermeidet Fehlurteile in der Einschätzung des Prüfungsergebnisses durch Laien.

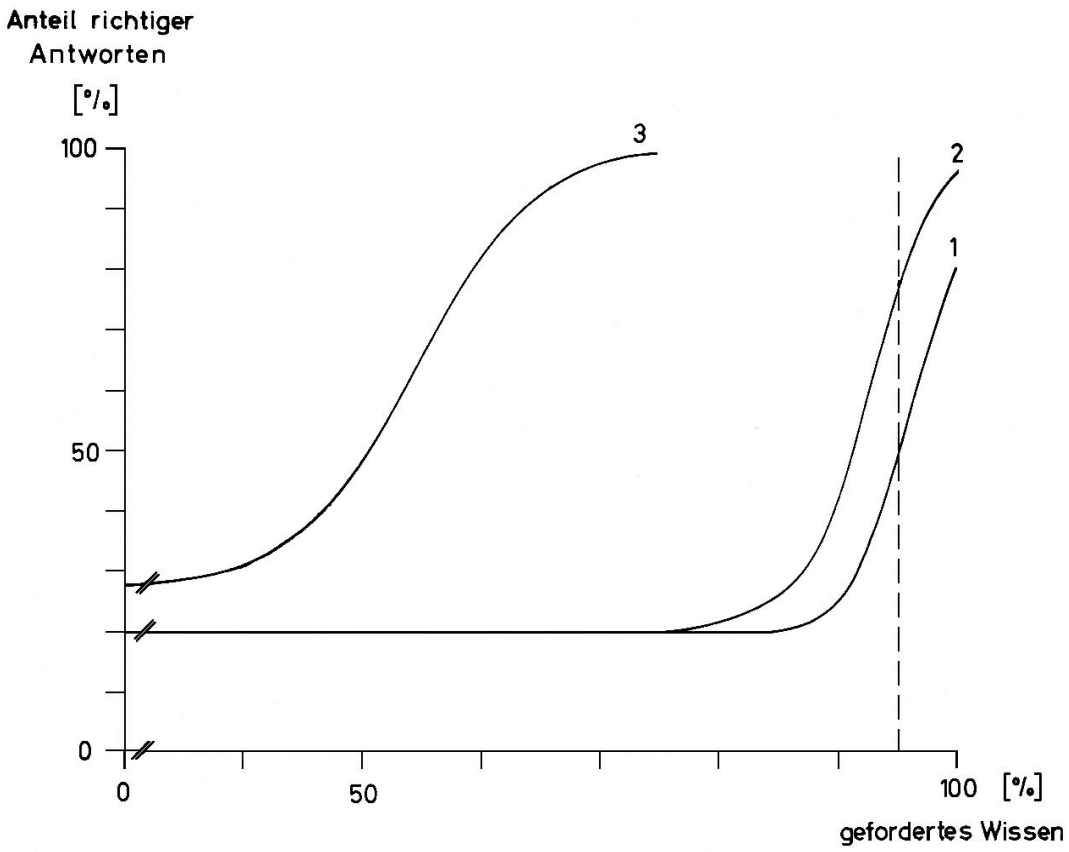

Abb. 1: **Schematische Skizze für verschiedene Kennlinien, die zeigen, wie der Anteil richtiger Antworten in einer M.C.-Prüfung mit einer aus fünf Aufgaben davon abhängt, inwieweit der Kandidat das geforderte Mindestwissen besitzt**
Die Ratewahrscheinlichkeit von 20 v. H. ist dann die Basislinie. Je schwieriger die Prüfung ist, umso mehr wird die Kennlinie nach rechts verschoben. Je schlechter die Trennschärfe der Prüfung ist, umso flacher wird die Kurve.
1 entspricht etwa dem Schwierigkeitsgrad der IMPP-Prüfungen;
2 günstigere Kennlinie einer etwas leichteren Prüfung (näheres siehe Text);
3 Kennlinie einer zu leichten und zu wenig trennscharfen Prüfung, bei der 50 v. H. richtige Antworten tatsächlich Halbwissen repräsentieren mit höherer Ratewahrscheinlichkeit und kompletter Lösung aller Aufgaben auch bei ungenügendem Wissen.
Man beachte, dass bei trennscharfen Prüfungen schon geringe Schwierigkeitsunterschiede bei gleichem Wissen große Unterschiede im Prüfungserfolg bewirken, wenn dieser fälschlicherweise unmittelbar aus dem Anteil richtiger Antworten abgeleitet wird.

In der Praxis sind allerdings erhebliche Schwierigkeiten zu überwinden, bis eine solche absolute Erfolgsmessung korrekt möglich ist. Die modernen Erkenntnisse der Messtheorie haben ergeben, dass ein einwandfreier Zusammenhang zwischen den Rohwerten einer Prüfung und einer Maßzahl für das Erreichen des Lehrziels nur durch Verwendung des zweikategorialen Messmodells nach Rasch zu errechnen ist [2].

Für jede Aufgabe eine Item-Charakteristik

Dieses Messmodell nach Rasch stellt sehr strenge Anforderungen an die Prüfungsaufgaben. Sie müssen nicht nur kontentvalide, sondern auch homogen sein, dass heißt die gleiche Eigenschaft des Kandidaten (und nur diese) messen. Für jede Aufgabe muss eine so genannte Item-Charakteristik berechnet werden, aus der die Wahrscheinlichkeit ihrer richtigen Beantwortung in Abhängigkeit von der Leistung des Kandidaten hervorgeht - also gewissermaßen die Kennlinie der einzelnen Fragen, ähnlich wie sie das IMPP für drei Aufgaben bereits exemplarisch präsentiert hat [4].

Verwertbar sind nur Fragen, deren Kennlinien in der Steigung übereinstimmen und zudem einer bestimmten mathematischen Funktion gehorchen (so genannte Ogiven). Die Kennlinien dürfen lediglich entsprechend den Schwierigkeitsunterschieden parallel verschoben sein.

Nur unter diesen Voraussetzungen ist es möglich:

1. die erreichte Zahl richtiger Antworten als erschöpfende Information über die Leistung eines Kandidaten zu betrachten, nur dann steckt hinter Kandidaten mit gleicher Erfolgsquote (innerhalb des Messfehlers) auch gleiche Leistung;

2. aus der Matrix der Antworten aller Kandidaten den Schwierigkeits-Parameter der Fragen und den Fähigkeits-Parameter der Kandidaten separat zu schätzen;

3. zu messen, wie groß der Abstand der Kandidatenleistung vom gesteckten Lehrziel ist und ob dieser Abstand für das Bestehen der Prüfung ausreicht.

Das jetzige Bewertungsverfahren ist rechtswidrig

Die erforderlichen Matrizenrechnungen werfen allerdings auch im Zeitalter der Großrechenanlagen angesichts der Zahl von Prüfungsaufgaben und Kandidaten erhebliche numerische Probleme auf. Nach bisherigen Erfahrungen ist auch damit zu rechnen, dass ein großer Teil des Fragenpools (bis zu einem Drittel?) den strengen Anforderungen des Rasch-Modells nicht genügt [2, 3, 9]. Darüber hinaus dürfte die Forderung der inhaltlichen Validität und Homogenität noch weitere Schwierigkeiten bereiten.

Schließlich muss dann noch auf der absoluten Skala von der zuständigen Instanz die richtige Maßzahl gesetzt werden, denn die korrekte Abbildung der Kandidatenleistung auf einer Skala beantwortet noch nicht die Frage, wo denn die Marke "100 v. H. Mindestwissen" im Sinn der staatlichen Prüfung auf dieser Skala liegt. Hier wird man vielleicht eine Anschlusseichung mit einer korrekt durchgeführten relativen Leistungsmessung zu Hilfe nehmen können.

Absolute kriterienbezogene Leistungsmessung in der Ärztlichen Prüfung ist also heute noch Utopia. Das jetzige rechtswidrige Bewertungsverfahren verdient dieses Etikett nicht.

Die wissenschaftlichen und technischen Voraussetzungen für eine Lösung der Probleme in der Zukunft liegen aber vor, so dass nach einem Ausweichen auf eine korrekte relative normbezogene Leistungsmessung der offene Weg ohne Hast beschritten werden kann.

Literatur

1. BMJFG: Niederschrift über das Sachverständigengespräch am 16. Dezember 1969
2. Fischer, G.: Einführung in die Theorie psychologischer Tests. Verlag Hans Huber, Bern, Stuttgart, Wien, 1974
3. Fricke, R.: Über Meßmodelle in der Schulleistungsdiagnostik. Verlag Schwann. Düsseldorf, 1972
4. IMPP: Aufgaben Entwicklung Analysen. Verlag Schmidt & Bödige, Mainz, 1976
5. IMPP: Ergebnisbericht über die schriftlichen Prüfungen nach der Approbationsordnung für Ärzte, März 1977
6. Klauer, K. J.: Einführung in die Theorie lehrzielorientierter Tests. In: Klauer, K. J. et al. Lehrzielorientierte Tests, Schwann Verlag, Düsseldorf, 1975, S. 16
7. Koransky, W.: Mündliche Mitteilung
8. Kuni. H., Becker, P.: Multiple Choice als Numerus clausus (7) Bestehensregel kann nicht bestehen (bleiben) "der arzt im krankenhaus" (1980) 590-595
9. Neumann, D.: Mündliche Mitteilung
10. Schumacher, Ch. F.: Auswertung und Analyse der Prüfung. In: Hubbard, J. P.: Erfolgsmessung der medizinischen Ausbildung, Verlag Hans Huber, Bern, Stuttgart, Wien 1974, S. 91

Anschrift der Verfasser

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

Rechtsanwalt Dr. Peter Becker, Gisonenweg 9, 35037 Marburg