Multiple Choice als Numerus clausus (7)

Bestehensregel kann nicht bestehen (bleiben)

Horst Kuni und Peter Becker (Marburg)

Dies sind die Ergebnisse, zu denen die Autoren in den sechs bisher erschienenen Artikeln dieser Serie gelangt sind: Das System der Multiple-Choice-Prüfungen ist in eine "grundgesetzwidrige Handhabung abgeglitten", weil mit seiner Hilfe die Numerus-clausus-Politik "mit anderen Mitteln" fortgesetzt wird (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 ärztliches Basiswissen (Heft 9/80, Seite 522ff.).

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.

Mit einer kritischen Betrachtung der Bewertungsregel der schriftlichen Prüfungen schließt sich der Kreis unserer Auseinandersetzung mit diesem Prüfungssystem insofern, als die unsachgemäße Veränderung der Bewertungsregel und ihre deletären Folgen unmittelbarer Auslöser für unsere Betrachtungen waren [4].

Die Unterwerfung des Menschen unter von Dritten gesetzte Bewertungsregeln wird offensichtlich so stark von früher Kindheit an geübt, dass man sich schon nicht mehr wundern darf, wie unreflektiert die Bewertungsregeln der Approbationsordnung für Ärzte hingenommen und wie irrational sie bisher diskutiert wurden. So fand in der so genannten Kleinen Kommission, die gemeinsam mit dem Bundesminister für Familie, Jugend und Gesundheit (BMFJG) die Approbationsordnung vorberaten hat, eine Auseinandersetzung um die Bewertungsregeln nicht statt [8].

Ziel: relative Bewertung des einzelnen Kandidaten

Bei der Übernahme des amerikanischen Prüfungssystems hätte es zunächst nahe gelegen, auch die Bewertungsregeln mit zu übernehmen.

Es handelte sich dabei um eine normbezogene Leistungsmessung, also eine relative Bewertung des einzelnen Kandidaten im Vergleich zur Gesamtheit aller Kandidaten. Die Prüfungsergebnisse sollen

nicht nur einen Vergleich aller Kandidaten eines Prüfungstermins untereinander ermöglichen, sondern von Termin zu Termin auch vergleichbar sein, damit der Entscheidung "bestanden/nicht bestanden" eine konstante Bedeutung zukommt.

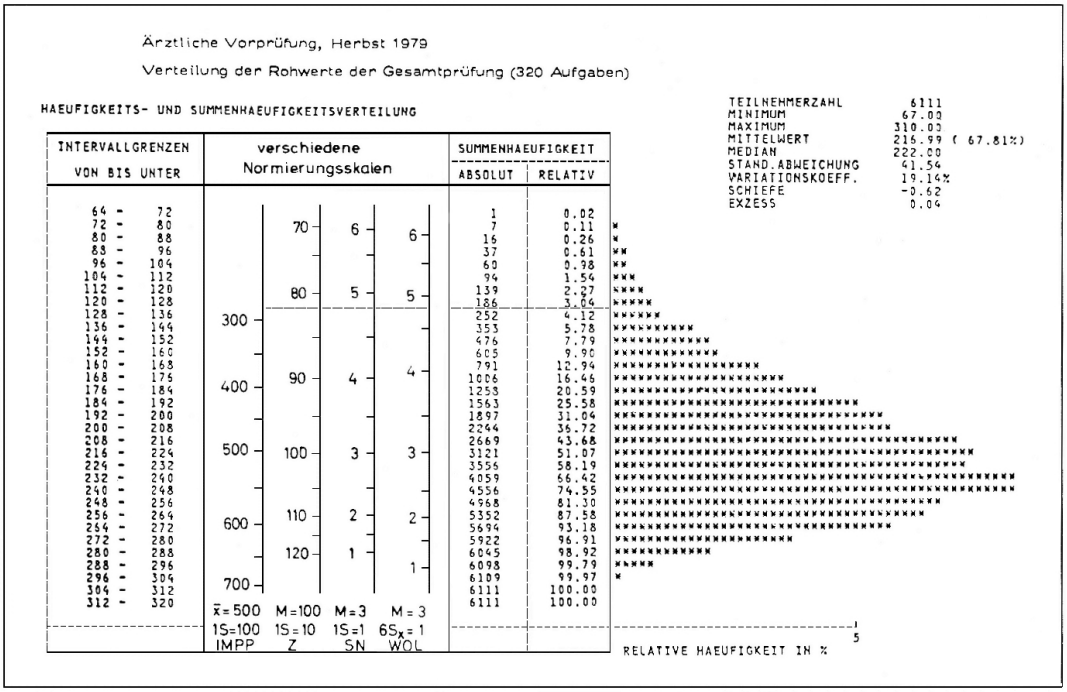

Abb. 1: **Beispiel für die Transformation von Rohwerten in Skalen für eine normbezogene Leistungsbeurteilung. Zugrunde liegt Tabelle 5 aus [1]**

- $\bar{x}$ Maßzahl für den Mittelwert bei linearer Transformation
- M Maßzahl für den Median bei Flächentransformation
- S dasselbe für einfache Standardabweichung der Rohwerte in der Kandidatenpopulation
- IMPP die in den Ergebnisberichten des IMPP (und des Medical Board) verwendete lineare Transformation
- Z in Deutschland übliche z-Skala, hier angelegt nach Flächentransformation, die die signifikante Abweichung der Häufigkeitsverteilung berücksichtigt [7]
- SN unter gleichen Gesichtspunkten angelegte Standard-Schulnotenskala.
- WOL Auf die Schulnotenskala übertragene Rechenregel nach Wolins [9], d.h. eine Notenklasse hat die Breite vom sechsfachen Messfehler
- S_x nach Wolins geschätzter Messfehler des individuellen Rohwertes.

Die Flächenanteile können unmittelbar der relativen Summenhäufigkeit entnommen werden. Dem Flächenanteil von 3 v. H. (beim Medical Board übliche) Misserfolgsquote entspricht auf der SN-Skala der Wert 4,8, das heißt 4,5+0,3s (Messfehler), und auf der z-Skala 82, das heißt 18 v. H. unter dem Mittelwert.
Man beachte, dass die Distanz 18 v. H. unter dem Mittelwert auf der normierten Skala einer höheren Misserfolgsquote entspricht, wenn sie ohne Berücksichtigung der Schiefe (und damit der Abweichung von der Normalverteilung) linear transformiert wurde.
Bei Anwendung auf die Rohwerteskala verliert die 18-Prozent-Grenze jeden rationalen Bezug und entspricht hier einer Misserfolgsquote von etwa 17 v. H. Aufgrund der zu diesem Zeitpunkt als Bewertungsschranke praktizierten 60 v. H.-Regel betrug die Misserfolgsquote tatsächlich 25,58 v. H., der Cut-off erfolge also etwa bei der Note 3,5!

Bei Anwendung der klassischen Testtheorie führt man deshalb eine Skalentransformation durch. Aus den Rohwerten werden Mittelwert und Standardabweichung errechnet. Die relative Position jedes

Kandidaten innerhalb der gesamten Population wird dann in Vielfachen der Standardabweichung vom Mittelwert festgelegt. Dabei verwenden Medical Board wie IMPP eine Skala, die eine (in der Tat nicht vorhandene) Normalverteilung oder eine (nicht vollzogene) Flächentransformation voraussetzt (Abb. 1, S. 2).

Man erkennt, dass die vom Medical Board praktizierte konstante Misserfolgsquote von 3 v. H. in den klinischen Prüfungen der Note 4,5 der Standard-Schulnotenskala zuzüglich einer Standardabweichung des Messfehlers (Reliabilität, s. [5]) (etwa ein Viertel Sigma) entspricht. Diese Schranke liegt auf der in Deutschland üblichen z-Skala 18 v. H. unter dem Mittelwert [7].

Ein verhängnisvolles Missverständnis

Damit wird auch die Herkunft der Bestehensgrenze klar, die in der ursprünglichen Fassung der Approbationsordnung enthalten war. In der Sachverständigenanhörung am 16. Dezember 1969 im Bundesministerium für Jugend, Familie und Gesundheit hatte nämlich Ministerialrat von Arnim als Alternative zur amerikanischen, nicht in Deutschland anwendbaren konstanten Misserfolgsquote vorgeschlagen:" … sich nach einer Durchschnittsleistung zu richten und dabei einen gewissen, unter dem Durchschnitt liegenden Prozentsatz festzulegen, der noch zum Bestehen der Prüfung ausreicht."

Auf den berechtigten Vorbehalt des Vertreters des Hochschul-Informationssystems (HIS), dass darin praktisch kein Unterschied bestehe, meint von Arnim, "… dass die Festlegung von Standardabweichungen ein größeres Maß an Chancengleichheit biete".

Betrachten wir nun die endgültige Fassung in der Approbationsordnung vom 28. Oktober 1970:

> "Paragraph 14, Absatz 5:"Die schriftliche Prüfung ist bestanden, wenn der Anteil der von dem Prüfling richtig beantworteten Fragen nicht mehr als 18 vom Hundert unter der durchschnittlichen Prüfungsleistung der Prüflinge des jeweiligen Prüfungstermins im gesamten Bundesgebiet liegt oder wenn der Prüfling mindestens 50 vom Hundert der Fragen zutreffend beantwortet hat."
>
> Die amtliche Begründung dazu lautet: "Absatz 5 befasst sich mit der Bewertung, der schriftlichen Prüfungen. Durch die Wahl eines relativen Bewertungssystems wird der Tatsache Rechnung getragen, dass mit den schriftlichen Prüfungen nach dem "Antwort-Wahl-Verfahren" etwas Neues in die staatlichen Prüfungen im Medizinstudium eingeführt wird.
>
> Es bestehen keine ausreichenden Erfahrungen, die der Festsetzung von absoluten Bewertungszahlen hatten zugrundegelegt werden können. Auch ist anzunehmen, dass die neue Prüfungsart den Prüflingen zunächst Schwierigkeiten machen wird, so dass eine Lösung angezeigt ist, die eine allzu strenge Auslese verhindert.
>
> Das BMJFG hat sich zu den mit den schriftlichen Prüfungen zusammenhängenden Fragen sehr eingehend von zahlreichen Sachverständigen des In- und Auslandes beraten lassen. Dabei hat sich gezeigt, dass unter den in der Bundesrepublik gegebenen Verhältnissen einem relativen Bewertungsmaßstab der Vorzug zu geben ist.

> Um zu verhindern, dass im Falle einer außergewöhnlich guten Durchschnittsleistung Prüflinge die Prüfung nicht bestehen, die die Hälfte oder mehr der Fragen richtig beantwortet haben, ist eine Alternativlösung vorgesehen. Insoweit ist eine absolute Bewertungszahl vertretbar."

Hier zeigt sich also ein verhängnisvolles Missverständnis über den Begriff der Standardabweichung. Denn in der schließlich in Kraft getretenen Fassung bezog sich die 18-Prozent-Schranke nicht mehr auf normalisierte Skalenwerte, sondern auf Rohwerte! Hierdurch ist zwar tatsächlich eine konstante Misserfolgsquote vermieden worden, der Wert "18 Prozent" hat aber zugleich jegliche rationale Grundlage verloren.

Nachteile der absoluten Leistungsmessung

Einmal mehr zeigt sich hier die mangelhafte Berücksichtigung der testtheoretischen Grundlagen, in deren Systemzusammenhang die Prüfungen angelegt waren: Je mehr Kandidaten durch Erreichen des Lehrziels einen hohen Durchschnittswert der Rohwerte erzielt hätten, um so mehr wäre die Varianz der Rohwerte zurückgegangen. Bei einer sachgerecht festgesetzten Schranke (zum Beispiel dem 1,8fachen des derzeitigen Variationskoeffizienten von etwa 20 v. H. = 38 v. H.) wäre bei einem Anstieg des Rohwertedurchschnitts auf etwa 85 v. H. kaum noch ein Kandidat durchgefallen.

Dieser Sachverhalt ist wegen der früher geschilderten "Nachjustierung" des Schwierigkeitsgrades [6] allerdings genauso hypothetisch wie das gegen die 18-Prozent-Schranke vorgebrachte Argument, eine bundesweite Absprache aller (!) Studenten eines Prüfungstermins könne zu einem ganz niedrigen Durchschnittswert der Rohwerte mit fehlender Varianz und damit ebenfalls zum Bestehen aller ohne Leistungsnachweis führen [2, 3].

Das gigantische Missverständnis bei der Verpflanzung der relativen Bewertungsregel vom Amerikanischen ins Deutsche hat völlig den Blick darauf verstellt, dass die normbezogene Leistungsmessung bei richtiger Anwendung in pragmatischer Weise eine Fülle von Schwierigkeiten aus dem Wege räumt, die - wie wir noch zeigen werden - sich bei der kriterienbezogenen, absoluten Leistungsmessung auftürmen und vom IMPP so rasch nicht zu bewältigen sind.

Es liegt auf der Hand, zunächst einmal den Anschluss an andere Prüfungsbewertungsregeln dadurch herzustellen, dass man allen Kandidaten die Prüfung als bestanden bescheinigt, die nach korrekt durchgeführter Skalentransformation auf der Standard-Schulnotenskala unter Berücksichtigung des Messfehlers noch die Note 4 ("ausreichend") erzielen.

Nun ist die Skala allerdings in zwei Parametern elastisch: Der Mittelwert und die Skaleneinheit (Standardabweichung) sind in gleicher Weise populationsbezogen. Die resultierende konstante Misserfolgsquote (jeweils die "unteren" drei Prozent der Kandidatenpopulation sind von ihr betroffen) könnte theoretisch auch einem Kandidaten mit absolut noch "ausreichenden" Kenntnissen den Examenserfolg versagen.

Wolins-Skala: Strenger - aber gerechter

Hier führt ein Vorschlag von Wolins weiter, der auch als Gast bei der oben genannten Sachverständigenanhörung am 16. Dezember 1969 in Bonn seine Stellungnahmen abgab (Wolins hatte damals eine Gastprofessur in Marburg): Nur der Mittelwert soll unmittelbar auf die Kandidatenpopulation bezogen werden, als Skaleneinheit soll die Standardabweichung des Messfehlers (Reliabilität) verwendet werden [9]. Sein Votum: Ein Kandidat soll der nächsten Notenklasse zugeordnet werden, wenn er hoch signifikant (= dreifacher Messfehler = 99 v. H. Vertrauensbereich) in der Leistung differiert.

Auf deutsche Gepflogenheiten der Notengebung angewendet, bedeutet das: dreifach Messfehler unter dem Mittelwert beginnt der Notenbereich "Ausreichend" (entspricht also 3,5), dreifach Messfehler über dem Mittelwert beginnt der Notenbereich "gut" (entspricht also 2,49). Ein deutscher Notenbereich entspricht also der Spanne des sechsfachen Messfehlers.

Wolins schätzt den Messfehler δ in guter Näherung mit der Formel

$$\delta_x = \sqrt{x(1-\frac{x}{n})}$$

wobei x der individuelle Rohwert des Kandidaten in einem Test mit n Fragen ist.

Wendet man die Wolins-Regel empirisch auf die Ergebnisse der schriftlichen Prüfungen an (Abb. 1, S. 2), ergeben sich nur geringe Verschiebungen gegenüber der Standard-Schulnotenskala. Die Wolins-Skala ist allerdings "strenger", da nun nicht zusätzlich der Messfehler geltend gemacht werden kann, aber insofern "gerechter", da bei ihrer Anwendung theoretisch alle Kandidaten bestehen können, falls sie "gut" genug sind.

So hätten beispielsweise im Dritten Abschnitt der Ärztlichen Prüfung August 1977, als die kleine Spitzengruppe von 811 Kandidaten antrat, bei Anwendung der Standard-Schulnotenskala die vorprogrammierten 3 v. H. der Kandidaten nicht bestanden, mit der Wolins-Skala wären aber alle Kandidaten erfolgreich gewesen (wie sie es in der Tat auch waren).

Die richtige Wahl einer relativen Bestehensschranke müsste deshalb wohl so erfolgen, dass ein Kandidat nach wenigstens einer der beiden Regeln mindestens die Note "ausreichend" erzielen muss, um das Examen zu bestehen.

Da die Grundlagen für ein korrektes Vorgehen also bereits vor Erlass der Approbationsordnung vom 28. Oktober 1970 dem BMFJG zugängig waren, ist es eine geradezu tragische Entwicklung, dass durch eine unsinnige Anwendung der relativen Bestehensregel eine als Auffangposition gedachte absolute Bewertungsschranke zum Zuge kam, die unsere weitere Analyse im folgenden Beitrag als nicht minder sinnlos und daher rechtswidrig entlarven wird.

Literatur

1. IMPP: Vorbericht. Ergebnisse der schriftlichen Prüfungen nach der Approbationsordnung für Ärzte, Herbst 1979

2. Krämer, H. J.: Mißerfolgsquote stark erhöht – wo liegen die Ursachen? Dtsch. Ärztebl. (1980) 544

3. Krämer, H. J.: Jeder fünfte fiel durch – warum? Klinikarzt 9 (1980) 139

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

5. Kuni. H., Becker, P.: Multiple Choice als Numerus clausus (2) Zu viele Medizinstudenten scheitern am Messfehler "der arzt im krankenhaus" (1980) 292-295

6. Kuni. H., Becker, P.: Multiple Choice als Numerus clausus (4) Prüfung nicht bestanden: die Kandidaten oder die Fragen? "der arzt im krankenhaus" (1980) 406-409

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

8. Sewering, H. J.: Mündliche Mitteilung

9. Wolins, L.: Methods for Basing Grades on Objective Test Scores. Manuskript, Juni 1970 (Auf dieses Manuskript machte uns dankenswerter Weise H. Kapuste aufmerksam.)

Anschrift der Verfasser

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

Rechtsanwalt Dr. Peter Becker, Gisonenweg 9, 35037 Marburg