Zuverlässig, objektiv und weitgehend gültig

5. schriftliche Facharztprüfung Allgemeinmedizin 2001 mit Kurzantwortfragen

P. Schläppi, S. Feller, B. Rindlisbacher, R. Bloch

Wenn medizinische Bildungsgänge und damit auch Facharztprüfungen einem internationalen Qualitätsvergleich standhalten wollen, müssen sie anerkannten wissenschaftlichen Ansprüchen genügen. Aber auch von der Öffentlichkeit wird zunehmend Qualität bzw. Qualitätssicherung in den Medizinalberufen gefordert (vorliegender Entwurf eines neuen Bundesgesetzes).

Die Schweizerische Gesellschaft für Allgemeinmedizin SGAM hat beschlossen, die schriftliche Facharztprüfung mit Kurzantwortfragen durchzuführen. Sie will damit v.a. das aktiv formulierte, praxisrelevante Wissen und die Problemlösungsfähigkeit von Titelanwärtern/-innen beurteilen. Die Methode mit Kurzantwortfragen (KAF), die solche Prüfungszielsetzungen erfüllen kann, ist in der Schweiz im Gegensatz zur Multiple-Choice-Methode (MC) und zu anderen Ländern auf Niveau Schluss- und Facharztprüfungen wenig verbreitet. Das hängt nicht zuletzt damit zusammen, dass die wissenschaftlichen Anforderungen an eine Prüfung mit KAF schwieriger zu erfüllen sind als mit MC-Fragen. Anforderungen wie Validität, Reliabilität, Objektivität und Durchführbarkeit müssen aber erfüllt werden, damit eine Prüfung mit Berufsausübungskonsequenzen ihren gesetzten Zielen fair und gerecht nachkommt.

Das Beispiel der schriftlichen Facharztprüfung Allgemeinmedizin der SGAM zeigt, dass die genannten Anforderungen auch für eine Prüfung mit Kurzantwortfragen in der Schweiz erreicht werden können. In einem langjährigen Prozess ist ein über die Sprachgrenzen hinweg gut funktionierendes Team aus Prüfungskommission, Examinatoren/-innen, Fragenautoren/-innen, Sekretärinnen und Prüfungsbearbeitern der AAE entstanden, das in diesem Jahr mit dem fünften Durchgang das Ziel praktisch erreicht hat, eine zuverlässige (reliable), objektive und weitgehend gültige (valide) schriftliche Facharztprüfung durchzuführen.

Korrespondenz:
Dr. med. Peter Schläppi
Medizinische Fakultät der Universität Bern
Institut für Aus-, Weiter- und Fortbildung IAWF
Abteilung für Ausbildungsforschung AAE
Inselspital 37a
CH-3010 Bern
E-mail: peter.schlaeppi@iae.unibe.ch

Qualität von Medizinalprüfungen

In der ärztlichen Aus- und Weiterbildung werden auch in der Schweiz in den letzten Jahren vermehrt Qualitätsfragen gestellt, nicht zuletzt wegen der Bemühungen um internationale Anerkennung der Aus- und Weiterbildung oder wegen der Kosteneffizienz der Bildung:

- Was wissen oder können Studienabgänger/innen bzw. was sollen oder müssen sie wissen oder können? Schweizerische Ausbildungsziele werden momentan erstmals für das gesamte Studium erarbeitet [1].
- Welche Kompetenzen müssen Ärzte/Ärztinnen am Schluss ihrer Weiterbildung haben? Die Weiterbildungsprogramme der Fachgesellschaften wurden erst in letzter Zeit ausführlicher niedergeschrieben, sie werden laufend revidiert [2].
- Wie soll die Qualität der Ausbildungsgänge belegt werden? Eine erste Akkreditierung der Medizinischen Fakultäten wurde im letzen Jahr durchgeführt [3], die gesamte medizinische Aus-, Weiter- und Fortbildung und ihre Qualitätssicherung wird gesetzlich neu geregelt mit einem Medizinalberufegesetz, das demnächst im Parlament behandelt wird [4].

Eine Möglichkeit, die Qualität der medizinischen Bildung zu überprüfen bzw. zu sichern, ist die Kontrolle, ob die Studienabgänger/innen bzw. die Fachärzte/-ärztinnen die gesteckten Bildungsziele erreichen. Gute Prüfungen sind deshalb sowohl gegenüber der Öffentlichkeit, den Bildungsverantwortlichen als auch gegenüber den Kandidaten/-innen wichtig.

Was sind gute Prüfungen? Prüfungen mit Konsequenzen für die Berufsausübung müssen aus wissenschaftlicher Sicht vor allem vier Anforderungen [5] genügen, die in Tabelle 1 aufgeführt sind.

Damit eine Prüfung diese Anforderungen erfüllen kann, muss sie einer gewissen Struktur folgen. Die wichtigsten Elemente dieser Struktur sind in Tabelle 2 zusammengefasst.

Im Zuge der Verbesserung von Medizinalprüfungen wurden in den Siebzigerjahren auch in der Schweiz die Multiple-Choice-Prüfungen eingeführt, die dieser Struktur folgen und – als Prüfung passiv wiedererkannten Wissens – die erwähnten Messanforderungen erfüllen können [6]. In Situationen mit komplexeren Prüfungsaufgaben ist diese Methode schwieriger anzuwenden, für die Beurteilung von Fertigkeiten, Fähigkeiten oder Verhalten der Prüfungskandidaten/-innen ist sie nicht geeignet. Andere Prüfungsformen, wie sie hierzulande üblich sind (z.B. Essayfragen oder das mündlich-praktische Examen), haben es schwer, die erwähnten Anforderungen zu erfüllen [7, 8]. In der Schweiz wurden diese Prüfungsarten bisher noch kaum auf ihre Qualität hin überprüft, weder bei den Schlussprüfungen noch bei den Facharztprüfungen.

Die Schweizerische Gesellschaft für Allgemeinmedizin SGAM will mit ihrer schriftlichen Facharztprüfung möglichst Wissen prüfen, das für die kom-

Tabelle 1
4 wichtige Anforderungen an einen Test.

Validität	Bei der *Gültigkeit* eines Testes geht es um die Genauigkeit, mit der ein bestimmtes Merkmal (z.B. die Kompetenzen von Hausärzten/-ärztinnen) tatsächlich gemessen wird, das mit der Prüfung gemessen werden soll.
Reliabilität	Hier wird die *Zuverlässigkeit* der Prüfung untersucht, mit der ein bestimmtes Merkmal gemessen wird. Mit welcher Wahrscheinlichkeit würde sich – bei einer Wiederholung der Facharztprüfung mit gleichwertigen Fragen – die gleiche Leistungsbeurteilung der Kandidaten/-innen ergeben?
Objektivität	Wenn die Beurteilung der Leistungen der Kandidaten/-innen unabhängig von den Prüfenden erfolgt, ist eine Prüfung objektiv.
Durchführbarkeit	Die Berufsorganisation kann den Aufwand für eine qualitativ optimale Prüfung leisten.

Tabelle 2
Strukturelemente einer guten Prüfung.

- Definition der zu prüfenden Kompetenzen
- gewichtetes Inhaltsverzeichnis (Blueprint) für die gewählte Prüfungsform
- Entwicklung von relevanten Prüfungsaufgaben mit definierten Kriterien (Fragen und Antworten)
- Standardsetzung (Bestimmen der Bestehensgrenze)
- Training der Examinatoren/-innen in der Prüfungsmethode
- Prüfungsveranstaltung
- Bewerten der Leistung der Kandidaten/-innen
- Analyse der Prüfung und der Prüfenden
- Nachlese, Feedback an alle Beteiligte

plexe alltägliche Praxisarbeit verfügbar sein muss und das die Prüflinge deshalb aktiv formulieren können müssen. Zum Beispiel:
- Wie ist ein Befund bei einem Patienten zu interpretieren und in dessen individuellen Behandlungsplan zu integrieren?
- Wie lösen angehende Hausärzte/-ärztinnen ein konkretes Problem, das sie in ihrer Praxis antreffen werden?

Die SGAM wählte für diese Prüfungszielsetzung die Methode mit Kurzantwortfragen, die z.B. in Kanada [9] oder Grossbritannien [10] in dieser Situation bereits etabliert ist.

Wie steht es nun aber mit der Qualität dieser neu eingeführten Facharztprüfung? Um diese Frage zu beantworten, werden die fünf Prüfungen (1997–2001) im folgenden auf dem Hintergrund der skizzierten Struktur (vgl. Tab. 2) und der erwähnten Prüfungsanforderungen (vgl. Tab. 1) analysiert.

Valide Beurteilung

Für die Validität einer Prüfung sind ein gewichtetes Prüfungsinhaltsverzeichnis (Blueprint) und Prüfungsaufgaben, die aus definierten Berufskompetenzen abgeleitet werden, entscheidend.

Im schriftlichen Facharztexamen Allgemeinmedizin werden Schlüsselkompetenzen aus konkreten «Praxisfällen» nach dem sogenannten «key-feature-approach» [11] in einem mehrschrittigen Konsensprozess entwickelt und in Prüfungsaufgaben umgesetzt. Diese Arbeit macht ein nach fünf Jahren gut harmonierendes, dreisprachiges Team aus etwa 40 Personen (Examinatoren/-innen, Fragenschreibern/-innen und Prüfungsbearbeitern der AAE). Dieses Team garantiert in diesem Prozess die Praxisrelevanz, die inhaltliche und formale Richtigkeit der Prüfungsaufgaben wie auch die faire Bewertung der Leistungen der Kandidaten/-innen in diesen Aufgaben. Die Prüfungskommission der SGAM und die Examinatoren/-innen entwickelten einen Blueprint als systematische Referenz für den jeweiligen Prüfungsinhalt.

Tabelle 3 beschreibt diesen Blueprint, wie er seit 1999 verwendet wird. Sie zeigt zudem die effektive Verteilung der jeweiligen Prüfungsaufgaben 1999–2001 auf die vier Dimensionen Allgemeinmedizin, Fächerwissen, Alter und Konsultationsart. Aus der Tabelle geht hervor, dass die Repräsentativität der Prüfungsaufgaben – gemessen an diesem Blueprint – noch nicht optimal ist. Ein wesentlicher Grund dafür ist der noch kleine Fall- bzw. Fragenpool, welcher nicht alle Gebiete des Blueprints genügend abdeckt. Wie Tabelle 4 aufzeigt, konnte den Prüflingen erst in den letzten zwei Jahren eine adäquate Fall- und Fragenanzahl vorgelegt werden. Seit diesen zwei Jahren ist die Prüfung im übrigen sanktionierend, d.h. Kandidaten/-innen, welche die Prüfung nicht bestehen, müssen sie wiederholen (frühestens ein Jahr später möglich).

Reliable Beurteilung

Die strukturierte Entwicklung und Anwendung von Prüfungsaufgaben minimiert viele Faktoren, welche die reliable Messung der Leistung der Kandidaten/-innen stören könnten. Beispiele solcher Störfaktoren sind unterschiedlich beurteilende Examinatoren/-innen, Kandidaten/-innen mit unterschiedlicher Vorbildung oder «Lehrmeinungen», die von Weiterbildungsstätte zu Weiterbildungsstätte differieren.

Die Prüfung scheint zunehmend nur noch *einen* Faktor zu messen, wird sie doch homogener, wie aus der Analyse nach Cronbach hervorgeht (Abb. 1). Das ist einerseits wünschenswert und die Annahme berechtigt, dass als wichtigster Faktor das allgemeinmedizinische Wissen der Kandidaten/-innen geprüft wird. Die Homogenisierung kommt dabei in erster Linie durch folgende Strukturelemente zustande:
- Erhöhung der Fall- und Fragenzahl (vgl. Tab. 4);
- Verlängerung der Prüfungszeit von zwei auf drei Stunden;
- Verbesserung des Konsens- und Korrekturprozesses.

Die Homogenisierung kann andererseits auch bedeuten, dass die Kurzantwortfragen sich immer stärker angleichen. Das wäre natürlich nicht erwünscht. Die

Tabelle 3
Blueprint und effektive Verteilung der Prüfungsaufgaben (Angaben in Prozent).

A. Dimension Allgemeinmedizin	soll	1999	2000	2001
1. allgemeinmedizinisches Fachwissen	40	41	31	36
2. Konsultation	35	59	68	64
2.1 Anamnese	10	9	16	8
2.2 Körperliche Untersuchungen	8	5	9	15
2.3 Zusatzuntersuchungen	6	11	8	14
2.4 Therapie, Management, Entscheidungen	8	34	35	26
2.5 Administration	3	0	0	1
3. Interaktion/Kommunikation/Arzt-Patient-Beziehung	15	0	0	0
4. Ethik	5	0	0	0
5. Recht	5	0	1	0
Total	100	100	100	100
B. Dimension Fächerwissen Allgemeinmedizin				
1. Chirurgie	10	7	8	16
2. Innere Medizin	30	66	38	41
3. Arbeitsmedizin	2	0	0	0
4. Dermatologie	5	0	10	12
5. Gynäkologie/Geburtshilfe	6	5	1	0
6. Neurologie	7	2	8	13
7. Ophthalmologie	4	0	5	3
8. Otorhinolaryngologie	7	9	4	0
9. Pädiatrie	8	7	10	8
10. Psychiatrie, Psychosoziale Medizin	9	4	7	5
11. Radiologie (Radiodiagnostik)	5	0	0	0
12. Rheumatologie	7	0	9	2
Total	100	100	100	100
C. Dimension Alter				
1. Säuglinge (<1)	5	0	0	8
2. Kinder (1–11)	10	9	13	4
3. Adoleszente (12–18)	10	0	0	6
4. Erwachsene (19–64)	45	66	67	78
5. ältere Menschen (>64)	30	25	20	4
Total	100	100	100	100
D. Dimension Konsultationsart				
1. Sprechstunde	70	66	74	73
2. Hausbesuch	10	9	3	5
3. Notfall	5	25	15	17
4. Telefon	10	0	8	5
5. Indirekte Konsultation (Gutachten)	5	0	0	0
Total	100	100	100	100

Abbildung 1
Entwicklung der Reliabilitätsmessung nach Cronbach in den fünf bisherigen Facharztprüfungen. Für eine gute Prüfung wird in der Literatur ein Cronbach-alpha von mindestens 0,8 gefordert.

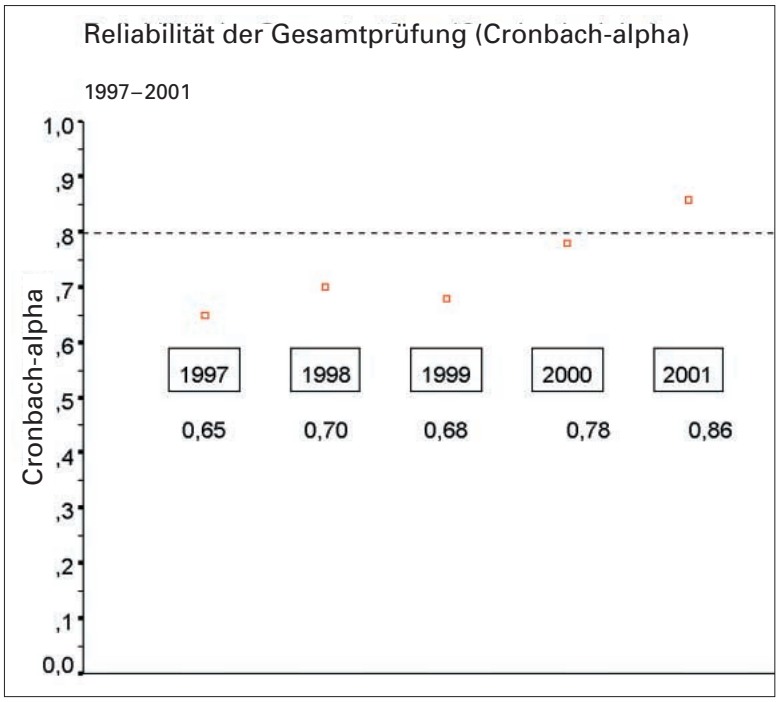

Tabelle 4
Anzahl Fälle, Fragen und Kandidaten/-innen der Facharztprüfungen.

	1997	1998	1999	2000	2001
Fälle	10	12	12	23	26
Fragen	28	43	44	77	78
Kandidaten/-innen	88	116	121	135	137

Praxisfälle sollten möglichst als komplexe Prüfungsaufgaben erhalten bleiben. Komplexität würde aber eher inhomogene Fragen erwarten lassen.

Die Reliabilitätsmessung nach Cronbach ist *eine* Möglichkeit, die aktuelle Reliabilität einer Prüfung zu beurteilen. In einer sanktionierenden Facharztprüfung sind allerdings andere Analysen wie Paralleltestmethoden oder eine Testwiederholung praktisch nicht durchführbar.

Hingegen kann die Übereinstimmung der Korrektur überprüft werden (Interraterreliabilität). Hier sieht es in der vorliegenden Prüfung gut aus: Die richtigen Antworten sind für die korrigierenden Examinatoren/-innen im Antwortschlüssel festgehalten, der von über 20 Examinatoren/-innen vor der Prüfung im Konsens akzeptiert wurde. In diesem Jahr wurden die Leistungen von 5 zufällig ausgewählten Kandidaten/-innen der Prüfung 2001 von 2 unabhängigen Examinatoren korrigiert. Die Übereinstimmung der Urteile ist hoch, wie aus Tabelle 5 hervorgeht (Intraclass Correlation Coefficient alpha = 0,975).

Objektive Beurteilung

Wird eine Prüfung entlang der beschriebenen Strukturelemente (vgl. Tab. 2) durchgeführt, ist sie unabhängig von einzelnen Prüfenden und damit objektiv.

Die Beurteilung der Leistung der Kandidaten/-innen in der Facharztprüfung Allgemeinmedizin geschieht dank dem erwähnten Konsens- und Korrekturprozess in einer Art und Weise (s. o.), dass sie nicht mehr abhängig ist von einzelnen Examinatoren/-innen. Zudem wird die Bestehensgrenze in einem inhaltsorientierten Standardsetzungsverfahren *vor* der Prüfung festgelegt, basierend auf Bestehensentscheiden einer Auswahl von Examinatoren/-innen, die bei jeder Frage entsprechend den festgelegten richtigen Antworten urteilen [12].

Durchführbarkeit der Prüfung

Dank dem unermüdlichen Einsatz der Prüfungskommission der SGAM (Präsident Prof. A. von Graffenried) hat die Prüfung in der nun vorliegenden Qualität entwickelt werden können. Gründliche Vorarbeit, gute Logistik (Sekretariat der SGAM: Luzia Schneider, Marlies Kara) und der Wille zu kontinuierlichen Verbesserungen waren dabei die Eckpfeiler. Nach zwei Jahren z. B. drohte das Unterfangen fast zu scheitern, weil zuwenig neue Fälle und Fragen geschrieben wurden. Ein Effort der SGAM mit Verbreiterung der Examinatoren/-innenschaft und Workshops zum Fragenschreiben umschiffte diese Klippe. Zeitweise wurden die personellen Ressourcen in der Romandie knapp. Insgesamt aber ist der Aufwand für diese Prüfung leistbar, wenn auch oft nur dank viel Goodwill der Mitarbeitenden.

Fazit

Nach drei Probe- und zwei sanktionierenden schriftlichen Facharztprüfungen Allgemeinmedizin mit Kurzantwortfragen sind die Validität, Reliabilität und Objektivität weit gediehen, wie die vorliegenden Analysen zeigen. Wichtigstes Element dabei ist das Examinatoren/-innenteam, welches strukturiert und praxisorientiert die Prüfung jedes Jahr entwickelt, durchführt und die Leistungen der Kandidaten/-innen in einem inhaltlich orientierten Standardsetzungsverfahren bewertet.

Tabelle 5
Interraterreliabilität in einer Stichprobe von 5 Kandidaten/-innen der Prüfung 2001.

Die von den 2 Examinatoren jeweils verteilten Punkte pro Antwort sind als Urteile (total 390) gegeneinander aufgelistet. Auf der Diagonalen liegen also die Anzahl übereinstimmender Urteile, die abweichenden Urteile sind in den darum herumliegenden Zellen ersichtlich.

Beispiel: Die Examinatoren waren sich 12mal einig, die Leistung eines/einer Kandidaten/-in mit 0 Punkten zu bewerten, je einmal gab Examinator A allerdings 1 bzw. 2 Punkte, 6mal gab Examinator B 1 Punkt.

Examinator A	0	1	2	3	4	5	6	verteilte Punkte
Examinator B								
0	12	1	1					
1	6	89	8	3				
2		5	108	8				
3			5	79	7			
4				4	24	3		
5					1	11	2	
6						2	11	
verteilte Punkte								390
	Anzahl Urteile							
Intraclass Correlation Coefficient alpha = 0,975								

Literatur

1 Arbeitsgruppe «Definition der Lernziele des Medizinstudiums» der Schweizerischen Medizinischen Interfakultätskommission SMIFK (Präsident: Prof. H. Bürgi).
2 vgl. www.fmh.ch/WBO oder www.fmh.ch/Weiterbildungsprogramme.
3 vgl. z.B. www.unibe.ch/faculties/akkredit/.
4 vgl. Gesetzesvorlagen «Fleiner I/II/III», www.admin.ch/bag/berufe/weiterbi/d/index.htm.
5 Lienert GA, Raatz U. Testaufbau und Testanalyse. Weinheim: Psychologie Verlags Union; 1994.
6 Norcini JJ, Swanson DB, Grosso LJ, Webster GD. Reliability, validity and efficiency of multiple choice questions and patient management problem item formats in assessment of clinical competence. Med Educ 1985;19:238-47.
7 Muzzin LJ, Hart L. Oral examinations. In: Neufeld VR, Norman GR (eds.). Assessing clinical competence. New York: Springer; 1985. p. 71-93.
8 Weingarten MA, Polliack MR, Tabenkin H, Kahan E. Variation among examiners in family medicine residency board oral examinations. Med Educ 2000;34:13-7.
9 Handfield-Jones R, Brown JB, Biehn J, Rainsberry P, Brailovsky CA. Certification examination of the Royal College of Family Physicians of Canada. Part 3: Short-answer management problems. Can Fam Physician 1996;42:1353-61.
10 Knox JD. What is ... a modified essay question? Medical Teacher 1989;11:51-7.
11 Page G, Bordage G. Developing key-feature problems and examinations to assess clinical decision-making skills. Acad Med 1995;70:194-201.
12 adaptiert nach Angoff WA. Scales, norms and, equivalent scores. In: Thorndike RL (ed.). Educational Measurement. Washington DC: American Council on Education; 1971. p. 514-5.