GMS Zeitschrift für Medizinische Ausbildung ISSN 1860-3572

Grundlegende quantitative Analysen medizinischer Prüfungen

Basic quantitative analyses of medical examinations

• Andreas Möltner[1] • Dieter Schellberg[2] • Jana Jünger[1]

Zusammenfassung:
Es werden die Auswertungsschritte beschrieben, die für eine einfache testtheoretische Analyse einer Prüfung notwendig sowie als Grundlage von Aufgabenrevisionen, Verbesserungen von Prüfungszusammenstellungen und Rückmeldung an Lehrbeauftragte und Curriculumsentwickler ausreichend sind. Diese Schritte umfassen die Ergebnisauswertung, die Analyse der Aufgabenschwierigkeiten und der Trennschärfen, sowie - wo angebracht - die entsprechenden Auswertungen der Einzelantworten. Vervollständigt wird das Vorgehen durch die Bestimmung der internen Konsistenz, durch die die Zuverlässigkeit und Aussagekraft (Reliabilität) der Prüfung abgeschätzt wird.
Schlüsselwörter:Messmethoden in der Ausbildung, Reproduzierbarkeit von Ergebnissen

Abstract:
The evaluation steps are described which are necessary for an elementary test-theoretic analysis of an exam and sufficient as a basis of item-revisions, improvements of the composition of tests and feedback to teaching coordinators and curriculum developers. These steps include the evaluation of the results, the analysis of item difficulty and discrimination and - where appropriate - the corresponding evaluation of single answers. To complete the procedure, the internal consistency is determined, which makes an estimate of the reliability and significance of the examination.
Keywords:educational measurement, reproducibility of results

Einleitung

• Wozu dient die quantitative Analyse einer Prüfung?

Nach der neuen ÄAppO werden im Abschlusszeugnis einer Universität mehr als 30 Noten aufgeführt [18]. Doch was sagen diese Noten überhaupt aus? Wer als Klinikleiter auf ein Zeugnis blickt, fragt sich, ob jemand mit einer Eins in einem bestimmten Fach wirklich mehr weiß oder kann als jemand mit einer Drei. Studierende mit einer guten Note wollen mit dieser auch reüssieren, die mit einer schlechten fragen sich, ob sie nicht - aus welchen Gründen auch immer - ungerecht behandelt worden sind. Von den letzteren werden möglicherweise einige versuchen, Prüfungen juristisch anzufechten, und - nach der Einführung von Studiengebühren - eventuell sogar auf Schadensersatz dringen. Ob es dann ausreichen wird, auf die **Bemühungen** zur Erstellung inhaltlich angemessener und aussagekräftiger Prüfungen hinzuweisen, ist fraglich, letztendlich ist man erst dann auf der sicheren Seite, wenn für die Einhaltung eines Mindeststandards der Prüfungsqualität auch ein **Nachweis** erbracht werden kann. Solch ein Nachweis bedarf jedoch einer quantitativen Analyse der Prüfungsresultate.

Neben solchen externen gibt es aber auch genügend andere Gründe für die quantitative statistische Analyse von Prüfungen: So ist etwa der Erfolg einer guten Prüfungsvorbereitung nie garantiert, es wird immer wieder vorkommen, dass einzelne Fragen oder Aufgaben sich erst während oder nach der Prüfung als zu schlecht konzipiert erweisen. Entfällt eine Kontrolle, kann man Schwachstellen einer Prüfung nicht identifizieren und damit auch nicht aus seinen Fehlern lernen. Letztendlich würde dies die Aufgabe des Anspruchs bedeuten, zuverlässig und gut zu prüfen.

Individuell wird mit einer Prüfung die Leistungsfähigkeit eines Studierenden beurteilt, bezogen auf die Gruppe aller Prüfungskandidaten prüft sie den Lehrerfolg. Kann die Mehrzahl der Studierenden eine Aufgabe nicht bewältigen, so wurden die zur Lösung notwendigen Kenntnisse oder Fertigkeiten offensichtlich nicht hinreichend vermittelt, dies sollte zu Konsequenzen hinsichtlich der Lehre führen. Für eine gute Rückmeldung von Prüfungsergebnissen an die Lehrenden ist jedoch eine differenzierte und klare Ergebnispräsentation notwendig, für die eine fundierte Prüfungsauswertung benötigt wird.

Dies sind alles gute Gründe für eine quantitative Analyse von Prüfungen. Wie geht man aber bei der Auswertung **praktisch** vor? Die nachfolgenden Ausführungen sollen hierzu eine Reihe von Hinweisen geben. Hinsichtlich einer detaillierten testtheoretischen Prüfungsauswertung sind diese sicherlich nicht vollständig, in der Praxis wird mit dem vorgestellten Vorgehen aber die überwiegende Zahl von Problemen bei Aufgaben und Prüfung quantitativ erfasst und damit die notwendige empirische Basis einer zielgerichteten Revision von Aufgaben und Prüfungszusammenstellung gebildet. Schließlich beinhalten sie bei guten Prüfungen auch den Nachweis ihrer Zuverlässigkeit (Reliabilität). Nicht angestrebt ist im Weiteren eine Darstellung der testtheoretischen Hintergründe, hierfür sei auf die umfangreiche Fachliteratur zu diesem Thema verwiesen (z. B. [7], [11][12]). Weiterhin sei angemerkt, dass sich die hier beschriebenen Verfahren im Rahmen der klassischen Testtheorie bewegen, deren Hauptziel in einer möglichst guten Differenzierung der Prüfungskandidaten besteht. Dies bedeutet nicht, dass wir der klassischen Testtheorie den Vorzug vor einer kriteriumsorientierten Leistungsmessung geben, viele der dort entwickelten Vorgehensweisen behalten jedoch auch in diesem Kontext ihren Wert.

[1] *Ruprecht-Karls-Universität Heidelberg, Kompetenzzentrum für Prüfungen in der Medizin - Baden-Württemberg, Heidelberg, Deutschland*

[2] *Universitätsklinikum Heidelberg, Psychosomatische und Allgemeine Klinische Medizin, Heidelberg, Deutschland*

Artikel online frei zugänglich unter http://www.egms.de/en/journals/zma/2006-23/zma000272.shtml
Bitte zitieren als: Möltner A, Schellberg D, Jünger J. Grundlegende quantitative Analysen medizinischer Prüfungen. GMS Z Med Ausbild. 2006;23(3):Doc53.
gQp020060705 31

• Übersicht

In den folgenden Abschnitten wird ein strukturiertes Vorgehen für eine quantitative Prüfungsauswertung vorgeschlagen, welches aus den Hauptpunkten Ergebnisübersicht, Analyse der Aufgabenschwierigkeiten, der Trennschärfen und schließlich der Zuverlässigkeit (Reliabilität) besteht. Voraussetzung dabei ist, dass die Bewertung der Prüfung durch die Vergabe von "Punkten" für Einzelaufgaben erfolgt und die Gesamtbeurteilung an Hand der Summe dieser Punkte erfolgt (der Einfachheit halber sei angenommen, dass die minimale Punktzahl einer Aufgabe mit 0 angenommen ist, Aufgaben mit möglichen "Strafpunkten" werden nicht in Betracht gezogen). Unter "Aufgaben" können Fragen einer Klausur, Stationen eines OSCE-Parcours, zu verschiedenen Zeitpunkten durchgeführte klinische Beobachtungen etc. verstanden werden, die Vorgehensweise ist in allen Fällen die gleiche. Neben einer "generellen" Auswertung hinsichtlich Schwierigkeit und Trennschärfe kann bei Aufgaben, bei denen die Antwort- oder Reaktionsmöglichkeiten in eine begrenzte Zahl von Kategorien eingeteilt werden können, eine detaillierte Antwortanalyse erfolgen. Bekanntestes Beispiel solcher Aufgaben sind Multiple-Choice-Aufgaben, in denen z. B. lediglich fünf Antwortmöglichkeiten vorgegeben sind. Detaillierte Analysen sind jedoch nicht nur auf solche "geschlossenen" Aufgaben (synonym: Aufgaben mit gebundener Beantwortung) beschränkt, sondern können auf alle Aufgaben angewandt werden, in denen die "Lösungsversuche" (richtige wie falsche) der Aufgabe von vorneherein oder auch erst post hoc sinnvoll kategorisierbar sind.

Den Abschluss bilden einige allgemeine Anmerkungen zu den Themen Objektivität, Validität und Reliabilität nicht-homogener Prüfungen, die über das hier gesteckte Ziel, den Inhalt einer grundlegenden Prüfungsauswertung darzustellen, hinausführen.

Auswertungsschritte

• Gesamtergebnis

Vorrangig interessiert alle Beteiligte selbstverständlich das Gesamtergebnis, das in Form einer tabellarischen Auflistung der Häufigkeiten der Gesamtpunktzahlen (siehe Tabelle 1) und als Histogramm (siehe Abbildung 1) dargestellt wird. Ergänzt werden können diese durch die Angabe der Bestehensgrenzen und des Notenschlüssels. Im aufgeführten Beispiel wird von einer Benotung an Hand einer klassischen Notenskala von 1 bis 5 ausgegangen, je nach Studien- oder Prüfungsordnung sind Modifikationen, etwa bei Vergabe von Zwischen- oder Dezimalnoten, möglich.

Die visuelle Inspektion des Histogramms gibt einen meist ausreichenden Überblick über Punkte- und Notenverteilung. Unserer Erfahrung nach sind Prüfungsergebnisse meist nur in grober Annäherung symmetrisch verteilt; bei auffälligen Abweichungen, wie z. B. mehrgipfligen Verteilungen, sollte eine Überprüfung der Ausgangsdaten erfolgen. Häufig zu beobachten sind Ausreißer nach unten, also Studierende mit auffällig wenigen Punkten. Das sind oft Prüfungsabbrecher (ob sie tatsächlich den Raum verlassen haben oder z. B. nur noch entmutigt vor ihrem Fragebogen sitzen ist diesbezüglich egal), von Bedeutung für die statistische Analyse sind sie deshalb, weil sie bei der Analyse von Trennschärfen und Reliabilität zu verzerrten Ergebnissen führen können und deshalb bei einer ernstzunehmenden Analyse von **weiteren Berechnungen ausgeschlossen werden sollten** (selbstverständlich muss auch bei diesen die Korrektheit der Angaben überprüft werden, triviale Fehler bei der Dateneingabe wie das Vergessen einer ganzen Aufgabe oder Aufgabengruppe sind häufiger als man vermutet). Die Identifikation von Ausreißern kann mit objektiven statistischen Methoden erfolgen, wobei zu beachten ist, dass die Normalverteilungsannahme für die Punkteverteilung häufig nicht erfüllt ist, (vgl. z. B. [17]), im Allgemeinen ist jedoch die "optische" Identifikation ausreichend.

Tabelle 1: Häufigkeitsverteilung der erreichten Punktzahlen mit Notenverteilung (Daten aus einer Prüfung im Fach Hygiene/Mikrobiologie/Virologie im Februar 2005 an der Universität Heidelberg). Teilgenommen haben 56 Studierende, maximal waren 136 Punkte erreichbar.

			Kumuliert		
Punkte	*N*	[%]	*n*	[%]	Note
56.0	1	1,8	1	1,8	5
82.5	2	3,6	3	5,4	4
...	...	...	...	...	
94.5	3	5,4	18	32,1	
95.0	1	1,8	19	33,9	3
...	...	...	...	...	
108.5	1	1,8	39	69,6	
109.0	2	3,6	41	73,2	2
...	...	...	...	...	
120.5	1	1,8	55	98,2	
121.5	1	1,8	56	100,0	1

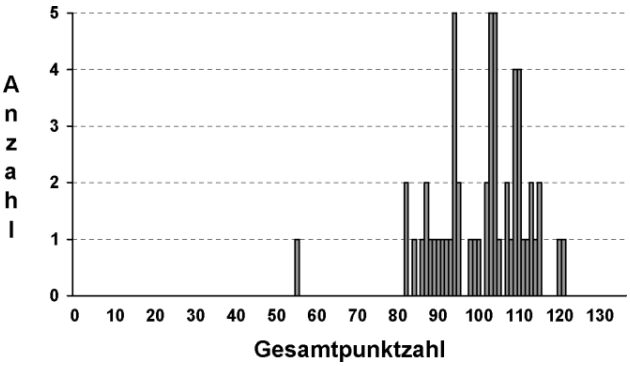


Abbildung 1: Häufigkeitsverteilung der erreichten Punktzahlen. Von weiteren Berechnungen ist der Datenpunkt mit dem niedrigem Gesamtscore S = 56 auszuschließen, da er als Ausreißer zu betrachten ist (Daten der Tab. 1).

• Teststatistische Analyse von Aufgaben

1. Aufgabenschwierigkeit

Nach dem Überblick über die Gesamtergebnisse der Prüfung erfolgt die Untersuchung der Aufgabenschwierigkeiten. In der Literatur wird meist der Begriff "Itemschwierigkeit" ("item difficulty") verwendet, der auf dessen Herkunft aus der psychologischen

Artikel online frei zugänglich unter http://www.egms.de/en/journals/zma/2006-23/zma000272.shtml
Bitte zitieren als: Möltner A, Schellberg D, Jünger J. Grundlegende quantitative Analysen medizinischer Prüfungen. GMS Z Med Ausbild. 2006;23(3):Doc53.

Testtheorie verweist. Da er aber auf jede Form von mit Punkten bewerteten Aufgaben übertragbar ist, soll hier der etwas allgemeinere Begriff "Aufgabe" verwendet werden.

Definiert ist die Aufgabenschwierigkeit als die mittlere bei dieser Aufgabe erreichte Punktzahl $\bar{x}$ (bei Aufgaben, für die bei korrekter Beantwortung genau ein Punkt und ansonsten kein Punkt vergeben wird, stimmt die Schwierigkeit mit der relativen Anzahl von Studenten, die richtig geantwortet haben, überein). Hohe Werte charakterisieren demnach eher leichte, niedrige Werte eher schwere Aufgaben, weshalb - semantisch korrekter - manchmal auch die Bezeichnung "item easiness" zu finden ist.

Zu beachten ist, dass die Schwierigkeit keine Eigenschaft der Aufgabe per se darstellt, sondern immer in Bezug auf die geprüfte Stichprobe oder eine Prüfungspopulation zu sehen ist. Die gleiche Aufgabe in einer Prüfung nach einem Einführungskurs besitzt sicherlich eine andere Schwierigkeit als bei Examenskandidaten. Bei Prüfungen, in denen bei den Aufgaben unterschiedliche Punktzahlen vergeben werden, ist die so definierte Aufgabenschwierigkeit zum Vergleich nicht geeignet, man verwendet stattdessen die "normierte Aufgabenschwierigkeit" $P = \bar{x}$ /max, die den relativen (oder den prozentualen) Anteil an der maximal erreichbaren Punktzahl max einer Aufgabe angibt (beachte: nicht der maximal erreichten Punktzahl in der Prüfung). Aufgaben mit einem Wert von $P = 1$ sind solche, bei denen immer die volle Punktzahl erreicht wurde, bei $P = 0,5$ wurde im Mittel die Hälfte der erreichbaren Punkte erzielt, $P = 0$ erhält man bei Aufgaben, bei denen niemand einen Punkt erreicht hat.

Für die Prüfungsersteller ist dringend zu empfehlen, schon während der Vorbereitung und während des Reviewprozesses eine grobe Einschätzung der Aufgabenschwierigkeiten (erwartete Schwierigkeit $\hat{P}$) zu geben. Dies dient dazu, den Erwartungshorizont hinsichtlich der Lösbarkeit der Aufgaben niederzulegen. Ein Vergleich der erwarteten Punktesummen mit den anzusetzenden Notengrenzen erlaubt eine grobe Vorstellung von den Endergebnissen, manche Prüfung mit überraschend hohen Durchfallquoten hätte dadurch vermieden werden können. Dies ähnelt in gewisser Hinsicht dem Vorgehen beim "Standard-Setting", im Unterschied zu diesem wird jedoch nicht festgelegt, welche Mindestpunktzahl von einem Studenten zum Bestehen erwartet (im Sinne von "gefordert") wird, sondern welche Punktzahl im Mittel erreicht wird.

Richtwerte

Als Richtwert für die Aufgabenschwierigkeit wird in der Literatur der Bereich von etwa 0,4 bis 0,8 empfohlen. Aufgaben mit höheren Werten als 0,8 gelten als zu leicht, aus Sicht der klassischen Testtheorie, in der vor allem der Aspekt einer guten Differenzierung betont wird, sind solche Aufgaben in Hinblick auf die Prüfungsökonomie überflüssig (Warum sollte gefragt werden, was ohnehin fast jeder beherrscht?). Im Sinne einer kriteriumsorientierten Prüfung ist es jedoch durchaus sinnvoll, wichtige basale Fertigkeiten oder Kenntnisse in einer Prüfung unabhängig von ihrer erwarteten Schwierigkeit abzufragen. Für zu schwere Aufgaben gilt ähnliches. Die Asymmetrie (0,4-0,8 statt 0,2-0,8) ist im Wesentlichen damit begründet, dass zu viele schwere Aufgaben in Prüfungen demotivierend wirken.

Die Auflistung der Aufgabenschwierigkeiten erfolgt am besten in einer Tabelle, etwa in der Form von Tab. 2 (s. unten: "Beispiele“), in der die erwarteten und tatsächlichen mittleren erreichten Punktzahlen $\bar{x}$ numerisch aufgeführt sind. Hilfreich ist eine zusätzliche graphische Darstellung der erwarteten und tatsächlichen Aufgabenschwierigkeiten $P = \bar{x}$ /max, da hier deutliche Abweichungen dieser Werte wie auch sehr leichte und schwere Aufgaben leicht zu identifizieren sind.

Von Interesse sind natürlich alle Aufgaben, bei denen deutliche Diskrepanzen zwischen Erwartung und Realität bestehen, sei es, dass sie unerwartet schwer oder unerwartet leicht sind. Für diese ist nach Erklärungen zu suchen, sie können in mangelhaften Aufgabenstellungen begründet sein (etwa den Distraktoren bei Multiple-Choice-Items, unklare Anforderungen bei OSCE-Stationen), oder in fehlerhaften Vorstellungen über die vermittelten Lerninhalte. Ein solches Vorgehen ist zugegebenermaßen ein mühseliges Unterfangen, letztendlich jedoch lohnenswert, da es einerseits eine empirische Basis für Aufgabenrevisionen liefert und zum anderen die Grundlage für die Feedbackschleife von Prüfung zu Curriculumsentwicklung schafft.

Zur zusammenfassenden Darstellung der normierten Aufgabenschwierigkeiten dient deren Histogramm (siehe Abbildung 2). Entsprechend den angegeben Richtwerten sollten diese vor allem im Bereich von 0,4 bis 0,8 streuen, bei feststehenden Bestehensgrenzen (z. B. bei Übernahme der 60%-Grenze der ÄAppO) ist diese zu berücksichtigen und deshalb mehr leichte Aufgaben in der Prüfung zu verwenden (in erster Näherung entspricht die Note, die für die mittlere erreichte Punktzahl vergeben wird auch dem Notendurchschnitt).

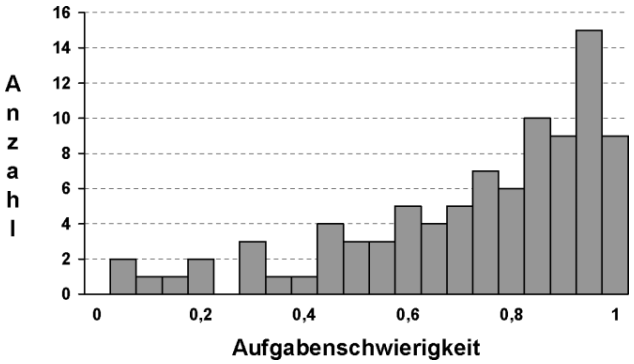


Abbildung 2: Beispiel für ein Histogramm der Aufgabenschwierigkeiten P einer Prüfung. Die Prüfung enthält sechs Aufgaben, die als sehr schwer anzusehen sind (P<= 0, 25, links in der Abb.) sowie eine Reihe von leichten Aufgaben mit Schwierigkeit über P = 0, 9 (Daten der Tab. 1).

Vorsicht ist geboten, wenn ein Teil der Prüfungskandidaten auf Grund der beschränkten Prüfungszeit nicht alle Aufgaben bearbeiten konnte, die daraus folgende Bewertung mit 0 Punkten ist in diesen Fällen nicht auf deren Schwierigkeit zurückzuführen. Auf Grund von Zeitmangel nicht bearbeitete Aufgaben sind bei diesen Prüflingen von der Auswertung auszuschließen. Im speziellen Fall von Multiple-Choice-Klausuren ist dies in der Praxis schwierig zu erfassen, da bei knapper Zeit Prüfungskandidaten verbliebene Fragen mit der Hoffnung auf Zufallstreffer "blind" kreuzen, in den meisten Prüfungen sollte jedoch ohnehin hinreichend Zeit für die Bearbeitung aller Aufgaben zur Verfügung stehen.

Artikel online frei zugänglich unter http://www.egms.de/en/journals/zma/2006-23/zma000272.shtml
Bitte zitieren als: Möltner A, Schellberg D, Jünger J. Grundlegende quantitative Analysen medizinischer Prüfungen. GMS Z Med Ausbild. 2006;23(3):Doc53.

2. Trennschärfe

Damit eine Aufgabe in einer Prüfung Sinn macht, muss sie zwischen guten und schlechten Prüfungskandidaten unterscheiden können (vgl. jedoch hierzu auch Abschnitt "Erhöhung der Reliabilität von Prüfungen"). Besitzt man ein messbares Außenkriterium für "gut" und "schlecht", so wird die Korrelation der bei einer Aufgabe erreichten Punktzahl mit diesem Kriterium als externe (Item-)Validität bezeichnet. Im Rahmen einer Prüfung steht ein solches Außenkriterium im Allgemeinen jedoch nicht zur Verfügung, weshalb man sich damit behilft, das Gesamtergebnis der Prüfung (also die Punktesumme) als Kriterium zu verwenden. Dabei wird angenommen, dass bei inhaltlich adäquater Wahl der Aufgaben diese "im Großen und Ganzen" ein geeignetes Kriterium für "besser" oder "schlechter" darstellen. Den Grad der Übereinstimmung von Aufgabe mit Gesamtpunktzahl bezeichnet man als **Trennschärfe**.

Sie ist die "Fähigkeit" einer Aufgabe zwischen guten und schlechten Prüfungskandidaten zu unterscheiden. Erreichen Kandidaten mit einer hohen Punktzahl in der gesamten Prüfung bei einer Aufgabe relativ viele Punkte, Kandidaten mit niedriger Gesamtpunktzahl nur wenige, so besitzt sie eine hohe Trennschärfe. Eine Aufgabe mit Trennschärfe um 0 wird von guten wie schlechten Prüfungskandidaten gleich gut oder schlecht beantwortet. Denkbar sind auch Aufgaben mit negativer Trennschärfe, diese sind solche, bei denen "paradoxerweise" gute Kandidaten wenig, schlechte Kandidaten hingegen viele Punkte erreichen.

Zur Charakterisierung der Güte, mit der eine Aufgabe gute und schlechte Kandidaten trennt, werden verschiedene Indizes verwendet (man beachte, dass in der Literatur die Terminologie nicht immer einheitlich ist).

Diskriminationsindex

Das anschaulichste Maß ist der Diskriminationsindex *D* ("index of discrimination" [7], [10]). Dieser beruht auf einer Aufteilung der Prüflinge in Gruppen "guter" und "schlechter" Kandidaten an Hand ihrer erreichten Gesamtpunktzahlen. Der Index ist dann definiert als Differenz der normierten mittleren Schwierigkeit der Gruppe der "guten" und der der "schlechten" Kandidaten. In der ursprünglichen Version von Kelley [10] wird dabei eine Einteilung der Prüfungskandidaten am unteren und oberen 27%-Perzentil in drei Gruppen vorgenommen (*D* ist damit die Differenz der 27% besten zu den 27% schlechtesten). Bezeichnet $p^{\uparrow}$ die mittlere erreichte normierte Punktzahl der Gruppe der besten Prüfungskandidaten und $p^{\downarrow}$ die der schlechtesten, so ist $D = p^{\uparrow} - p^{\downarrow}$.

Es gibt verschiedene Varianten zu D, so kann die Gruppeneinteilung an den Terzilen (unteres und oberes 33%-Perzentil statt der 27%-Perzentile) oder eine Einteilung in nur zwei Gruppen am Median erfolgen. Unseres Erachtens ist eine Einteilung in drei Gruppen (gut/mittel/schlecht) vorzuziehen, ob diese an den 33%- oder den 27%-Perzentilen erfolgt ist unwesentlich. Trennt die Aufgabe gut, so ist die Differenz *D* offensichtlich groß, trennt sie schlecht, liegt sie nahe bei 0, negative *D* erhält man bei den bereits erwähnten paradoxen Antwortmustern.

(Die Zahl von 27% ist darauf zurückzuführen, dass *D* ursprünglich als Schnellschätzung für eine Korrelation entwickelt worden ist und eine Aufteilung an den 27%-Perzentilen gewisse Optimalitätseigenschaften aufweist. Als Korrelationsschätzung ist *D* jedoch an zusätzliche Voraussetzungen gebunden; wir behandeln *D* als eigenständiges Maß.)

Ein Nachteil von *D* besteht darin, dass bei der Aufteilung in gute und schlechte Prüfungsabsolventen an Hand der Gesamtpunktzahl, der Punktwert, der bei der betrachteten Aufgabe erreicht wird, diese Aufteilung mit beeinflusst. Es ist deshalb angebracht, die Einteilung auf Basis der erreichten Punktsumme aller anderen Aufgaben vorzunehmen (korrigierter Diskriminationsindex *D*'). Bei Prüfungen, die aus vielen Aufgaben bestehen, ist der Unterschied zwischen *D* und *D*' vernachlässigbar, jedoch sollte die Korrektur z. B. bei OSCE-Prüfungen, die selten mehr als 20, meist jedoch weniger Stationen umfassen, unbedingt durchgeführt werden.

Bei der Ergebnisdarstellung in Tab. 2 (s. unten: "Beispiele") sind die mittleren erreichten Punktzahlen in den Gruppen mit aufgeführt, *D*' ergibt sich optisch als Unterschied zwischen der Schwierigkeit in der Gruppe "gut" und der Gruppe "schlecht".

Korrelationsmaße

Häufiger als *D* werden Korrelationsmaße zur Charakterisierung der Trennschärfe verwendet, insbesondere der übliche Korrelationskoeffizient r nach Pearson-Bravais. Die Trennschärfe ist dann die Korrelation der in der Aufgabe erreichten Punktzahlen mit den Gesamtpunkten in der Prüfung. Bei "unschön" verteilten Daten (stark schiefe Verteilungen, Ausreißer) sind nicht-parametrische Korrelationskoeffizienten (z. B. Spearmans ρ' oder Kendalls τ) als Kennzahlen auf Grund ihrer geringeren Abhängigkeit von Extremwerten oft besser geeignet, nach der in Abschnitt 2 beschriebenen Ausreißerbereinigung sind die Unterschiede zu *r* aber eher gering.

Wie bei *D*' sollte der in der betrachteten Aufgabe erreichte Punktwert nicht bei der Gesamtpunktzahl berücksichtigt werden, analog ist die korrigierte Trennschärfe *r*' (oder ρ') definiert als Korrelation der erreichten Punktzahl mit der Summe der Punkte in allen anderen Aufgaben (die Korrektur wird auch als "part-whole-correction" bezeichnet).

Als Standardvorgehen ist die Verwendung der korrigierten Pearson-Korrelation *r*' zu empfehlen und diese auf Übereinstimmung mit einer nicht-parametrischen Korrelation, z. B. ρ' nach Spearman, zu prüfen. Bei auffälligen Differenzen ($|r' - \rho'| > 0,20$) sollte eine Nachkontrolle bzgl. Ausreißer erfolgen.

Unterschiedliche Bedeutung von *D* und *r*

Die Koeffizienten *r* und *D* unterscheiden sich insofern, als r eher die "Schärfe" der Trennung von guten und schlechten Kandidaten angibt, *D* hingegen eher die "Größe" des Unterschieds. Damit repräsentiert *r* die Trennschärfe besser; in der Praxis zeigt sich jedoch meist, dass das anschaulichere Maß *D* zur Aufgabenbeurteilung im Allgemeinen ausreichend ist.

Aus mathematisch-statistischer Sicht sind Korrelationsmaße als Kenngrößen der Trennschärfe dem "Index of discrimination" *D*

Artikel online frei zugänglich unter http://www.egms.de/en/journals/zma/2006-23/zma000272.shtml
Bitte zitieren als: Möltner A, Schellberg D, Jünger J. Grundlegende quantitative Analysen medizinischer Prüfungen. GMS Z Med Ausbild. 2006;23(3):Doc53.

vorzuziehen, da sie die "Fähigkeit" einer Aufgabe, zwischen guten und schlechten Kandidaten zu unterscheiden auch dann anzeigen, wenn bei den Bewertungen der Aufgabe in der Prüfung nur ein Teilbereich zwischen 0 und maximal erreichbarer Punktzahl verwendet wird.

(Beispiel: Sind bei einer OSCE-Station maximal 20 Punkte zu erreichen und werden vom besten Drittel der Studierenden zwischen 19 und 20 Punkte, vom mittleren Drittel zwischen 18 und 19 und vom schlechtesten Drittel der Studierenden zwischen 17 und 18 Punkte erreicht, so ist die Trennschärfe gemessen mit dem Korrelationskoeffizienten *r* sehr gut ($r \approx 0,8$), gemessen mit *D* jedoch eher mäßig ($D \approx 0,1$).)

Richtwerte

Trennschärfen *r'* von über 0,3 gelten als gut und zwischen 0,2 und 0,3 als akzeptabel. Trennschärfen zwischen 0,1 und 0,2 können noch als marginal angesehen werden, Werte unter 0,1 sind schlecht, hier haben die bei den Aufgaben erzielten Punktzahlen kaum oder nichts mehr etwas mit "guten" oder "schlechten" Prüfungskandidaten zu tun. Aufgaben mit negativen Trennschärfen wirken sich in jedem Fall schädlich auf die Prüfungsgenauigkeit aus. Wie bei den Aufgabenschwierigkeiten dient ein Histogramm der Trennschärfen aller Prüfungsaufgaben als zusammenfassender Überblick (siehe Abbildung 3).

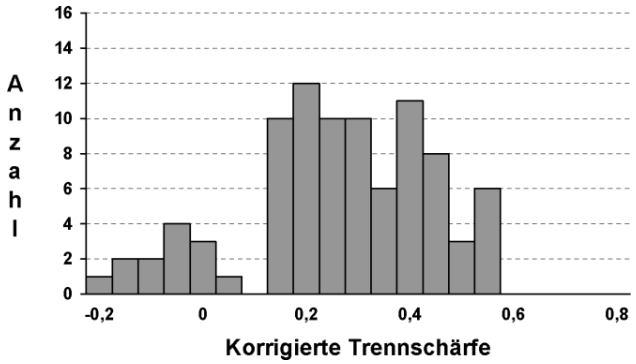


Abbildung 3: Beispiel für ein Histogramm der Trennschärfen *r´* einer Prüfung. Die Aufgaben links in der Abbildung weisen unzulängliche Trennschärfen *r´*<0,1 aus, Aufgaben mit *r´*≥0,2 sind akzeptabel und mit *r´*≥0,3 gut (Daten der Tab. 1).

Man beachte, dass die Trennschärfe einer Aufgabe nicht nur von der Prüfungspopulation abhängig ist (wie die Aufgabenschwierigkeit), sondern auch von der Gesamtheit aller Prüfungsaufgaben. Wird z. B. eine Aufgabe in einer Klausur für Innere Medizin **völlig identisch** in einer Prüfung für Biochemie gestellt und beantwortet, so bleibt deren Schwierigkeit gleich, die Trennschärfe kann jedoch völlig anders sein. Im Rahmen einer Prüfung mit vielen "schlechten" Aufgaben wird auch eine "sehr gute" Aufgabe eine niedrige Trennschärfe aufweisen!

Beispiele

Tabelle 2 illustriert die in diesem Abschnitt beschriebene Kennmaße:

Aufgabe 1 besitzt eine Schwierigkeit von 0,704 und eine gute Trennschärfe ($r' \geq 0,3$). Erwartet wurde von den Prüfern ein Mittel von 7,5 Punkten, die Einschätzung liegt in der Größenordnung des tatsächlichen Werts. Die gute Trennschärfe wird durch den Diskriminationsindex ebenfalls deutlich: Die Gruppe der besten Studenten erreicht 2,08 Punkte ($= D' \cdot max$) oder 0.208 normierte Punkte mehr als die der schlechtesten.

Bei Aufgabe 2 wurden im Mittel statt der erwarteten 8 Punkte 9,32 Punkte erzielt, Aufgabe 2 ist zu leicht ($P > 0,9$), sie weist zudem eine deutlich niedrigere Trennschärfe auf ($r' = 0,177$), gute und schlechte Studenten unterscheiden sich im Mittel um 0,63 Punkte (0.063 normierte Punkte).

Aufgabe 3 ist sehr schwer ($P < 0,4$) und besitzt sogar eine negative Trennschärfe ($r' < 0$), die schlechten Studenten erreichen im Mittel etwas mehr (normierte) Punkte als die guten.

• Einzelantwort- oder Distraktorenanalyse

1. Häufigkeitsanalyse der Antworten

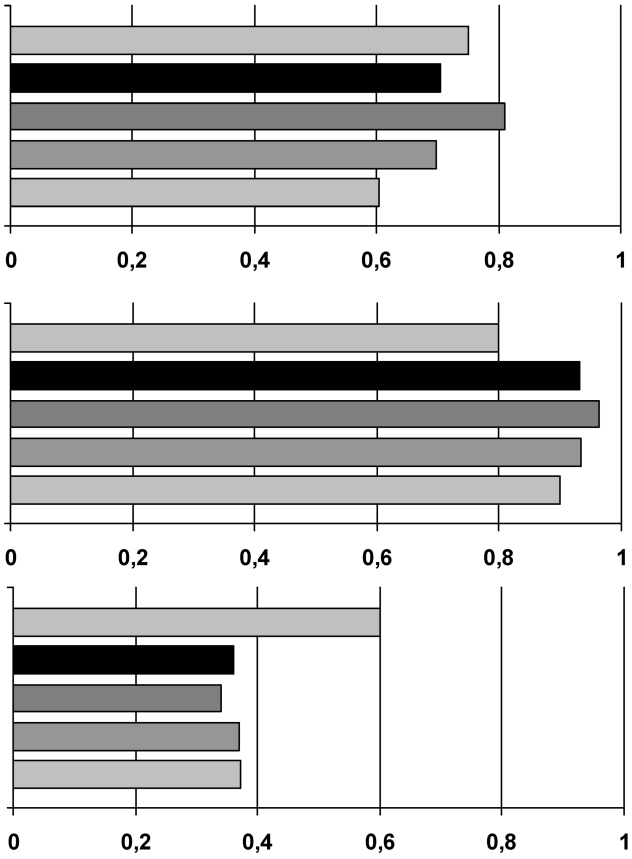
Bei Aufgaben mit beschränkter Zahl möglicher Antworten (wie z. B. Multiple-Choice-Fragen oder Long-Menu-Items) oder nachträglicher Klassifizierung der gegebenen Antworten ist über die Aufgabenschwierigkeiten hinaus eine detailliertere Häufigkeitsanalyse aller Antwortmöglichkeiten, also auch der falschen, von Bedeutung (im Kontext von MC-Fragen werden diese als Distraktoren bezeichnet). Sie dient bei geschlossenen Aufgaben in erster Linie zur Identifikation untauglicher Antwortalternativen, allgemein zur Feststellung, welche Art von Fehlern bei Aufgaben den Studierenden häufig unterlaufen (z. B. kontraindizierte Maßnahmen bei praktischen Prüfungen).

Paradebeispiel ist ein klassisches MC-Item vom Typ A ("Eins aus Fünf"). Die entsprechende Tabelle enthält alle 5 Antwortalternativen und ihre absoluten wie relativen Häufigkeiten, in Tabelle 3 sind zwei Beispiele aufgeführt. Genauer zu betrachten sind hier Aufgaben, in denen eine der Falschkategorien häufiger gewählt wird als die korrekte Antwort. Hier ist nach den Gründen zu fragen, warum die Studierenden eine Falschantwort öfter angeben als die richtige. Ursache dafür ist vielfach eine unklare Abgrenzung der Fehlantwort/des Distraktors von der richtigen Antwort oder die Tatsache, dass die korrekte Antwort doch nicht eindeutig die beste Antwort ist. Weiter gibt die Tabelle Auskunft darüber, welche Falschkategorien/Distraktoren überhaupt oder wie häufig sie gewählt wurden.

Artikel online frei zugänglich unter http://www.egms.de/en/journals/zma/2006-23/zma000272.shtml
Bitte zitieren als: Möltner A, Schellberg D, Jünger J. Grundlegende quantitative Analysen medizinischer Prüfungen. GMS Z Med Ausbild. 2006;23(3):Doc53.

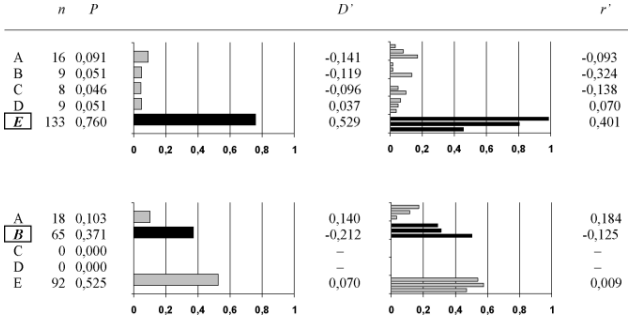
Tabelle 2

Aufgabe		$\bar{x}$	*max*		Kennmaße
1	erwartet	7,50	10	$\hat{P}$	
	gesamt	7,04	10	P	P = 0,704
	gut	8,11	10	$P^{\uparrow}$	D' = 0,208
	mittel	6,99	10	$P^{\updownarrow}$	r' = 0,318
	schlecht	6,03	10	$P^{\downarrow}$	ρ' = 0,256
2	erwartet	8,00	10	$\hat{P}$	
	gesamt	9,32	10	P	P = 0,923
	gut	9,63	10	$P^{\uparrow}$	D' = 0,063
	mittel	9,33	10	$P^{\updownarrow}$	r' = 0,177
	schlecht	9,00	10	$P^{\downarrow}$	ρ' = 0,168
3	erwartet	6,00	10	$\hat{P}$	
	gesamt	3,61	10	P	P = 0,361
	gut	3,40	10	$P^{\uparrow}$	D' =-0,033
	mittel	3,70	10	$P^{\updownarrow}$	r' = -0,089
	schlecht	3,73	10	$P^{\downarrow}$	ρ' =-0,069



Auflistung erwarteter und tatsächlicher Aufgabenschwierigkeiten ($\hat{P}$ bzw. P) in der Gesamtgruppe und nach Aufteilung in gute, mittlere und schlechte Kandidaten ($P^{\uparrow}$, $P^{\updownarrow}$ bzw. $P^{\downarrow}$). Der obere graue Balken visualisiert die von den Prüfungserstellern erwartete normierte Aufgabenschwierigkeit, der schwarze Balken darunter die tatsächliche Schwierigkeit in der Prüfung. Letztere entspricht der zu beurteilenden Schwierigkeit P, die i. A. zwischen 0,4 und 0,8 liegen sollte. Aus dem P der guten und dem der schlechten Kandidaten errechnet sich der Diskriminationsindex D'. Als weitere Trennschärfemaße sind r' und ρ' angegeben (Erläuterungen zu den 3 Beispielen: s. Text).

Tabelle 3: Häufigkeitsanalyse der Antworten (Distraktoranalyse). Oben das Beispiel eines "guten" Items (mit der korrekten Antwort E), bei dem die Schwierigkeit P = 0, 76 im Bereich zwischen 0,40 und 0,80 liegt und alle Distraktoren auch gewählt wurden. Die Trennschärfe ist mit r' = 0, 401 bzw. D' = 0, 529 sehr gut, die Distraktoren weisen negative Trennschärfen oder eine Trennschärfe um 0 (bei Antwort D) auf. Unten ein "schlechtes" Item (korrekte Antwort B): Das Item ist eher schwer (P = 0, 371), Distraktor E wurde häufiger als die richtige Antwort B, die Distraktoren C und D werden hingegen überhaupt nicht gewählt. Die Trennschärfe der korrekten Antwort ist negativ (r' = -0, 125), ein Distraktor (Falschantwort) besitzt eine positive Trennschärfe.



Bei geschlossenen Aufgaben mit wenigen Distraktoren ist bei nie oder nur selten gewählten Falschantworten zu vermuten, dass diese auch bei Unkenntnis der richtigen Antwort schon von vornherein ausgeschlossen werden können. Dadurch erhöht sich die Wahrscheinlichkeit, die richtige Lösung zu erraten, deutlich, weshalb diese Antwortmöglichkeiten einer näheren Prüfung hinsichtlich ihrer Tauglichkeit zu unterziehen sind.

Erfahrungsgemäß werden bei dieser Antwortanalyse die meisten fehlerhaften Aufgaben identifiziert, die im Nachhinein eine Korrektur der Bewertungsvorgaben erforderlich machen.

2. Trennschärfenanalyse von Antworten/Distraktoren

Bei einer detaillierten Antwortanalyse werden analog zur Aufgabenschwierigkeit nicht nur die korrekten sondern auch die falschen Antworten (bei MC-Items: Distraktoren) ausgewertet.

(Anmerkung: Mathematisch lassen sich "Antworten" nicht mit einer Punktzahl korrelieren, man bildet vielmehr für jede Antwortmöglichkeit (z. B. A, B, C, D und E) eine sog. Indikator-Variable, die für jeden Prüfungskandidaten den Wert 1 aufweist, falls dieser

Artikel online frei zugänglich unter http://www.egms.de/en/journals/zma/2006-23/zma000272.shtml
Bitte zitieren als: Möltner A, Schellberg D, Jünger J. Grundlegende quantitative Analysen medizinischer Prüfungen. GMS Z Med Ausbild. 2006;23(3):Doc53.

die Antwort gewählt hat, ansonsten den Wert 0. Die Berechnungen von Trennschärfemaßen beziehen sich auf diese Indikator-Variablen.)

Korrekte Antwortmöglichkeiten sollten eine hohe Diskrimination aufweisen, falsche Antwortmöglichkeiten eine negative, d. h. "gute" Studierende wählen richtige Antworten häufiger und falsche Antworten seltener als schlechte Studierende. Für korrekte Antwortmöglichkeiten gilt das im letzten Abschnitt Erwähnte, niedrige Trennschärfen zeigen einen ungenügenden Zusammenhang zwischen Antwort und Gesamtscore.

Anders ist es bei den falschen Antwortmöglichkeiten, hier sind positive Trennschärfen auffällig, man sollte überlegen, warum gute Studenten eine solche (falsche) Antwort häufiger wählen als schlechte.

Die Ergebnisdarstellung enthält wie bei der allgemeinen Aufgabenanalyse eine graphische Darstellung der relativen Häufigkeiten in den drei Gruppen gut/ mittel/schlecht (s. Tab. 3).

(Man beachte, dass bei den klassischen MC-Items vom Typ A ("Eins aus Fünf") die Trennschärfe des Items mit der Trennschärfe der korrekten Antwort übereinstimmt, dies gilt jedoch nicht für alle Antwortformate. Bei Items vom Typ K' (für die Typenbezeichnungen von MC-Items s. [2]) sind auf eine Frage vier Ja-Nein-Alternativen zu beantworten, ein Punkt wird für das Item nur gegeben, wenn alle vier Antworten korrekt sind. In diesem Fall ist die Trennschärfe des Gesamtitems etwas anderes als die Trennschärfen der korrekten Antworten auf einzelnen Unterfragen.)

• Reliabilität

1. Bedeutung der Reliabilität

Die Reliabilität bezeichnet die Zuverlässigkeit oder Reproduzierbarkeit von Prüfungsergebnissen [4]. Um die inhaltliche Bedeutung der Zahlenwerte zu veranschaulichen, sei ein fiktives Beispiel vorangestellt. Ein Prüfer habe einen hinreichend großen Fragenpool zur Verfügung und beabsichtigt, die Abschlussprüfung seines Fachgebiets mit 20 Fragen durchzuführen. Um eine Abschätzung darüber zu gewinnen, ob die Fragenzahl für eine zuverlässige Prüfung ausreicht, stellt er beim ersten Mal zwei Prüfungen mit je 20 verschiedenen Fragen aus seinem Pool zusammen und setzt alle 40 Fragen bei 250 Studierenden ein. Danach bewertet der die beiden Prüfungsteile unabhängig voneinander und vergleicht die dabei erzielten Noten.

Für die verschiedenen Reliabilitätswerte 0,5, 0,65, 0,8 und 0,9 zeigt die Tabelle 4 die zu erwartenden Häufigkeiten der Notenkombinationen.

Tabelle 4: Beispiel zur Übereinstimmung von Noten bei zwei Prüfungen in Abhängigkeit von ihrer Reliabilität (Erläuterung siehe Text).In den Spalten sind für die Reliabilitätswerte *r* = 0,5, 0,65, 0,8 und 0,9 die zu erwartenden Notenkombinationen bei zwei Prüfungen aufgeführt. Z. B. erreichen bei *r* = 0,50 12 Studierende in beiden Prüfungen die Note Eins, 20 jeweils die Note Zwei usw., insgesamt erhalten 84 Studierende (etwa ein Drittel) in beiden Prüfungen die gleiche Note. 23 Prüfungskandidaten erhalten einmal eine Eins und einmal eine Zwei usf. und schließlich ein Student sogar einmal eine Eins und einmal eine Fünf!

Noten	*r*=0,50	*r*=0,65	*r*=0,80	*r*=0,90
zweimal 1	12	15	20	23
zweimal 2	20	23	28	35
zweimal 3	27	31	37	46
zweimal 4	17	19	24	31
zweimal 5	8	11	14	17
gleiche Note	84	99	123	152
eine 1, eine 2	23	23	22	18
eine 2, eine 3	39	39	37	32
eine 3, eine 4	35	36	35	30
eine 4, eine 5	17	18	18	15
Differenz 1	114	116	112	95
eine 1, eine 3	14	10	5	1
eine 2, eine 4	17	12	6	1
eine 3, eine 5	12	9	4	1
Differenz 2	43	31	15	3
eine 1, eine 4	4	2	0	0
eine 2, eine 5	4	2	0	0
Differenz 3	8	4	0	0
eine 1, eine 5	1	0	0	0
Differenz 4	1	0	0	0

(Dem Beispiel liegt folgendes Modell zu Grunde: Die Punktwerte sind bivariat normalverteilt mit Erwartungswerten 0 und Standardabweichungen 1 bei einer Korrelation von r . Als Notengrenzen sind für die beiden Prüfungen jeweils die Werte x_{45} = -1, 28155, $x_{34} = x_{45}$ +0, 8, $x_{23} = x_{45}$ +1, 6 und $x_{12} = x_{45}$ +2, 4 gewählt, so dass die Notenverteilung der beiden Prüfungen für die Noten 1 - 5 etwa 13%, 24%, 31%, 22% und 10% ist. Dies entspricht einer realistischen Notenverteilung in einer Prüfung. Die Erwartungswerte der Notenkombinationen ergeben sich aus der bivariaten Normalverteilung in Abhängigkeit von der Korrelation *r*.)

Aus der Tabelle 4 geht hervor, dass bei einer Reliabilität von 0,5 lediglich 84 der 250 Prüflinge die gleiche Note erhalten, bei 114 Studierenden differieren die beiden Notenwerte um eine Stufe. Während Notenunterschiede von einer Stufe noch nicht als gravierender Mangel angesehen werden brauchen, stellt die Tatsache, dass bei mehr als 20% der Prüfungskandidaten ein Unterschied von mindestens zwei Notenstufen vorliegt, den Wert einer solchen Prüfung und ihrer Benotung sicher in Frage. Besonders auffallend sind natürlich die Fälle, in denen einmal eine Fünf und einmal eine Note gleich oder sogar besser als Drei vergeben wird (17 Prüflinge, entsprechend 6,8%). Eine zuverlässige Identifikation ungenügend vorbereiteter Kandidaten ist mit einer solchen Prüfung offensichtlich nicht möglich.

Nun ist bei schriftlichen Prüfungen eine Reliabilität von 0,5 oder niedriger unserer Erfahrung nach eher selten, die in der zweiten

Artikel online frei zugänglich unter http://www.egms.de/en/journals/zma/2006-23/zma000272.shtml
Bitte zitieren als: Möltner A, Schellberg D, Jünger J. Grundlegende quantitative Analysen medizinischer Prüfungen. GMS Z Med Ausbild. 2006;23(3):Doc53.

Spalte der Tabelle angegebenen Werte für eine Reliabilität von 0,65 sind in der Praxis jedoch durchaus häufig zu finden. Die Benotung ist hier schon deutlich stabiler, aber bei immer noch 35 Personen - das sind 14% der Studierenden - ist die Differenz der beiden Noten 2 oder 3.

In der Literatur wird für relevante Prüfungen eine **Mindestreliabilität von 0,8** angegeben, die dabei zu erwartenden Kombinationen sind in der dritten Spalte aufgeführt. Auch hier erhält noch nicht einmal die Hälfte der Personen die gleiche Note, der Anteil von Differenzen um zwei Notenstufen beträgt noch 6%. Höhere Differenzen treten nur noch mit einer zu vernachlässigenden Wahrscheinlichkeit auf. Erst ab einer Reliabilität von 0,9 steigt der Anteil von gleichen Noten auf mehr als 2/3 der Prüflinge (siehe letzte Spalte der Tabelle 4).

Bedenkt man die Bedeutung, die den Noten beigemessen werden und die Konsequenzen für den Studenten, der eine Prüfung nicht besteht, so dürfte der Mindestwert von 0,8, der allgemein für aussagekräftige Prüfungen gefordert wird, sicherlich nicht zu niedrig angesetzt sein (jeder Prüfungsersteller sollte für die in Tabelle 4 angegebenen Zahlenwerte überlegen, ob er seine eigenen Leistungen mit einem Test beurteilt wissen will, der die genannten Güteeigenschaften aufweist).

2. Formen der Reliabilität und interne Konsistenz

Um die Zuverlässigkeit einer Prüfung festzustellen, wäre das Naheliegendste natürlich eine Prüfungswiederholung. Das ist bei Lernstoffen jedoch nicht praktikabel, da sich die Prüfungskandidaten an die einzelnen Aufgaben erinnern und die zugehörigen Antworten lernen (in anderen Zusammenhängen ist die Bestimmung einer solchen Re-Test-Reliabilität durchaus sinnvoll). Eine andere Möglichkeit besteht in der Konstruktion eines "Paralleltests", also einer Prüfung, die die selben Prüfungsinhalte, aber mit anderen Fragen oder Aufgaben enthält. Die Reliabilität ist dann die Korrelation der beiden Prüfungsergebnisse. Hierzu sind keine zwei Prüfungstermine notwendig, die Aufgaben der beiden Prüfungen könnten direkt nacheinander oder ineinander verschränkt dargeboten werden (der fiktive Prüfer im vorhergehenden Abschnitt hat diesen Weg gewählt). Praktisch ist dieser Weg zwar gangbar, erfordert aber eine entsprechende Konstruktion der Prüfung mit genauer Definition der Prüfungsziele und des -inhalts ("Blue-Print"). Ohne Paralleltestkonstruktion kann man die Prüfung nach einem bestimmten vorgegebenen Schema (z. B. Prüfung 1 besteht aus den ungeraden Prüfungsfragen, Prüfung 2 aus den geraden) oder nach Zufall in zwei Hälften teilen und die Korrelation der Ergebnisse der beiden Hälften bestimmen ("Split-Half-Reliabilität", meist werden die beiden Testhälften als gleich groß angenommen, was zwar gewisse Optimalitätseigenschaften beinhaltet, jedoch nicht zwingend notwendig ist).

Für die beiden Prüfungshälften hat man je Kandidat zwei Prüfungsergebnisse S_1 und S_2, die Split-Half-Reliabilität ist die Korrelation der beiden Werte und kann als Zuverlässigkeit der Prüfung 1 wie auch der Prüfung 2 interpretiert werden. Für die Beurteilung der Kandidaten wird man die beiden Prüfungspunktwerte S_1 und S_2 zusammenfassen, ob man die beiden Werte summiert oder mittelt, macht keinen Unterschied. Mathematisch lässt sich unter relativ schwachen Modellannahmen zeigen, dass die Verwendung der zusammengefassten Punktzahl die Zuverlässigkeit erhöht, die entsprechende Formel, mit der aus der Korrelation der beiden Einzelprüfungen auf die Zuverlässigkeit der zusammengefassten geschlossen werden kann, ist als Spearman-Brown-Formel bekannt ([12], vgl. auch den nächsten Abschnitt).

Es bleibt jetzt noch eine kleine Unschönheit im genannten Vorgehen zu behandeln: Die Aufteilung in zwei Hälften ist willkürlich, bei einer anderen Aufteilung würde man auch einen anderen Wert für die Reliabilität erhalten. Abhilfe schafft hier Cronbachs α-Koeffizient (sog. interne Konsistenz)t: Bestimmt man - so wie im letzten Abschnitt angegeben - die Reliabilität des Gesamttests aus allen möglichen Testhalbierungen, so ist dieser Mittelwert identisch mit α.

Eine Anmerkung ist noch vonnöten: Die Reliabilität ist mathematisch nicht mit der internen Konsistenz identisch. In vielen Veröffentlichungen wird α mit der Reliabilität gleichgesetzt, dies ist falsch. Cronbachs α ist tatsächlich eine untere Grenze für die Reliabilität, d. h. es gilt allgemein $\alpha \leq r$. Die Reliabilität kann jedoch u. U. deutlich höher als α sein. Die Verwendung besserer Abschätzungen erfordert jedoch einen höheren Aufwand und statistische Fachkenntnisse, für das hier angestrebte Ziel eines How-to-do einer einfachen Prüfungsauswertung ist α in jedem Fall adäquat, zumal man sich mit einem ausreichend hohen α auf der sicheren Seite befindet.

Zur praktischen Berechnung ist die Verwendung eines Statistik-Programmpakets wie z. B. SPSS, SAS oder das frei erhältliche R, dringend zu empfehlen, diese liefern neben der Reliabilität auch noch ein weiteres Maß zur Beurteilung der einzelnen Aufgaben. Eine "gute" Aufgabe steuert Informationen zur Differenzierung der Prüfungskandidaten bei, als einfaches Maß dient dabei die Trennschärfe. Gleichzeitig dient sie zur Erhöhung der Messzuverlässigkeit, weshalb diese sinken sollte, wenn man die Aufgabe in der Prüfung weggelassen hätte. Umgekehrt dürfte eine schlechte Aufgabe die Messzuverlässigkeit absinken lassen, ihre Herausnahme aus der Prüfung würde die Zuverlässigkeit sogar erhöhen. Aus diesem Grund bestimmt man für jede Aufgabe die Zuverlässigkeit der restlichen Prüfung, das "α if deleted". Als problematisch sind dann all die Aufgaben anzusehen, bei denen das "α if deleted" höher ist als das α der Gesamtprüfung, diese Aufgaben wirken **reliabilitätsmindernd** (sofern man Cronbachs α als Reliabilitätsmaß verwendet, s. hierzu noch die Anmerkungen zur Reliabilität nicht-homogener Tests weiter unten) Praktisch stimmt dieses Kriterium mit dem der Trennschärfe nahezu überein, sollte der Vollständigkeit halber jedoch mitbestimmt werden.

3. Erhöhung der Reliabilität von Prüfungen

Wie bereits erwähnt, wird in der Literatur ein Mindestwert von 0,8 für die Reliabilität angegeben. Die Praxis zeigt, dass dieser Wert jedoch bei vielen Prüfungen unterschritten wird. Zunächst sei bemerkt, dass - etwas vereinfachend - die Reliabilität mit den mittleren Korrelationen der Punktwerte der Aufgaben untereinander und der Anzahl der Aufgaben ansteigt. Damit sind auch die beiden Möglichkeiten angegeben, die Reliabilität von Prüfungen zu steigern: Eine Verbesserung der Aufgabenqualität führt meist zu deutlichen Erhöhungen der Interkorrelationen und dadurch zur Reliabilitätssteigerung. Ein Beispiel für den Erfolg solcher Bemühungen ist der kontinuierliche Anstieg der Zuverlässigkeiten der an der medizinischen Fakultät Heidelberg durchgeführten OSCE-Prüfungen, bei der die geforderte Mindestgrenze von 0,8 nahezu erreicht wurde [16].

Artikel online frei zugänglich unter http://www.egms.de/en/journals/zma/2006-23/zma000272.shtml
Bitte zitieren als: Möltner A, Schellberg D, Jünger J. Grundlegende quantitative Analysen medizinischer Prüfungen. GMS Z Med Ausbild. 2006;23(3):Doc53.
gkp(c)20060703.1)

Eine andere Möglichkeit zur Reliabilitätserhöhung ist, mehr Aufgaben zu stellen. Dies stößt natürlich auf praktische Grenzen, so dass unseres Erachtens immer der Qualitätsverbesserung der Aufgaben der Vorzug gegeben werden sollte. Um eine Abschätzung zu gewinnen, wie sich eine Veränderung der Aufgabenzahl unter sonst gleichen Bedingungen auswirkt, ist nachstehende Formel hilfreich: Hat eine Prüfung n Aufgaben und ist r_n ihre Reliabilität, so ist

$$r_m = \frac{m\, r_n}{n + (m-n)\, r_n}$$

die entsprechende Reliabilität einer Prüfung mit m Aufgaben (für $m = 2$ und $n = 1$ erhält man die oben erwähnte Formel von Spearman-Brown). Durch Umstellen der Formel kann man aus r_n und n die notwendige Fragenzahl m bei angezielter Reliabilität r_m bestimmen:

$$m = \frac{r_m\,(1-r_n)}{r_n\,(1-r_m)}\, n$$

Verwenden wir das am Anfang des Abschnitts angegebene Beispiel einer Prüfung mit $n = 20$ und Reliabilität $r_n = 0,5$. Um eine Reliabilität von 0,8 zu erzielen, benötigt man

$$m = \frac{0,8(1-0,5)}{0,5(1-0,8)}\,20 = \frac{0,4}{0,1}\,20 = 80$$

Aufgaben, will der Prüfer eine Vervierfachung seiner Prüfungsdauer vermeiden, muss er die Prüfungsqualität erheblich erhöhen.

Natürlich ist es auch möglich, "eine" Prüfung an mehreren Terminen durchzuführen (z. B. vier Teilprüfungen mit je 20 Aufgaben statt einer einzigen Abschlussprüfung mit Aufgaben). Die Zuverlässigkeit der Gesamtbewertung zusammengefasster Teilprüfungen bestimmt sich aus allen Einzelaufgaben.

Die Reliabilität kann als das quantitative Hauptkriterium einer Prüfung angesehen werden, da sie angibt, wie zuverlässig die Prüfungsergebnisse sind. Man wird einwenden, dass die Validität eigentlich das Hauptkriterium darstellen sollte. Hierfür hat man aber bei Prüfungen meist keine geeignete quantitative Abschätzung, es kann i. A. nur eine hinlängliche Inhaltsvalidität über eine repräsentative Abdeckung der Lehrinhalte durch die Prüfungsfragen gesichert werden (siehe weiter unten), dennoch können empirisch nicht belegbare Spekulationen über die Validität einer Prüfung eine mangelhafte Messzuverlässigkeit nicht heilen.

Auf ein damit zusammenhängendes Problem sollte noch hingewiesen werden: Prüfungen dienen nicht allein dazu, zuverlässig Unterschiede zwischen Studierenden zu quantifizieren, sie sollen und **müssen** auch das für das Fach notwendige Basiswissen und die grundlegenden Fertigkeiten abprüfen. Eine gute Lehre wird diese auch erfolgreich vermitteln und dann werden auch die meisten Studierenden die grundlegenden Kenntnisse beherrschen und die entsprechenden Aufgaben erfolgreich bearbeiten. Solche Aufgaben weisen aber schlechte Kennwerte auf: Sie sind viel zu leicht, besitzen nur eine geringe Trennschärfe und vermindern die interne Konsistenz der Prüfung (man beachte: sie verringern **nicht** die Reliabilität, steigern sie aber auch nicht). Daraus folgt, dass die oben beschriebenen Gütemaße nicht alleiniges Kriterium für die Auswahl von Aufgaben sein können; das Ziel einer hohen Reliabilität darf nicht dazu führen, eine fachlich inadäquate Prüfung durchzuführen (Stichwort "Kolibri-Fragen"). Jeder Prüfer sollte aber wissen, welche Aufgaben "schlecht" im Sinne der klassischen Testtheorie sind und, - wenn Aufgaben mit unzureichenden Gütemaßen in Prüfungen eingesetzt werden - gute Gründe für ihre Verwendung nennen können (siehe hierzu auch "… und was ist mit der Validität" weiter unten).

Zusammenfassung

Nachfolgend sind die einzelnen Schritte der hier vorgestellten quantitativen Prüfungsbeurteilung zusammen gefasst (Tabelle 5 enthält noch einmal alle zu berücksichtigenden Kennwerte)

Tabelle 5: Übersicht der grundlegenden Gütemaße

Kennwerte der Aufgaben	
P	Aufgabenschwierigkeit: mittlere, auf das Intervall zwischen 0 und 1 normierte Punktzahl
D'	Diskriminationsindex: Differenz der normierten Punktzahl zwischen „guten“ und „schlechten“ Prüfungsabsolventen
r'	korrigierte Trennschärfe: Korrelation der Punktzahl mit Punktsumme aller anderen Aufgaben
ρ'	Rangkorrelation der Punktzahl mit Punktsumme aller anderen Aufgaben (zur Berücksichtigung von Ausreißern)
„α if deleted“	interne Konsistenz bei Nichtberücksichtigung der Aufgabe
Kennwerte der gesamten Prüfung	
Cronbachs α	interne Konsistenz: untere Grenze der Reliabilität (Reliabilität ist mindestens so hoch wie α)
$m_{0.8}$	Geschätzte Anzahl von Aufgaben zum Erreichen einer Reliabilität von 0.8

- 1. Ergebnisübersicht
- 1.1 Häufigkeitstabelle der Punktwerte und Noten
- 1.2 Histogramm der Punktwerte
- 1.3 Identifikation von Ausreißern
- 2. Teststatistische Analyse der Aufgaben
- 2.1 Berechnung der Aufgabenschwierigkeiten P
- (mit graphischer Darstellung der P)
- 2.2 Berechnung der Trennschärfeindizes D', r', ρ' der Aufgaben
- (mit graphische Darstellung der Schwierigkeiten in den Untergruppen gut/mittel/schlecht)
- 2.3 Histogramme der Aufgabenschwierigkeiten P und Trennschärfen r'
- 3. Einzelantwort-/Distraktoranalyse (sofern sinnvoll)
- 3.1 Bestimmung der Häufigkeiten der Einzelantworten
- (mit graphischer Darstellung)
- 3.2 Bestimmung der Diskrimination der Einzelantworten r' und D'
- (mit graphischer Darstellung in den Untergruppen)
- 4. Reliabilität
- 4.1 Bestimmung von Cronbachs α (nach Ausschluss von Ausreißen)
- 4.2 Bestimmung von "α if deleted" für jede Aufgabe
- 4.3 Berechnung der notwendigen Zahl $m_{0.8}$ von Prüfungsfragen für Mindestreliabilität 0,8.

Nach der in Schritt 1 erfolgenden Übersicht über die Prüfungsresultate, die auch einer weiteren Kontrolle der Bewertung dient,

Artikel online frei zugänglich unter http://www.egms.de/en/journals/zma/2006-23/zma000272.shtml
Bitte zitieren als: Möltner A, Schellberg D, Jünger J. Grundlegende quantitative Analysen medizinischer Prüfungen. GMS Z Med Ausbild. 2006;23(3):Doc53.

liefern die Schritte 2 und 3 die empirische Basis für die Revision von Aufgaben und Prüfungszusammenstellung. Revisionsbedürftig sind zu leichte, zu schwere oder wenig trennscharfe Aufgaben. Diese Analyseschritte sollten vor Bekanntgabe der Prüfungsergebnisse erfolgen, da hier problematische Aufgaben (unklare Aufgabenstellungen, mehrdeutige oder sogar fehlerhaften Antwortmöglichkeiten) meist auffallen und bei der Korrektur berücksichtigt werden können. Dies ist in jedem Fall besser als die nachträgliche Berücksichtigung gerechtfertigter Einsprüche von Studenten gegen Prüfungsaufgaben.

Der abschließende Schritt 4 dient zur Kontrolle der Zuverlässigkeit der Prüfung; bei ungenügender Reliabilität muss eine Verbesserung der Aufgabenqualität und/oder Vergrößerung der Aufgabenanzahl erfolgen, was natürlich erst bei den folgenden Prüfungen zum Tragen kommt (s. "Erhöhung der Reliabilität von Prüfungen").

Wichtig für die Praxis ist eine kompakte Darstellung der Ergebnisse der Aufgabenauswertung, die Standardausgaben von Programmpaketen sind meist zu "unhandlich" und zu umfangreich. Es ist einigermaßen mühsam, sich bei der Betrachtung der Ergebnisse durch Dutzende Seiten Papier oder Bildschirmdarstellungen durchzuarbeiten.

Unserer Erfahrung nach hat sich bei der Standardauswertung für den ersten Schritt eine einfache tabellarische Darstellung bewährt, in der je Aufgabe die Kennwerte P, D', r', α' und "α if deleted" sowie ein „Warnhinweis" aufgeführt werden. Gewarnt wird bei zu schweren oder sehr leichten Aufgaben ($P < 0{,}4$ bzw. $P > 0{,}85$) und bei zu geringer Trennschärfe ($r' < 0{,}2$) sowie diskrepanten Trennschärfen $|r' - \alpha'| > 0{,}2$. Damit lassen sich bei einer Prüfung mit 60 Aufgaben die wesentlichen Ergebnisse leicht auf zwei Seiten darstellen. Im zweiten Schritt werden nur für die problematischen Aufgaben die Details betrachtet (hier erweisen sich die graphischen Darstellungen als sehr hilfreich, **nur** für eingefleischte Zahlenliebhaber sind diese platzbedürftigeren Darstellungen überflüssig).

Ähnlich geht man bei der Einzelantwort-/Distraktorenanalyse vor, die umfangreicheren Darstellungen werden nur bei den Aufgaben verwendet, die hinsichtlich der Häufigkeiten oder Trennschärfen kritisch sind.

Natürlich gibt es hier bei den Prüfern auch unterschiedliche Vorlieben, manche bevorzugen auch bereits im ersten Schritt die graphischen Varianten der Tabellen 2 und 3 .

Anmerkung

Zu den klassischen Gütekriterien zählen neben der Reliabilität noch Objektivität und Validität. Da hier im Wesentlichen die quantitative Auswertung von Prüfungsdaten thematisiert ist, sollen hier nur einige Anmerkungen angefügt werden, bei den standardmäßig durchgeführten Prüfungen liegen nämlich weder für Objektivität noch Validität überhaupt empirische Daten vor.

1. Objektivität

Bei der Objektivität unterscheidet man üblicherweise Durchführungs-, Auswertungs- und Interpretationsobjektivität. Erstere lässt sich durch eine hinreichende Standardisierung des Prüfungsablaufs gewährleisten, bei schriftlichen Klausuren ist diese üblicherweise gegeben, für mündliche und praktische Prüfungen ist eine genügende Durchführungsobjektivität bei OSPE/OSCE, OSLER, strukturierten mündlichen Prüfungen u. a. Prüfungsformaten zu sichern. Kritischer ist die Auswertungsobjektivität, bei der bei Prüfungen noch zwischen der Objektivität der Leistungsfeststellung und der der -beurteilung zu unterscheiden ist. Zu deren Sicherung sollte - soweit logistisch und personell möglich - wenigstens eine stichprobenartige Überprüfung durch den Einsatz voneinander unabhängiger Bewerter (Korrektoren oder Prüfern) erfolgen.

2. . . und was ist mit der Validität?

Die Reliabilität gibt nur an, wie genau die Ergebnisse sind, nicht jedoch, ob das was gemessen wird, auch das ist, was man überhaupt messen will. Bekannt ist hierzu die Diskussion um den Stellenwert von Multiple-Choice-Klausuren, da sich bei diesen die Frage stellt, inwieweit damit wirklich die für die ärztliche Tätigkeit notwendigen Kenntnisse in einem medizinischen Teilgebiet gemessen werden oder schlicht nur die mittelfristige Merkfähigkeit der Studierenden.

Wenn dem so ist, führen dann Bemühungen zur Reliabilitätssteigerung nicht windschief an der eigentlichen Zielsetzung der Prüfung medizinischer Kenntnisse und Fertigkeiten vorbei? Richtig an dieser kritischen Frage ist, dass natürlich das **wesentliche Ziel "guter Prüfungen" eine möglichst hohe Validität** ist, andererseits ist die Reliabilität eine notwendige Bedingung, um mit einer Prüfung überhaupt etwas zu erfassen **(eine Prüfung die nicht reliabel ist, ist auch nicht valide!).**

Das Problem der Validität besteht darin, objektive und nachprüfbare Außenkriterien zur Validitätsprüfung anzugeben (externe oder Kriterienvalidität, "Gold-Standard"), weshalb als Validitätskonzept in der Forschung die Konstruktvalidität in den Vordergrund getreten ist. (vgl. [8][15] und speziell in Bezug auf medizinische Prüfungen [6], [19], [21]).

(In der Literatur finden sich neben den Begriffen der Kriterien- (externen) und Konstruktvalidität auch noch etwa die Augenschein- ("face-validity") und die Kontent- oder Inhaltsvalidität. Beide Begriffe sind auf Grund ihres subjektiven Charakters lediglich als heuristische Kriterien verwendbar, für eine wissenschaftliche Argumentation jedoch kaum von Nutzen.)

Hierzu sind für medizinische Prüfungen und deren Aussagekraft für praktische ärztliche Tätigkeiten eine Reihe von Untersuchungen veröffentlicht worden (Reviews der Literatur finden sich etwa in [9], [14], [20]), die einerseits durchaus deren Validität belegen, andererseits aber zeigen, dass Prüfungsformen nicht per se eine höhere oder geringere Aussagekraft aufweisen, sondern die Qualität der einzelnen Prüfungen ausschlaggebend ist und erst eine Kombination der Ergebnisse verschiedener Prüfungsformate (für theoretisches Wissen, praktische Fertigkeiten usw.: "Triangulation") eine hohe Validität sicher stellen kann (vgl. z. B. [1]).

Die empirische Überprüfung der Validität kann nicht Aufgabe der einzelnen Prüfungsersteller sein, diese muss einer fundierten Prüfungsforschung vorbehalten bleiben. Dem Prüfer obliegt es, durch eine sorgfältige Aufgabenauswahl (**Vermeidung von Konstruktunterrepräsentation** durch eine hinreichend breite und repräsen-

Artikel online frei zugänglich unter http://www.egms.de/en/journals/zma/2006-23/zma000272.shtml
Bitte zitieren als: Möltner A, Schellberg D, Jünger J. Grundlegende quantitative Analysen medizinischer Prüfungen. GMS Z Med Ausbild. 2006;23(3):Doc53.

tative inhaltliche Abdeckung des Themengebiets und **Vermeidung von konstruktirrelevanter Varianz** durch Klarheit der Aufgabenstellungen) und der **Sicherung von Objektivität und Reliabilität** die wesentlichen Voraussetzungen einer validen Prüfung zu schaffen (vgl. [3], [5]).

3. Reliabilität nicht-homogener Tests

Die Ausführungen zur Reliabilität basieren auf mathematisch relativ einfachen Annahmen, die jedoch bei Prüfungen in der Medizin nicht erfüllt zu sein brauchen. Insbesondere ist die Homogenität der Prüfung kaum anzunehmen (d. i. die Annahme, dass die abzuprüfende Fertigkeit nur eindimensional ist, und sich nicht in Unterdimensionen aufspalten lässt - wie z .B. bei der Intelligenz in sprachliche, logische und weitere Intelligenzfaktoren).

Bei nicht-homogenen Tests kann die Abschätzung der Reliabilität mittels Cronbachs α zu einer deutlichen Unterschätzung führen. Viele Befürworter von praktischen Prüfungen führen deren höhere Inhaltsvalidität ins Feld und stufen die Bedeutung der Reliabilität deshalb herab, dabei ist jedoch zu fragen, ob sie tatsächlich die Reliabilität meinen oder Cronbachs α, also die interne Konsistenz. Verwechselt werden beide Begriffe deshalb, weil in der Mehrzahl aller praxisorientierten Veröffentlichungen α mit der Reliabilität gleichgesetzt wird, was nur unter bestimmten Voraussetzungen zutrifft, im Allgemeinen jedoch definitiv falsch ist.

Auswertungsverfahren, die die Nicht-Homogenität von Prüfungen berücksichtigen, sind z. B. faktorenanalytische Verfahren (vgl. z. B. [13]). Die Anwendung dieser Methoden erfordert jedoch eine hinreichende statistische oder biometrische Fachkompetenz und ist mit einem nicht unerheblichen Zeitaufwand verbunden, ein einfaches "How-to-do", wie es die vorliegende Arbeit für die basale Prüfungsauswertung darstellt, kann beim augenblicklichen Kenntnisstand über die Struktur medizinischer Prüfungen nicht erstellt werden.

Fazit

Zum Abschluss sollte noch einmal betont werden, dass die beschriebenen Auswertungen keinen Selbstzweck darstellen, sondern die objektive und empirisch fundierte Basis einer kontinuierlichen Entwicklung und Verbesserung medizinischer Prüfungen darstellen, die immer in engem Zusammenhang mit der Festlegung und der Vermittlung von Lehrinhalten stehen muss.

Eine wesentliche Aufgabe für die Zukunft ist die Etablierung einer medizinischen Prüfungsforschung, die natürlich methodisch über die dargestellten Verfahren hinausgehen muss und in Zusammenarbeit mit Einzelprüfern und Prüfungsverbünden eine Sicherung der Validität medizinischer Prüfungen zum Ziel hat.

Danksagung

Wir danken allen Kolleginnen und Kollegen des Kompetenzzentrums für Prüfungen in der Medizin sowie Herrn Prof. H. Geiss für die konstruktive Hilfe bei der Erstellung des Manuskripts.

Korrespondenzadresse:

• Dr. phil. Andreas Möltner, Ruprecht-Karls-Universität Heidelberg, Kompetenzzentrum für Prüfungen in der Medizin - Baden-Württemberg, Im Neuenheimer Feld 346, 69120 Heidelberg, Deutschland, Tel.: 06221/56-8249, Fax.: 06221/56-7175 andreas.moeltner@med.uni-heidelberg.de

Literatur:

[1] Auewarakul C, Downing SM, Jaturatamrong U, Praditsuwan R. Sources of validity evidence for an internal medicine student evaluation system: an evaluative study of assessment methods. Med Educ. 2005;39:276-283.

[2] Bloch R, Hofer D, Krebs R, Schläppi P, Weis S, Westkämper R. Kompetent prüfen. Handbuch zur Planung, Durchführung und Auswertung von Facharztprüfungen. Medizinische Fakultät Universität Bern, Bern/Wien: Institut für Aus-, Weiter- und Fortbildung; 1999.

[3] Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. Med Educ. 2004;38:327-333.

[4] Downing SM. Reliability: on the reproducibility of assessment data. Med Educ. 2004;38:1006-1012.

[5] Downing SM. Threats to the validity of clinical teaching assessments: what about rater error? Med Educ. 2005;39:353-355.

[6] Downing SM. Validity: On the meaningful interpretation of assessment data. Med Educ. 2003;37:830-837.

[7] Ebel RL, Frisbie DA. Essentials of educational measurement. Englewood Cliffs, N.J.: Prentice Hall; 1991.

[8] Grotjahn R. Testtheorie: Grundzüge und Anwendungen in der Praxis. In: Wolff A, Tanzer H (Hrsg.), Materialien Deutsch als Fremdsprache (Bd. 53). Regensburg: FaDaF. 2000:S.304-341.

[9] Hutchinson L, Aitken P, Hayes T. Are medical postgraduate certification processes valid? A systematic review of the published evidence. Med Educ. 2002;36:73-91.

[10] Kelley TL. The selection of upper and lower groups for the validation of test items. J Educ Psych. 1939;30:17-24.

[11] Lienert GA, Raatz A. Testaufbau und Testanalyse. Weinheim: Beltz; 1994.

[12] Lord FM, Novick MR. Statistical theory of mental test scores. Reading: Addison-Wesley; 1968.

[13] Lucke JF. The a and of congeneric test theorie: An extension of reliability and internal consistency to heterogeneous tests. Appl Psych Meas. 2005;29:65-81.

[14] Lynch DC, Surdyk PM, Eiser AR. Assessing professionalism: a review of the literature. Med Teach. 2004;26:366-373.

[15] Messick S. Validity. In: Linn RL (Hrsg.) Educational Measurement. New York: American Council of Education/McMillan Publishing Company. 1989:S13-103.

[16] Nikendei C, Jünger J. OSCE - praktische Tipps zur Implementierung einer klinisch-praktischen Prüfung. GMS Z Med Ausbild. 2006;23(3):Doc.46.

[17] Rousseeuw PJ, Leroy AM. Robust Regression and Outlier Detection. New York: John Wiley & Sons Inc; 1987.

[18] Schulze J, Drolshagen S, Nürnberger F, Siegers CP, Syed ALI S. Prüfen und Prüfungen nach der neuen Approbationsordnung - Grundsätze und Rahmenbedingungen. Med Ausbild. 2004;21:30-34.

[19] Van der Vleuten C, Schuwirth L. Assessing professional competence: from methods to programmes. Med Educ. 2005;39:309-317.

[20] Veloski JJ, Fields SK, Boex JR, Blank LL (2005). Measuring professionalism: a review of studies with instruments reported in the literature between 1982 and 2002. Acad Med. 2005;80:366-370.

[21] Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. Lancet. 2001;357:945-949.

Artikel online frei zugänglich unter http://www.egms.de/en/journals/zma/2006-23/zma000272.shtml
Bitte zitieren als: Möltner A, Schellberg D, Jünger J. Grundlegende quantitative Analysen medizinischer Prüfungen. GMS Z Med Ausbild. 2006;23(3):Doc53.
gQp020060705 31