Erfahrungen mit dem Fragentyp „k aus n" in Multiple-Choice-Klausuren

R. Blasberg[1], U. Güngerich, W. Müller Esterl, D. Neumann[2], S. Schappel

[1] Institut für Physiologische Chemie und Pathobiochemie (IPCP) der Johannes-Gutenberg-Universität Mainz

[2] Institut für medizinische und pharmazeutische Prüfungsfragen (IMPP) Mainz

Zusammenfassung: In einem gemeinsamen Projekt des Instituts für physiologische Chemie in Mainz und des Instituts für Medizinische und Pharmazeutische Prüfungsfragen wurde ein neuer Typ von Multiple-Choice-Fragen erprobt. Der bisher verwendete Typ der Aussagenkombination enthält zu einer Frage mehrere Aussagen und eine Reihe von Kombinationen dieser Aussagen; genau eine der vorgebenen Kombinationen ist die richtige Antwort. Der hier untersuchte Typ „k aus n" enthält hingegen nur mehrere Aussagen (n = 5...7) zu einer Frage; jede der richtigen Aussagen (k = 2...4) ist einzeln als zutreffend anzukreuzen. In einer Pflichtklausur (274 Teilnehmer) zum Praktikum Biochemie im Sommersemester 1998 wurden die Ergebnisse, also die erzielten Punktzahlen dieser beiden Fragentypen verglichen. Zwei Versionen der Klausur enthielten jeweils 20 übereinstimmende Aufgaben anderen Fragentyps, 15 Aufgaben vom herkömmlichen Typ der Auswahlkombination und 15 vom Typ „k aus n". Bei letzteren war die Zahl der zutreffenden Aussagen jeweils angegeben. Die 15 Aufgaben des Typs „k aus n" der ersten Version der Klausur entsprachen inhaltlich den 15 Aufgaben des Typs mit Aussagenkombination der zweiten Version und vice versa. Wenn die Aufgaben des Typs „k aus n" nur als insgesamt richtig gelöst bewertet wurden, fielen die erzielten Punktzahlen deutlich niedriger aus als für die korrespondierenden Aufgaben mit Aussagekombinationen. Einige zum Vergleich herangezogene Bewertungsmodi, bei denen für partiell richtige Beantwortung Teilpunkte vergeben wurden, führten zu deutlich höheren Punktzahlen.

Experiences with Item Type „k from n" in Multiple-choice Tests: In cooperation between the institute for medical biochemistry, university of Mainz, and the German Central Institute for Medical Examinations a new type of multiple-choice-questions was tested. The common used type of combined questions contains several statements concerning to a question and a number of combinations of these statements. Exactly one of the given combinations is the correct answer. The tested type „k of n" contains only several statements (n = 5...7) concerning a question. Each (k = 2...4) of the correct statements has to be marked seperately. In an examination paper to the biochemical course (274 students) in summer 1998 the results of both kinds of questions were compared. Each of two versions of this paper consisted of 20 identical questions of other types, 15 questions combined type and 15 questions type „k of n". The number of correct answers was given for each of the questions type „k of n". The 15 question type „k of n" in the first version of the paper had the same biochemical content as the 15 questions combined type in the second version and vice versa. When the 15 questions type „k of n" were rated only as in total correct answered, the ratings were lower than the ratings for the corresponding combined questions. Some tested modi of ratings with fraction points for partially correct answered questions resulted in higher numbers of points.

Key words: Multiple-choice tests – Multiple true/false item format – Scoring – Evaluation

Einleitung

Prüfungen im Studium allgemein, Klausuren im Besonderen, haben die Aufgabe, zu überprüfen, ob die erforderlichen Kenntnisse, Fertigkeiten und Fähigkeiten in ausreichendem Maße vorhanden sind. Die Form der Prüfungen hat sich an dem Inhalt des Geprüften zu orientieren. Die Wahl des Aufgabenformats sollte im Idealfall dem geprüften Inhalt angepasst sein. Die Ärztliche Approbationsordnung (ÄAppO) setzt dem Gestaltungsspielraum enge Grenzen, denn vorgegeben sind nicht nur die Art der Aufgaben, nämlich multiple choice, sondern auch ihre Bewertung: richtig oder falsch. Nicht ganz eindeutig ist die ÄAppO in Bezug auf die Anzahl zutreffender Antworten. Das Institut für Medizinische und Pharmazeutische Prüfungsfragen (IMPP) hat sie bisher so interpretiert, dass nur eine der vorgegebenen Antworten zutreffen darf. Das führte, da viele wichtige Fragestellungen mehr als eine zutreffende Antwort haben, z.B. eine Frage nach diagnostischen Maßnahmen bei vorgegebener Symptomatik, zur Einführung des Aufgabentyps „Aufgaben mit Aussagenkombinationen", hier als Typ „*D*" bezeichnet. Es handelt sich dabei um eine Form des MTF-Formats, die durch Vorgabe von Aussagenkombinationen, von denen genau eine zutrifft, an die Vorgaben der ÄAppO angepasst wurde.

Von Teilnehmern an den schriftlichen Prüfungen nach der ÄAppO wurde häufig Unbehagen über die Aufgaben mit Aussagenkombinationen geäußert; hinzu kommt Kritik aus psycho-

Med Ausbild 2001; 18: 73–76

ISSN 0176-4772

metrischen Fachkreisen. Erfahrungen des IMPP zeigen in Übereinstimmung mit der Fachliteratur, dass Aufgaben dieses Formats im Vergleich zu Einfachauswahlaufgaben nicht nur im Mittel schwerer sind, sondern, und das wiegt viel schwerer, auch weniger genau messen. Als Erklärung dafür wird angegeben, dass die Vorgabe einer eingeschränkten Anzahl von Antwortkombinationen Lösungshinweise gibt, von denen leistungsschwächere Prüflinge mehr profitieren als leistungsstärkere. Prüflinge, deren Kenntnisse nicht ausreichen, jede einzelne Aussage als richtig oder falsch zu erkennen, können ihr Teilwissen einsetzen, um die angebotenen Aussagenkombinationen zu prüfen. Die Vorgabe einer eingeschränkten Anzahl von Kombinationen nützt somit eher den Prüflingen mit Teilwissen als denen mit vollständigem Wissen, was zu geringerer Trennschärfe und damit geringerer Zuverlässigkeit führt.

Im Hinblick auf eine geplante Novellierung der ÄAppO hat das IMPP Projekte initiiert, in denen neue Aufgabenformen erprobt werden. Der Direktor des IMPP, Herr Professor Boelcke, hat hierüber anlässlich des Medizinischen Fakultätentages 1998 in Frankfurt berichtet. Das IMPP möchte die gewonnenen Erfahrungen nutzen, um in künftigen Prüfungen die Validität weiter verbessern zu können. In einem gemeinsamen Projekt des Instituts für Physiologische Chemie und Pathobiochemie der Johannes Gutenberg-Universität Mainz und des IMPP wurde der Aufgabentyp „k aus n" erprobt. Parallel dazu wurden Projekte im Fach Anatomie in Gießen, Heidelberg und Mainz durchgeführt. Das prinzipielle Vorgehen wird anhand des erstgenannten Projektes nachfolgend beschrieben.

Der Aufgabentyp „k aus n" unterscheidet sich von dem sonst verwendeten Typ der „Aussagenkombinationen" dadurch, dass die zutreffenden Aussagen einzeln zu markieren sind, die Antworten also nicht aus vorgegebenen Aussagenkombinationen bestehen, von denen genau eine Kombination richtig ist.

Methode

In einer Pflichtklausur (274 Teilnehmer) zum Praktikum Biochemie im Sommersemester 1998 wurden die Ergebnisse, also die von den Teilnehmern erzielten Punktzahlen dieser beiden Fragentypen verglichen. Die Klausur wurde in zwei Versionen aus je 50 Aufgaben zusammengestellt; hiervon waren 20 Aufgaben in beiden Versionen identisch vom Typ „Einfachauswahl" und vom Typ „Verknüpfungsfrage". Die übrigen 30 Aufgaben verteilten sich auf je 15 vom Typ „Aussagenkombinationen" (Typ „*D*") und vom Typ „k aus n", letztere mit bis zu acht Aussagen. In der Fragestellung war jeweils angegeben, wie viele (k = 2...4) der Aussagen zutreffen. Die Aufgaben des Typs „k aus n" der zweiten Version entsprachen inhaltlich den Aufgaben des Typs „*D*" der ersten Version und vice versa. Somit enthielt jede Version der Klausur insgesamt 35 Fragen, bei denen jeweils *nur eine* Antwort zutrifft und 15 Fragen des Typs „k aus n", bei denen jeweils *mehr als eine* Antwort zutrifft. Die Studierenden wurden über die unterschiedlichen Fragentypen anhand eines Beispieles ausführlich instruiert.

Beispielaufgabe (aus der Klausur), Formulierung als Typ „*D*":

Welche Aussagen zur Transkription von DNA in RNA sind richtig?
1. Die Transkription von Eukaryonten findet im Zytoplasma statt.
2. Ein Startsignal für die Transkription ist eine AATAAA-Box.
3. Das entscheidende Enzym ist eine RNA-abhängige RNA-Polymerase.
4. An der Termination der Transkription in Prokaryonten ist eine selbstkomplementäre Haarnadelschleife beteiligt.
5. Amanitin und Actinomycin D sind Inhibitoren der Transkription.
   a) nur 1 und 3 sind richtig
   b) nur 2 und 3 sind richtig
   c) nur 3 und 4 sind richtig
   d) nur 3 und 5 sind richtig
   e) nur 4 und 5 sind richtig.

Dieselbe Aufgabe als Typ „k aus n"

Zwei Aussagen zur Transkription von DNA in RNA sind richtig:
a) Die Transkription von Eukaryonten findet im Zytoplasma statt.
b) Ein Startsignal für die Transkription ist eine AATAAA-Box.
c) Das entscheidende Enzym ist eine RNA-abhängige RNA-Polymerase.
d) An der Termination der Transkription in Prokaryonten ist eine selbstkomplementäre Haarnadelschleife beteiligt.
e) Amanitin und Actinomycin D sind Inhibitoren der Transkription.

In den Klausurbogen war deutlich angegeben, welche Fragen zu Typ „*D*" (nur *eine* von fünf Antworten trifft zu) und welche zum Typ „k aus n" (*bis zu vier* Antworten waren anzukreuzen) gehören. Die beiden Fragentypen wurden außerdem auf Papier unterschiedlicher Farbe gedruckt.

Ergebnisse

Im ersten Schritt wurden die Aufgaben vom Typ „k aus n" gemäß ÄAppO als insgesamt richtig oder falsch bewertet, d. h. keine Teilpunkte vergeben. Die Ergebnisse der Klausur im Fach Biochemie in Mainz sind in Tab. **1** dargestellt.

Tab. 1 Schwierigkeitsgrade der Klausur Biochemie nach Aufgabentyp und Version.

Aufgabentyp	Version A (n = 138)	Version B (n = 136)
„*D*"	54	46
„k aus n"	35	41

In beiden Versionen sind die Aufgaben des Typs „k aus n" im Mittel schwerer als die vom Typ „*D*", allerdings ausgeprägter in Version A. Interessanter ist der Überkreuzvergleich der Ergebnisse, da die Aufgaben vom Typ „k aus n" in Version B inhaltlich den Aufgaben des Typs „*D*" in Version A und vice versa entsprechen. Ein direkter Schwierigkeitsvergleich ist möglich, da die Analyse der in beiden Versionen identischen Aufgaben nahezu dieselben Ergebnisse für beide Versionen zeigt (hier nicht aufgeführt). Danach wurden die Aufgaben des Typs „*D*" in Version A nach Umwandlung in den Typ „k aus n" im Mittel um 13 Prozentpunkte schwerer (54 – 41), die Aufgaben des Typs „*D*" in Version B nach Umwandlung um 11 Prozentpunkte schwerer (46 – 11).

Tab. 2 Schwierigkeitsgrade des Tests Anatomie nach Aufgabentyp, Hochschule und Version.

Aufgabentyp	Gießen Version A (n = 172)	Version B (n = 163)	Heidelberg Version A (n = 106)	Version B (n = 103)	Mainz Version A (n = 216)	Version B (n = 211)
„*D*"	37	34	49	49	58	54
„k aus n"	18	21	26	27	36	40

Für einen Test in Anatomie mit 19 Aufgaben des Typs „*D*" und der gleichen Anzahl inhaltlich gleicher Aufgaben des Typs „k aus n" in derselben Versuchsanordnung, der in Zusammenarbeit mit den Hochschulen in Gießen, Heidelberg und Mainz durchgeführt wurde, sehen die Ergebnisse ähnlich aus, wie Tab. **2** zu entnehmen ist.

Der Überkreuzvergleich der Versionen zeigt eine Abnahme des Schwierigkeitsgrades inhaltlich gleicher Aufgaben von Typ „*D*" zu Typ „k aus n" zwischen 16 (Gießen) und 23 Prozentpunkten (Heidelberg). Die Höhe der Abnahme liegt noch über den für die Biochemieklausur gefundenen Werten und bestätigt damit die Aussage, dass *Aufgaben des Typs „k aus n" erheblich schwerer sind als Aufgaben des Typs „D" gleichen Inhalts, wenn sie ohne Teilpunktvergabe als richtig oder falsch bewertet werden.*

Im zweiten Schritt wurden die Aufgaben des Typs „k aus n" mit vier ausgewählten Modi mit Teilpunktvergabe bewertet. (Diese Art der Bewertungen ist mit der derzeitigen Fassung der ÄAppO nicht vereinbar.) Folgende Bewertungsmodi wurden verwendet:

1. Für jede richtig erkannte zutreffende Antwort werden 1/k Punkte vergeben, wenn k die Anzahl richtiger Antworten ist. Die Punktzahl der Aufgabe ist die Summe der Teilpunkte über die k Aussagen. Wählt der Prüfling mehr als k Aussagen, wird die Aufgabe mit 0 Punkten gewertet, weniger als k Antworten sind zulässig.
2. Für jede zutreffende Entscheidung werden 1/n Punkte vergeben, wenn n die Anzahl der vorgegebenen Aussagen ist. Eine zutreffende Entscheidung liegt vor, wenn eine richtige Aussage gewählt oder eine falsche Aussage nicht gewählt wird. Die Punktzahl der Aufgabe ist die Summe der Teilpunkte über die n Aussagen. Werden mehr als k Aussagen ausgewählt, wird die Aufgabe mit 0 Punkten bewertet. Weniger als k Antworten des Prüflings sind zulässig.
3. Die Bewertung entspricht der in 2. genannten. Zusätzlich wird die berechnete Punktzahl dichotomisiert. Ist die erzielte Teilpunktzahl größer als $^1/_2$, wird ein Punkt vergeben, sonst 0 Punkte.
4. Für jede zutreffende Entscheidung werden 1/n Punkte vergeben, für jede falsche Entscheidung werden 1/n Punkte abgezogen. Die Punktzahl der Aufgabe ergibt sich als Summe der Teilpunkte über jede Aussage. Ist diese negativ, wird sie auf 0 gesetzt.

Tab. **3** zeigt sowohl für die Biochemieklausur, als auch für die Tests in Anatomie die damit erzielten Ergebnisse.

Ein Vergleich mit den jeweils letzten Zeilen der Tab. **2** und **3** zeigt, dass die Ergebnisse der Aufgaben des Typs „k aus n" für alle vier Bewertungsmodi mit Teilpunktvergabe höher sind als bei Bewertung als richtig oder falsch. Mit Ausnahme des Bewertungsmodus 4 (Teilpunktabzug bei falschen Entscheidungen), sind die Ergebnisse auch höher als bei den inhaltlich gleichen Aufgaben des Typs „*D*".

Tab. 3 Schwierigkeitsgrade der Aufgaben des Typs „k aus n" bei Bewertung mit Teilpunktvergabe.

Bewertungsmodus (s. Text)	Biochemie		Anatomie					
	Mainz		Heidelberg		Mainz		Gießen	
	A	B	A	B	A	B	A	B
1	72	71	54	61	64	73	47	59
2	70	72	62	64	71	75	57	61
3	80	85	73	70	83	83	66	65
4	47	52	41	43	50	56	33	37

Im dritten Schritt der Analyse wurde der Einfluss des Aufgabentyps und des Bewertungsmodus auf die Messgenauigkeit (Zuverlässigkeit) der Tests untersucht. Als Indikator der Messgenauigkeit wurde der Reliabilitätskoeffizient (Cronbach α) verwendet. Tab. **4** enthält die Reliabilitätskoeffizienten für alle Tests und Bewertungsmodi.

Bei der Bewertung ohne Teilpunktvergabe sind die Reliabilitätskoeffizienten der Teiltests mit Aufgaben des Typs „k aus n" höher als diejenigen der inhaltsgleichen Teiltests mit Aufgaben des Typs „*D*" (Überkreuzvergleich), d.h. *mit den Aufgaben des Typs „k aus n" wird zuverlässiger gemessen als mit den Aufgaben des Typs „D"*. Mit Ausnahme von Bewertungsmodus 3 sind die Reliabilitätskoeffizienten der Teiltests mit Aufgaben des Typs „k aus n" bei den Bewertungen mit Teilpunktvergabe nochmals zum Teil deutlich höher. Die zuverlässigsten Ergebnisse werden mit den Bewertungsmodi 1 (Anteil gewählter an den richtigen Aussagen) und 2 (Anteil richtiger Entscheidungen) erzielt.

Diskussion

Nach den vorliegenden Ergebnissen erscheint eine Verwendung des Aufgabentyps „k aus n" dann sinnvoll, wenn eine Bewertung der Aufgaben mit Teilpunktvergabe möglich ist. Dies lässt die ÄAppO nicht zu, deshalb scheidet ein Einsatz dieses Aufgabentyps in den bundeseinheitlichen Prüfungen gegenwärtig aus. Seinem Einsatz in Klausuren steht nichts entgegen, wobei die Konstruktion von Aufgaben „k aus n" allerdings eine Abstimmung von Aufgabeninhalt und Bewertungsmodus voraussetzt und damit erhöhte Anforderungen an die Aufgabenautoren stellt. Zu beachten ist außerdem, dass die Einzelaussagen wie beim Typ „*D*" in jedem Fall eindeutig entscheidbar richtig oder falsch sind; dies ist schwierig im klinischen Bereich.

Tab. 4 Reliabilitätskoeffizienen (Cronbach α) nach Aufgabentyp, Bewertungsmodus und Version.

Aufgabentyp	Bewertungsmodus (s. Text)	Biochemie		Anatomie					
		Mainz		Heidelberg		Mainz		Gießen	
		A	B	A	B	A	B	A	B
		0-1-Bewertung (keine Teilpunktvergabe)							
„*D*"	ÄAppO	0,51	0,52	0,64	0,65	0,65	0,60	0,59	0,42
„k aus n"	ÄAppO	0,59	0,69	0,63	0,74	0,74	0,74	0,59	0,71
		Teilpunktvergabe							
„k aus n"	1	0,76	0,76	0,81	0,84	0,77	0,80	0,76	0,83
„k aus n"	2	0,69	0,76	0,84	0,83	0,77	0,80	0,78	0,80
„k aus n"	3	0,61	0,62	0,79	0,71	0,58	0,62	0,68	0,66
„k aus n"	4	0,64	0,72	0,73	0,80	0,78	0,78	0,66	0,77

Bei einer Bewertung mit Teilpunktvergabe kann die Entscheidung, ob ein Prüfling die Prüfung bestanden hat oder nicht, zu einer schmalen Gratwanderung werden. Im Extremfall entscheidet ein Unterschied von Bruchteilen eines Punktes über den Ausgang der Prüfung. Eine Feststellung, Kandidat 1 habe mit 143,3 Punkten die Prüfung bestanden, während Kandidat 2 mit 142,2 Punkten durchgefallen ist, ist vernünftig nicht zu plausibilisieren und dürfte in einem Rechtsstreit kaum durchsetzbar sein.

Hinzuweisen ist noch auf technische Aspekte der Testauswertung. Um zu gewährleisten, dass die vom Belegleser erkannten Markierungen mit den Markierungen auf den Antwortbogen übereinstimmen, war ein erheblicher Kontrollaufwand erforderlich. Da beim konventionellen Aufgabentyp „*D*" für jede Aufgabe genau eine Antwort zu markieren ist, konnte sich die Kontrolle auf diejenigen Aufgaben beschränken, bei denen der Belegleser keine oder mehr als eine Markierung erkannt hatte. Bei Fragen des Typs „k aus n" mussten bei der Biochemieklausur praktisch alle Belege visuell nachkontrolliert werden. Tatsächlich ergab die Nachkontrolle bei etwa 6% der Belege Unstimmigkeiten zwischen Markierungen auf dem Beleg und vom Belegleser erkannten Markierungen. In einem Fall war daraufhin die Entscheidung über die Scheinvergabe in Biochemie zugunsten des Teilnehmers zu revidieren. Der extrem hohe Kontrollaufwand für Aufgaben des Typs „k aus n" könnte durch Umgestaltung des Antwortbeleges auf ein vertretbares Maß reduziert werden. Erforderlich wären für jede Aussage zwei Markierungen für richtig oder falsch, die Prüfungsteilnehmer müssten also für jede Aussage die Entscheidung über richtig oder falsch im Beleg markieren.

Rolf Blasberg

Institut für Physiologische Chemie und Pathobiochemie (IPCP)
der Johannes-Gutenberg-Universität
55099 Mainz

E-mail: blasberg@mail.uni-mainz.de