UNIVERSITÄT
BERN

Medizinische Fakultät

Institut für Medizinische Lehre IML

Abteilung für Assessment und Evaluation AAE

Anleitung zur Herstellung von MC-Fragen und MC-Prüfungen für die ärztliche Ausbildung

René Krebs

Kontakt:
rene.krebs@iml.unibe.ch

Download:
www.iml.unibe.ch

INHALT

	Was lernen Sie in dieser Anleitung?	1
1.	Welche Anforderungen sollen MC-Items erfüllen?	2
2.	Worauf zielt das Fragenthema?	4
3.	Wozu eignet sich welcher Fragetyp?	5
3.1	Positive Einfachwahl aus fünf Wahlantworten (Typ A)	6
3.2	Negative Einfachwahl aus fünf Wahlantworten (Typ Aneg)	9
3.3	Zuordnung (Typ B)	10
3.4	Erweiterte Zuordnung (Typ R)	11
3.5	Wahl einer angegebenen Zahl bester Antworten (Typ PickN)	12
3.6	Vierfache Entscheidung richtig/falsch (Typ Kprim/K')	13
3.7	Kausale Verknüpfung (Typ E)	15
4.	Wie werden Fragen formuliert	16
4.1	Was ist bei der Formulierung des Stammes zu beachten?	17
4.2	Was ist bei der Formulierung der Antworten zu beachten?	17
	UTC-Kurztest	19
5.	Wie werden Fragen überprüft?	22
6.	Wie wird eine Prüfung zusammengestellt?	23
	Verwendete Quellen; Literatur	24
	Anhang 1: Grundlagen für die Herstellung standardisierter Fragen	25
	Anhang 2: Fragenformular	26
	Anhang 3: Allgemeine Hinweise zur korrekten Beantwortung	27
	Anhang 4: Instruktionen zur Beantwortung der einzelnen Typen	28
	Glossar prüfungstechnischer Fachbegriffe	29

Bern, Juli 2004

Was lernen Sie in dieser Anleitung?

Sie erfahren,

- **wie relevante Itemthemen* bestimmt werden**
- **wofür sich welcher Itemtyp eignet**
- **worauf bei der Itemformulierung zu achten ist**
- **wie neu verfasste Items zu überprüfen sind**
- **was bei der Prüfungszusammenstellung zu beachten ist**

Wenn Sie bei der Herstellung von MC-Items nach dieser Anleitung vorgehen, sollten diese in der Regel mehr prüfen als isoliertes Faktenwissen und weitgehend frei sein von formalen Fehlern.

Aber auch hier gilt: Übung macht den Meister. Es ist deshalb zu empfehlen, dass neue Fragenautoren in einem mindestens eintägigen Workshop das Herstellen und Revidieren von MC-Items gemeinsam erarbeiten. Zudem sollten Autoren kontinuierlich über das Ergebnis der Revision ihrer Fragen informiert werden, damit sie aus Fehlern lernen können.
Ferner gilt: Gut Ding will Weile haben. Auch geübte Autoren rechnen durchschnittlich mit einer Stunde Arbeit, um ein MC-Item herzustellen.

Abschnitt 1. beschreibt, worauf zu achten ist, wenn MC-Items zur Gültigkeit (Validität) und Zuverlässigkeit (Reliabilität) einer Prüfung beitragen sollen. Die Abschnitte 2 - 5 bilden den Hauptteil und beschäftigen sich mit der konkreten Entwicklung der Items: Dabei geht es zuerst um die Wahl des Itemthemas und des geeigneten Itemtyps, wobei auch schon besondere itemspezifische Anforderungen genannt werden. Es folgen allgemeine Hinweise zur korrekten Formulierung und zur Vermeidung ungewollter Lösungshinweise (sog. Cues). Schliesslich wird der Revisionsprozess beleuchtet. Der 6. Abschnitt beschreibt, worauf zu achten ist, wenn Items zu einer Prüfung zusammengestellt werden.

* Ein Glossar zu allen prüfungstechnischen Fachbegriffen, die in dieser Anleitung verwendet werden, finden Sie auf den Seiten 29/30.

1. Welche Anforderungen sollen MC-Items erfüllen?

Die MC-Prüfung als Instrument zur Qualitätssicherung

Wir gehen davon aus, dass mit der geplanten MC-Prüfung ein wichtiger Qualitätsaspekt der Ausbildungsabsolventen kontrolliert werden soll, nämlich die Verfügbarkeit relevanten Fachwissens und die Kompetenz, dieses in konkreten Problemsituationen anzuwenden.

Vorbedingung: Inhaltsdefinition und -gewichtung

Vorbedingungen, dass diese Kontrolle in der Prüfung auf repräsentative Weise erfolgen kann, sind die möglichst präzise Definition des relevanten Fachwissens, idealerweise in Form eines Lernzielkataloges, und die Erstellung eines Blueprints, d.h. eines gewichteten Verzeichnisses des Prüfungsstoffes.

Hauptbedingung: gute Items

Letztlich steht oder fällt die Validität und Reliabilität aber mit der Qualität der einzelnen Items. Items sind mit grosser Wahrscheinlichkeit für den Einsatz in einer Prüfung geeignet, wenn sie nachfolgend aufgeführten Kriterien genügen:

So messen Items gültig ...

Ein Item trägt dann zur Validität der Prüfung bei,

- wenn das gewählte **Fragenthema relevant** ist **im Hinblick auf künftige Anforderungen**, sei es in der weiteren Ausbildung oder der praktischen beruflichen Tätigkeit

  Spitzfindigkeiten sind ebenso zu vermeiden wie Trivialitäten. Mögliche Relevanzkriterien s. Seite 4

- wenn das **Anspruchniveau stimmt**

  Aus Items, die lediglich die Reaktivierung auswendig gelernter Einzelfakten erfordern, entsteht keine gültige Prüfung, wenn das Prüfungsziel primär Verständnis und Anwendungskompetenz von Wissen vorsieht. Solange Inhalte so abgefragt werden, wie sie im Lehrbuch oder Skript stehen, prüft man auf der untersten Taxonomiestufe **Kennen.**

  Verstehen kann geprüft werden, wenn die Kandidaten zur Beantwortung der Items

  - Informationen analysieren, verknüpfen, integrieren,
  - Informationen transformieren (z.B. Tabelle, Grafik → Sprache),
  - Zusammenhänge erfassen,
  - Schlussfolgerungen ziehen müssen.

  Anwenden und Beurteilen kann geprüft werden, wenn die Kandidaten zur Beantwortung eines Items

  - erworbenes Wissen auf neue Situationen übertragen (abstrahieren, transferieren, generalisieren),
  - erworbenes Wissen beim Lösen von Problemen anwenden,
  - Informationen (Gegebenheiten, Ergebnisse) beurteilen, bewerten, gewichten und Folgen abschätzen müssen.

- wenn es auf einen klar umschriebenen Inhalt resp. ein Problem **fokussiert** ist und in sich **ein geschlossenes Ganzes** bildet

  Alle Wahlantworten sollen in die gleiche inhaltliche Kategorie fallen. Abwägen zu müssen, ob eine vorgeschlagene Problemursache wahrscheinlicher richtig ist als eine Begriffsdefinition oder eine Massnahme, hat keinen Bezug zur Berufsrealität.

- wenn es **eine eindeutig beste Lösung** gibt

  Inhalte mit kontroversen Lehrmeinungen sind für die MC-Methode ungeeignet oder es werde denn ausdrücklich nach einer bestimmten Lehrmeinung gefragt.

... und so messen Items zuverlässig

Ein Item trägt dann zur **Reliabilität** der Prüfung bei, wenn es im Sinne der Prüfung differenziert, d.h. wenn es Trennschärfe besitzt. Dies ist dann der Fall,

- wenn es bezüglich Schwierigkeit **der Zielgruppe angemesen** ist

  Insbesondere problematisch sind zu schwere Items, bei denen auch gute Kandidaten aufs Raten angewiesen sind.

- wenn es **sprachlich einfach, klar formuliert** ist

  Es soll Wissen geprüft werden und nicht Sprachverständnis oder Interpretationsglück.

- wenn es **keine ungewollten Lösungshinweise** (sog. Cues) enthält

  Es soll Wissen geprüft werden und nicht MC-Prüfungserfahrung und 'Testknackerfähigkeiten'.

Diese drei Punkte sind nicht nur wichtig, um eine gute Reliabilität zu erzielen, sie bilden auch die Voraussetzung, damit die Prüfung gültig messen kann. Der korrekten Formulierung und der Vermeidung von Cues ist das Kapitel 4 gewidmet.

Neben diesen allgemeinen gibt es noch einige **typenspezifische Anforderungen.** Diese werden bei der Erläuterung von Besonderheiten der einzelnen Typen aufgeführt.
Ebenfalls sei hier schon darauf hingewiesen, dass es bezüglich Messqualität **typenabhängige Unterschiede** gibt. Auch darauf wird bei der Beschreibung der Typen eingegangen.

2. Worauf zielt das Fragenthema?

vom Themenbereich zum Itemthema

Blueprintkapitel und Lernziele geben mögliche Themenbereiche an, zu denen Items kreiert werden können. Für ein Fragenthema ist eine weitere Einschränkung auf einen Teilaspekt erforderlich, z.B. auf die (wahrscheinlichste) Ursache eines Problems oder das richtige oder effizienteste Vorgehen zur Ursachenlokalisierung oder die einzig richtige Massnahme resp. das aussichtsreichste Vorgehen zur Problembehebung. Wenn als passende Frage zu einem gewählten Thema nur formuliert werden kann: "Welche der folgenden Aussagen zu XY trifft zu?", ist das Thema mit grosser Wahrscheinlichkeit zu breit, zu heterogen für ein gutes MC-Item.

Kriterium Relevanz

Innerhalb eines Themenbereichs ergeben sich relevante Items vor allem aus Teilaspekten,

- mit denen man in diesem Inhalts- resp. Arbeitsbereich besonders **häufig** konfrontiert ist
- bei denen Fehler **gravierende Folgen** haben können
- bei denen **Fehlmeinungen verbreitet** sind
- die **für das Verständnis späterer Lerninhalte entscheidend** sind

Überprüfen Sie die Relevanz der vorgesehenen Themen mit der Frage: "Wie wichtig ist es, dass ein Ausbildungsabsolvent dieses Problem selbstständig lösen resp. die sich ergebende Frage richtig beantworten kann?"

eigene Erfahrungen als Inspirationquelle

Von selbst erlebten Problemen auszugehen, kann sehr anwendungsbezogene, relevante Items ergeben. Man muss sich aber hüten, 'interessante' Sonderfälle auszuwählen.

Verwendung von Lehrbüchern

Die Verwendung von Lehrbüchern ist wichtig zur Absicherung und Dokumentierung der fachlichen Richtigkeit einer Frage. Lehrbücher können auch hilfreich sein zum Finden guter Distraktoren. Als Inspirationsquelle für Fragenthemen sind sie aber nicht zu empfehlen. Es entstehen daraus sehr gerne im negativen Sinn ‘akademische’ Fragen.

Kriterium Anwendungsbezug

Überprüfen Sie die vorgesehenen Fragenthemen auf ihren Anwendungsbezug mit der Frage: Stellt sich dieses Problem/diese Frage einem Ausbildungsabsolventen in seiner weiteren Ausbildung und/oder praktischen Tätigkeit auf diese Weise?

strukturierte Sammlung von Aussagen zu einem Thema

Manchmal ergibt sich aus einem gewählten Thema ganz spontan eine gute Idee für ein konkretes MC-Item. Häufig ist es sinnvoll, sich nicht zu früh auf einen bestimmten Itemtyp zu versteifen, sondern erst einmal Aussagen zu sammeln. Dazu kann ein Raster nützlich sein, wie das im Anhang 1 (S. 25) gezeigte.

3. Wozu eignet sich welcher Fragetyp?

Beschränkung auf wenige Typen

Seit Beginn der Entwicklung objektiver schriftlicher Prüfungen gegen Ende des 19. Jahrhunderts sind viele verschiedene Itemtypen kreiert worden. Für die MC-Prüfungen des amerikanischen National Board of Medical Examiners wurde ursprünglich eine Auswahl von über zehn Typen vorgeschlagen[1]. Diese wurden mit Buchstaben etikettiert, welche heute noch (auch in dieser Anleitung) verwendet werden. Zunehmend wurde aber erkannt, dass es unter den Aspekten der Herstellung, der Beantwortung und der Messqualität besser ist, sich auf wenige Typen zu beschränken.

Best-Antwort-Typen ...

Von der Aufgabenstellung her lassen sich zwei Grundtypen unterscheiden: Best-Antwort-Typen und Richtig/Falsch-Typen. Bei der Gruppe der **Best-Antwort-Typen** wird die Zahl der auszuwählenden Antworten angegeben; meist ist es nur eine. Die falschen Antworten, sog. Ablenker oder Distraktoren müssen sich nicht schwarz/weiss von der richtigen Antwort oder den richtigen Antworten abheben. Es ist möglich, die Kandidaten eine intellektuell anspruchsvollere und oft auch realitätsnähere Abwägung von Graustufen vornehmen zu lassen.
Folgende Best-Antwort-Typen werden im Weiteren vorgestellt:

Typ	Aufgabe
A(pos)	positive Einfachwahl aus fünf Wahlantworten
Aneg	negative Einfachwahl aus fünf Wahlantworten
B	Zuordnung der richtigen aus fünf Wahlantworten zu mehreren Fragen
R	wie Typ B aber mit bis zu 26 Wahlantworten (A-Z)
PickN	wie Typ A(pos) oder Typ R aber mit mehr als einer auszuwählenden 'besten' Antwort

... vor Richtig/ Falsch-Typen

Bei der Gruppe der **Richtig/Falsch-Typen** müssen die Kandidaten für jede einzelne Antwort oder Aussage eine Ja/Nein-Entscheidung treffen. Es sind deshalb nur Antworten tauglich, bei denen dies eindeutig möglich ist. Richtig/Falsch-Typen bergen damit die Gefahr, reines Faktenwissen zu prüfen.
Folgende Richtig/Falsch-Typen werden im Weiteren vorgestellt:

Typ	Aufgabe
Kprim	vierfache Entscheidung richtig/falsch
E	Beurteilung zweier Aussagen und ihrer kausalen Verknüpfung

☞ **Best-Antwort-Items sollen in einer MC-Prüfung quantitativ klar dominieren.**

3.1 Positive Einfachwahl aus fünf Wahlantworten (Typ A)

Definition

Auf eine Frage oder unvollständige Aussage folgen fünf Wahlantworten oder Ergänzungen, aus welchen die **einzig richtige** oder die **beste** auszuwählen ist.

Der Stamm kann aus einem Satz ...

Itemstämme können aus einem einzigen Fragesatz oder einer Aussage bestehen. Mit solchen Items kann durchaus anwendungsrelevantes Wissen geprüft werden, in der Regel handelt es sich allerdings lediglich um Faktenwissen.

... oder einer Vignette plus Fragestellung bestehen.

Mit MC-Items können und sollen möglichst häufig auch die Fähigkeiten geprüft werden, Informationen zu interpretieren und zu integrieren und theoretische Kenntnisse auf ein konkretes Problem anzuwenden. Dafür sind zweiteilige Fragenstämme erforderlich. In einem ersten längeren Teil wird ein Problem aus dem Berufsfeld beschrieben. Wir sprechen von einer Fall- oder Problemvignette. Davon abgetrennt erfolgt die kurze Fragestellung.

Patientenvignetten sollen einige oder alle der folgenden Elemente in der angegebenen Reihenfolge enthalten:

1. Alter, Geschlecht, evtl. Beruf, Herkunft/Ethnie *(Ein 45-jähriger Wirt ...)*
2. Ort der Behandlung *(... wird in die Notfallabteilung gebracht ...)*
3. Konsultationsgrund *(... wegen starker Kopfschmerzen, ...)*
3. Dauer *(... die seit 2 Tagen andauern.)*
4. Anamnese, evtl. Familienanamnese (relevante Punkte)
5. Status, physische Befunde
6. evtl. Resultate der diagnostischen Untersuchungen
7. evtl. Anfangsbehandlung, Folgebefunde, etc.

Der Informationsteil kann auch Tabellen und Abbildungen beinhalten (Fotografien, Röntgenbilder, Grafiken) oder Zitate aus Artikeln.

richtige Antwort und Ablenker oder Distraktoren

Die fünf Wahlantworten sollen sowohl inhaltlich wie formal möglichst homogen sein und sich gleichartig auf den Stamm beziehen. Die vier Distraktoren müssen nicht völlig falsch sein. Die richtige Antwort muss sich aber eindeutig positiv abheben.

Beispiele

1. Eine 60-jährige Frau hat Schwierigkeiten aus einem Sessel aufzustehen und sich aufzurichten. Sie hat dagegen keine Mühe auf ebenem Boden zu gehen und kann auch ohne weiteres ihre Beine in den Hüftgelenken beugen.

Welcher der folgenden Muskeln ist am wahrscheinlichsten insuffizient?

(A) M. glutaeus maximus
(B) M. glutaeus medius
(C) M. obturatorius externus
(D) M. obturatorius internus
(E) M. psoas major

Anatomie (2. Jahr) Key: A

2. *Ein 65-jähriger, übergewichtiger Mann leidet an zunehmenden Atembeschwerden. Die Spirometrie ergibt folgende Resultate:*
   - *eingeschränkte Lungenvolumina*
   - *normales Verhältnis zwischen Erstsekundenvolumen und Vitalkapazität (FEV1/VK)*
   - *stark verminderte Lungendehnbarkeit*

   Die Blutgasanalyse zeigt eine arterielle Hypoxämie und eine Hypokapnie während der Arbeit.
   Welches ist die wahrscheinlichste Ursache der Beschwerden?

   (A) ein obstruktives Lungenemphysem

   (B) eine diffuse interstitielle Lungenfibrose

   (C) eine gesteigerte Lungendurchblutung wegen Links-Rechts-Shunt

   (D) eine Adipositas

   (E) eine chronische Bronchitis

   Pathophysiologie (3. Jahr): P 81, R 0.20 [*] Key: B

3. *Eine Fall-Kontroll-Studie liefert folgende Resultate:*

	Krankheit +	*Krankheit –*
Exposition +	*23*	*27*
Exposition –	*77*	*173*

   Wie ist in diesem Fall die Odds-Ratio zu berechnen?,

   (A) $\frac{77}{23+77} = 0.77$

   (B) $\frac{173}{27+173} = 0.87$

   (C) $\frac{23}{23+27} = 0.46$

   (D) $\frac{23/27}{77/173} = 1.91$

   (E) $\frac{23/(23+27)}{23/(77+173)} = 5.00$

   z.B. SPM (3. Jahr): P 87, R 0.25 Key: D

4. *Bei einem Bankett bestand das Menu aus gebratenem Huhn, Bratkartoffeln, Erbsen, Schokoladeneclair und Kaffee. Innerhalb von zwei Stunden erkrankten die meisten Teilnehmenden sehr heftig an Übelkeit, Erbrechen und Bauchschmerzen.*
   Welchen der folgenden Organismen wird man bei der Analyse des Essens am wahrscheinlichsten in grosser Zahl finden?

   (A) Escherichia coli

   (B) Proteus mirabilis

   (C) Salmonella typhimurium

   (D) Staphylococcus aureus

   (E) Streptococcus faecalis

   Mikrobiologie (3. Jahr) Key: D

[*] P = % richtige Antworten (Item-Schwierigkeit), R = Item-Trennschärfe (vgl. Glossar S. 29)

5. *Ein 53-jähriger Mann leidet seit 2 Tagen unter zunehmender Dyspnoe und Husten mit eitrigem Auswurf. Seit 30 Jahren raucht er täglich ein Päckchen Zigaretten.*
Seine Temperatur beträgt 37.2 °C. Die Atemgeräusche sind abgeschwächt mit ein wenig Giemen und Pfeifen. Die Leukozytenzahl beträgt 9000/mm³ (9.0 G/l) mit normaler Differenzierung. Die Gramfärbung des Sputums zeigt zahlreiche Neutrophile und gramnegative Diplokokken. Die Thoraxröntgenaufnahme zeigt eine Überblähung.
Welches ist die wahrscheinlichste Diagnose?

   (A) Asthma
   (B) Bronchitis
   (C) Bronchiektasen
   (D) Lungenembolie
   (E) Streptokokkenpneumonie

   Innere Medizin (Schlussprüfung) Key: B

6. *Ein 47-jähriger Patient erscheint nach einem Velosturz auf der Notfallstation. Er ist agitiert und weist einen starken Foetor aethylicus auf. Er klagt über Thoraxschmerzen links basal und über Schmerzen im linken Hypochondrium. Der Glasgow-Koma-Score beträgt 15. Er hat keine erkennbare Wunde. Die Atemfrequenz beträgt 38, Puls 140, arterieller Blutdruck 90/60 mmHg. Welches ist die wahrscheinlichste Erklärung der hämodynamischen Situation?*

   (A) kardiogener Schock
   (B) Hypovolämie
   (C) Lungenembolie
   (D) Folge der Alkoholintoxikation
   (E) vagaler Reflex im Anschluss an den Unfall

   Chirurgie (Schlussprüfung): P 88, R 0.25 Key: B

7. *Ein 64-jähriger Mann liegt im Spital im terminalen Stadium eines Lungenemphysems. Seine Angehörigen stellen fest, dass er sie zwar erkennt, geistig wach und orientiert zu sein scheint, dass aber bei ihren Besuchen sein Interesse und seine Teilnahme nachgelassen haben.*
Welches ist die wahrscheinlichste Erklärung dieses Verhaltens?

   (A) Entwicklung eines Deliriums
   (B) Entwicklung einer psychotischen Störung
   (C) Verschlimmerung einer Persönlichkeitsstörung
   (D) Übersedierung
   (E) Rückzug

   Psychosoziale Medizin (3. Jahr) Key: E

Eignung

Dies ist nach wie vor der Standardtyp der MC-Methode. Er hat sich international unter verschiedenen Aspekten, insbesondere auch unter messtechnischem Gesichtspunkt bestens bewährt und sollte anteilsmässig in jeder MC-Prüfung klar dominieren.

Wichtige Formulierungshinweise

Detaillierte Formulierungshinweise zu diesem Standardtyp finden Sie im Kapitel 4 ab Seite 16.

3.2 Negative Einfachwahl aus fünf Wahlantworten (Typ Aneg)

Definition

Auf eine Frage oder unvollständige Aussage folgen fünf Wahlantworten oder Ergänzungen, aus welchen die **Ausnahme** oder die **am wenigsten zutreffende** auszuwählen ist.

Beispiele

1. *Welches der folgenden Antibiotika darf einem 18 Monate alten Kind mit akuter Otitis media **nicht** verabreicht werden?*
   (A) Amoxycillin
   (B) Cefaclor
   (C) Co-trimoxazol
   (D) Doxycyclin
   (E) Erythromycin
   Pädiatrie (Schlussprüfung): P 81, R 0.25 Key: D

2. *Welche der folgenden Veränderungen **fehlt** bei der Mitralstenose?*
   (A) Hypertrophie des linken Vorhofes
   (B) Hypertrophie der linken Kammer
   (C) Hypertrophie der rechten Kammer
   (D) chronische Blutstauung der Lungen
   (E) Hämosiderinpigment in der Lunge
   Pathologie (3. Jahr) P 87, R 0.30 Key: B

Eignung

Sicher angezeigt ist dieser Itemtyp in den seltenen Fällen, in denen das Kennen einer wichtigen Ausnahme entscheidend ist. Viel häufiger wird er aber anders eingesetzt. Eigentlich soll das Kennen der vier positiven Antworten geprüft werden, die "richtige" Antwort ist bloss das Abfallprodukt der Lösungen. Wenn es dabei um einen schwarz/weiss-Entscheid geht, wäre von der Prüfungsabsicht her ein Kprim-Item logischer (s. 3.6). Fragen Sie sich aber zuerst grundsätzlich, ob nicht das Kennen des wichtigsten, wahrscheinlichsten, gefährlichsten XY relevanter und anwendungsnäher, ein positives A-Item also besser wäre.

Hinsichtlich Schwierigkeit und Trennschärfe schneiden Aneg-Items im Durchschnitt zwar praktisch gleich gut ab wie die positiven. Unter dem Aspekt der Validität sollte eine Prüfung aber nicht viele Aneg-Items enthalten.

Wichtige Formulierungshinweise

- Die Negation muss durch **Fettdruck** oder <u>Unterstreichung</u> hervorgehoben werden.
- Alle Antworten müssen positiv formuliert werden, da sonst doppelte Negationen zu beurteilen sind.
  Aus diesem Grund ist auch 'Keine der genannten Antworten' hier als Wahlantwort nicht zulässig.

3.3 Zuordnung (Typ B)

Definition

Bei Typ B werden eingangs fünf mit den Buchstaben A-E bezeichnete Wahlantworten vorgegeben. Oft eingeleitet durch einen Verbindungssatz, der die Zuordnungsaufgabe präzisiert, folgen zwei bis maximal fünf Fragen oder Aussagen. Jeder davon muss die **einzig richtige** oder die **beste** der Wahlantworten zugeordnet werden, wobei eine Antwort mehr als einmal die richtige sein kann.
Es handelt sich also eigentlich um eine Serie von Fragen des Typs A mit den jeweils gleichen Wahlantworten.

Beispiel

(A) Captopril
(B) Diazoxid
(C) Isosorbiddinitrat
(D) Nifedipin
(E) Propranolol

Welche Substanz hemmt

1. *die Reninfreisetzung?*
   (A) (B) (C) (D) (E) Key: E
2. *den Abbau von Bradykinin?*
   (A) (B) (C) (D) (E) Key: A
3. *die Öffnung von Kalziumkanälen?*
   (A) (B) (C) (D) (E) Key: D

Pharmakologie (3. Jahr): 1. P 61, R 0.40; 2. P 90, R 0.51; 3. P 92, R 0.40

Eignung

In der dargestellten Form ist der Typ B sowohl unter dem Validitäts- wie dem Reliabilitätsaspekt gleich gut geeignet wie der positiv formulierte Typ A. Es sollte von der Breite und Bedeutung des zugrunde liegenden Themas her entschieden werden, ob dafür ein einzelnes A-Item ausreicht, oder ob ein B-Komplex mit mehreren Items angemessener ist. Abzuraten ist von Aufgaben, bei welchen als "Fragen" Definitionen oder Eigenschaftsbeschreibungen vorgelegt werden, denen als "Antwort" der treffende Begriff zugeordnet werden muss. Solche Kreuzworträtsel-Fragen prüfen Begriffs- und Faktenwissen auf tiefster Stufe.

Wichtige Formulierungshinweise

- Inhaltlich heterogene Fragenkombinationen, bei denen von vornherein bei einzelnen Fragen gewisse Antworten ausser Betracht fallen, sind zu vermeiden.
- Wenn einer Frage eine bestimmte Antwort zugeordnet werden kann, sollte sie nicht dadurch als richtige Antwort auf andere Fragen von vornherein ausser Betracht fallen.

  Bei Verletzung dieser beiden Punkte wird die Ratewahrscheinlichkeit erhöht und damit die Messzuverlässigkeit beeinträchtigt.
- In der Regel sollen nicht mehr als drei Fragen zu einem B-Komplex kombiniert werden, um eine inhaltliche Übergewichtungen zu vermeiden und um obigen zwei Gefahren vorzubeugen.

3.4 Erweiterte Zuordnung (Typ R)

Definition

Die Struktur ist grundsätzlich gleich wie die des Typs B, nur kann die Antwortliste bis zu 26 Wahlantworten (A-Z) enthalten.

Beispiel

(A) Adenovirus
(B) Aspergillus fumigatus
(C) Bacillus anthracis
(D) Candida albicans
(E) Chlamydia psittaci
(F) Coccidioides immitis
(G) Coronavirus
(H) Corynebacterium diphteriae
(I) Coxiella burnetii
(J) Coxsackievirus
(K) Epstein-Barr Virus
(L) Haemophilus influenzae
(M) Histoplasma capsulatum
(N) Mycobacterium tuberculosis
(O) Mycoplasma pneumoniae
(P) Neisseria gonorrhoeae
(Q) Neisseria meningitidis
(R) Pneumocystitis carinii
(S) Rhinovirus
(T) Streptococcus pneumoniae
(U) Streptococcus pyogenes (Gruppe A)

Wählen Sie für die folgenden Patienten mit Fieber je den Mikroorganismus, der ihre Krankheit am wahrscheinlichsten verursacht hat.

1. *Ein 7-jähriges Kind hat hohes Fieber und Halsweh. Der Rachen ist gerötet, die rechte Rachenmandel geschwollen, mit schmierigem weisslichem Exsudat, und rechtsseitig besteht eine schmerzhafte submandibuläre Lymphadenopathie. Die Kultur des Rachenabstriches auf Agarplatten ergibt zahlreiche kleine ß¬hämolytische Kolonien, die durch Bacitracin gehemmt werden.* Key: U

2. *Seit einer Woche leidet ein 18-jähriger Patient an Fieber, Halsweh und Unwohlsein. Er hat geschwollene, von Exsudat bedeckte Rachenmandeln sowie eine diffuse zervikale Lymphadenopathie und eine Splenomegalie. Es besteht eine Lymphozytose mit atypischen Lymphozyten. Ein Test für heterophile Antikörper ist positiv.* Key: K

Mikrobiologie: übersetztes Beispiel aus Case, Swanson (1998), p. 77

Eignung

Der Typ hat in diversen Vergleichsstudien mit andern Fragetypen messtechnisch gut abgeschnitten. Es wird erhofft, dass es bei entsprechend langen Antwortlisten für die Prüfungskandidaten unökonomisch wird, diese nach der richtigen Lösung zu durchsuchen, dass sie diese selbst entwickeln und sie dann gezielt in der Liste suchen (=Annäherung an Fragen mit freier Beantwortung).

Wichtige Formulierungshinweise

- Zuerst wird - bezogen auf ein ausgewähltes Thema - die Antwortliste kreiert. Als Thema gut geeignet sind etwa unspezifische Hauptbeschwerden wie Bauchschmerzen, Kopfweh, Rückenschmerzen, Fieber (s. obiges Beispiel), für Antwortlisten z.B. anatomische Strukturen, Diagnosen, diagnostische Untersuchungen, Medikamente. Erst dann werden die Fälle/Fragen entwickelt.
- Die Antworten müssen inhaltlich homogen sein und sollten im Idealfall Wörter/Begriffe sein, in Ausnahmefällen sehr kurze Sätze. Denkbar sind auch bezeichnete Gebiete in einem Bild oder einer Grafik.
- Antworten möglichst alphabetisch oder logisch anordnen.

3.5 Wahl einer angegebenen Zahl bester Antworten (Typ PickN)

Definition

Dieser Typ kann entweder aufgebaut sein wie Typ A(pos) mit verlängerter Antwortliste oder Typ R. Der Hauptunterschied zu diesen Typen ist, dass die Kandidaten aufgefordert werden, eine angegebene Zahl bester Antworten (2 - 5) auszuwählen.

Beispiel

(A) Eisen
(B) Fluorid
(C) Folsäure
(D) Kalzium
(E) Vitamin A
(F) Vitamin B1 (Thiamin)
(G) Vitamin B6
(H) Vitamin B12 (Cyanocobalamin)
(I) Viatamin C
(J) Vitamin D
(K) Vitamin E

Wählen Sie für jedes Kind die angemessenen Vitamin- oder Mineralzusätze.

1. *Ein 1-monatiger Säugling wird dem Arzt zur Kontrolluntersuchung gebracht. Er wird ausschliesslich gestillt und bei der Untersuchung liegen normale Befunde vor. (Wählen Sie 2 Zusätze)* Key: B,J
2. *Ein 6-jähriges Mädchen leidet unter zystischer Fibrose. Es nimmt keine Medikamente ein. (Wählen Sie 3 Zusätze)* Key: E,J,K

Pädiatrie: übersetztes Beispiel aus Case, Swanson (1998), p.100

Eignung

Der Typ ist für Problemstellungen geeignet, bei denen es mehrere wichtige Optionen gibt, die sich deutlich von anderen abheben, wie in folgender Darstellung die Antworten B und H.

F C	I E D	A	G	H	B
schlechteste Wahl				beste Wahl	

Wichtige Formulierungshinweise

- Es gelten die gleichen Formulierungsregeln wie bei Typ R.
- PickN sollte nur verwendet werden, wenn es wirklich um eine Graustufenabwägung geht (wichtigste Tests, wichtigste Massnahmen, wahrscheinlichste Diagnosen). Für blosse richtig/ falsch-Entscheide ist der Typ Kprim adäquat (s. 3.6).

Auswertung

Untersuchungen haben ergeben, dass bei mehr als zwei verlangten besten Antworten eine Teilpunktebewertung sinnvoll sein kann (wie beim Typ Kprim). Damit kann die statistische Schwierigkeit diese Typs reduziert und die Trennschärfe verbessert werden.

z.B. halber Punkt für: 2 von 3 richtig
3 von 4 richtig
3 von 5 richtig

3.6 Vierfache Entscheidung richtig/falsch (Typ K', genannt Kprim)

Definition

Auf eine Frage oder unvollständige Aussage folgen vier Antworten oder Ergänzungen. Für jede muss entschieden werden, ob sie **richtig** oder **falsch** ist. Um einen Punkt zu erhalten, müssen alle vier Beurteilungen korrekt sein. Es hat sich als sinnvoll erwiesen, drei korrekte Beurteilungen bereits mit einem halben Punkt zu honorieren.[2,3]

Beispiele

1. *Eine Anreicherung der Inspirationsluft mit 5 % CO_2 führt beim Gesunden zu:*
   (A) arterieller Hypoxämie
   (B) Ventilationssteigerung
   (C) Zunahme der Gehirndurchblutung
   (D) Abnahme des Plasmabikarbonats

   Pathophysiologie (3. Jahr): *GP P 46, R 0.32* *HP P 65, R 0.33* Key: -++-

   GP = 1 Punkt für 4 richtige Beurteilungen, 0 Punkte für weniger als 4.
   HP = 1 Punkt für 4 richtige, 1/2 Punkt für 3 richtige Beurteilungen.

2. *Sie lesen das Thorax-Röntgenbild eines Patienten, bei welchem ein Emphysem vermutet wird.*
   Was entspricht dieser Verdachtsdiagnose?
   (A) Zunahme des Retrosternalraumes
   (B) stumpfer sterno-diaphragmatischer Winkel
   (C) horizontale Rippen
   (D) vermehrte pulmonale Vaskularisation in der Peripherie

   Chirurgie (Schlussprüfung): *GP P 66, R 0.23* *HP P 80, R 0.27* Key: +++-

Eignung

Inhaltlich betrachtet ist der Typ K' angezeigt, wenn es um einen Sachverhalt geht, bei dem mehrere Aspekte bedeutsam sein können, resp. ein Problem, zu dessen richtiger Lösung mehrere Elemente gehören können. Alle Antworten müssen schwarz/weiss beurteilbar sein.
Der Typ K' sollte nicht missbraucht werden, um völlig heterogene Aussagen zu einem breiten Thema in einem Item zusammenzuwürfeln.

Unter messtechnischem Aspekt zeigen K'-Items etwas häufiger Probleme als Items der Typen A und B und müssen eliminiert werden, weil eine einzige nichtfunktionierende Teilantwort ein Item zu Fall bringen kann. Hinsichtlich Trennschärfe sind K'-Items mit der Alles-oder-Nichts-Auswertung aber den Typen A und B ebenbürtig. Mit der HP-Auswertungsvariante, bei welcher drei richtige Teilantworten bereits mit 0.5 Punkten belohnt werden, sind sie diesen gar tendenziell überlegen. In der Alles-oder-Nichts-Form sind sie deutlich schwerer als Einfachwahlfragen und bei den Kandidaten unbeliebt. Durch die Halbpunkt-Auswertungsversion wird beides positiv verändert.[2,3]

Kein Typ K mehr!

Beim Typ K wurden zu vier Aussagen 1–4 fünf mögliche Lösungsmuster vorgegeben:

(A) 1+2+3 (B) 1+3 (C) 2+4 (D) nur 4 (E) alle vier

Dieser Typ schneidet messtechnisch wesentlich schlechter ab als der Typ Kprim und sollte nicht mehr verwendet werden. Wird eine Aussage als falsch identifiziert, scheiden drei Lösungsmuster aus und die Ratewahrscheinlichkeit beträgt bereits 50 %.[3]

Wichtige Formulierungshinweise

- Die **Formulierung des Stammes muss offen lassen, wieviele der vier Aussagen richtig sind**.

  Früher wurde dies oft durch leseunfreundliche Formulierungen gewährleistet in der Art:

  'Welche/r Zusammenhang/Zusammenhänge trifft/treffen zu?'

  Heute weisen wir zum einen in der Instruktion für die Kandidaten darauf hin, dass unabhängig von der Einzahl- oder Mehrzahlformulierung 0-4 Aussagen richtig sein können. Zum andern wählen wir möglichst quantitätsneutrale Stammformulierungen, wie:

  'Für dieses XY gilt: .. spricht: .. trifft zu:

  oder einer quantitätsneutralen Frage, wie:

  'Was gehört zu den XY?', 'Was kommt als Ursache in Betracht', 'Was ist als Massnahme geeignet?,' 'Wann ist ein XY angezeigt?'

  Kann als inhaltlich passender Stamm nur formuliert werden*:*

  'Beurteilen Sie folgende Aussagen zu XY:'

  sollte kritisch überprüft werden, ob die Frage inhaltlich nicht allzu heterogen ist.

- Der **Stamm** muss **immer positiv** formuliert werden; in den Antworten sind Negationen wenn möglich auch zu vermeiden.

  Bei der Beurteilung entstehen doppelte Negationen (Es ist falsch, dass es falsch ist.), die verwirrlich sind und die Messabsicht stören können.

- **Jede Aussage** muss **eindeutig richtig oder falsch** sein.

- Es sollen **nie zwei Aussagen in eine gepackt** werden (z.B. Aussage mit Begründung).

  Es kann unklar werden, was zu beurteilen ist.

- **Jede Antwort** muss **unabhängig von allen andern** sein. Insbesondere sind Antworten zu vermeiden, die sich gegenseitig ausschliessen.

- **Vage Begriffe** wie 'gewöhnlich', 'häufig', 'oft', 'assoziiert mit' sind möglichst zu **vermeiden**.

  Untersuchungen haben gezeigt, dass solche Begriffe mit sehr unterschiedlichen Quantifizierungen verbunden werden. Damit ist es u.U. nicht mehr nur vom Fachwissen abhängig, ob einer solchen vagen Aussage zugestimmt wird.[4,5]

- **Richtige und falsche Aussagen** sollten **ausbalanciert** sein.

  Es besteht bei Autoren eine klare Tendenz, auch in der Prüfung noch Wissen vermitteln zu wollen und daher positive Aussagen vorzuziehen. Da die Mehrzahl der Kandidaten dies zu erwarten scheint und bei Nicht-Wissen eher auf eine positive Antwort tippt, weisen diese tendenziell eine schwächere Trennschärfe auf als negative.

3.7 Kausale Verknüpfung (Typ E)

Definition

Zwei Aussagen sind durch das Wort 'weil' verknüpft. Es sind zuerst unabhängig voneinander die beiden Aussagen als **richtig** oder **falsch** zu beurteilen. Wenn beide richtig sind, ist zusätzlich die Berechtigung der weil-Verknüpfung zu beurteilen. Das ergibt fünf Antwortmöglichkeiten:

(A)	+weil+	Beide Aussagen stimmen, die weil-Verknüpfung ist berechtigt.
(B)	+/+	Beide Aussagen stimmen, deren weil-Verknüpfung ist falsch.
(C)	+/-	Die erste Aussage ist korrekt, die zweite ist falsch.
(D)	-/+	Die erste Aussage ist falsch, die zweite ist korrekt.
(E)	-/-	Beide Aussagen sind falsch.

Beispiel

Die Verminderung von 2,3-DPG in den Erythrozyten begünstigt die Gewebsoxygenation,

weil

die Dissoziationskurve des Hämoglobins bei Verminderung des 2,3-DPG in den Erythrozyten nach rechts verschoben wird.

(A)	*(B)*	*(C)*	*(D)*	*(E)*
+weil+	*+/+*	*+/-*	*-/+*	*-/-*

Pathophysiologie (3. Jahr): P 86, R 0.32 Key: E

Eignung

Inhaltlich betrachtet wären Items vom Typ E in Gebieten angezeigt, in denen kausale Zusammenhänge bedeutsam sind. Tatsächlich werden sie von Fragenautoren häufig als Ausweg benutzt, wenn sie zu einem Thema keine vier plausiblen Distraktoren für ein Item vom Typ A finden.

Unter messtechnischem Gesichtspunkt sind E-Items ziemlich problematisch. Kausalitäten sind recht selten schwarz/weiss zu beurteilen. (Ist die Kausalität z.B. zu bejahen, wenn die zweite Aussage nur ***ein*** Grund unter mehreren ist?) Die Entscheidung zwischen den Antworten A und B ist damit oft auch eine Ermessensfrage und nicht nur vom Fachwissen abhängig. Anderseits ist es schwierig, E-Items zu konstruieren, in denen alles plausibel erscheint, obwohl eine oder gar beide Aussagen falsch sind. E-Items sind damit anfällig auf Cues, welche ihre Trennschärfe beeinträchtigen.[3]

☞ **Typ E-Items** sollten deshalb - **wenn überhaupt - nur sehr sparsam eingesetzt** werden.

Wichtige Formulierungshinweise

- Beide Aussagen müssen in sich geschlossene Aussagen sein und alle notwendigen Informationen enthalten, um unabhängig voneinander auf ihre Richtigkeit hin beurteilt werden zu können.
- Bei Items mit zwei richtigen Aussagen muss die Kausalität eindeutig richtig oder falsch und nicht Ermessensfrage sein.
- Unabhängig davon, was richtig und falsch ist, soll die Kausalität für Uninformierte immer plausibel erscheinen.
- Es soll vermieden werden, dass E-Items überwiegen, bei denen alles inkl. kausaler Verknüpfung korrekt ist oder bei denen nur der erste Teil richtig ist.

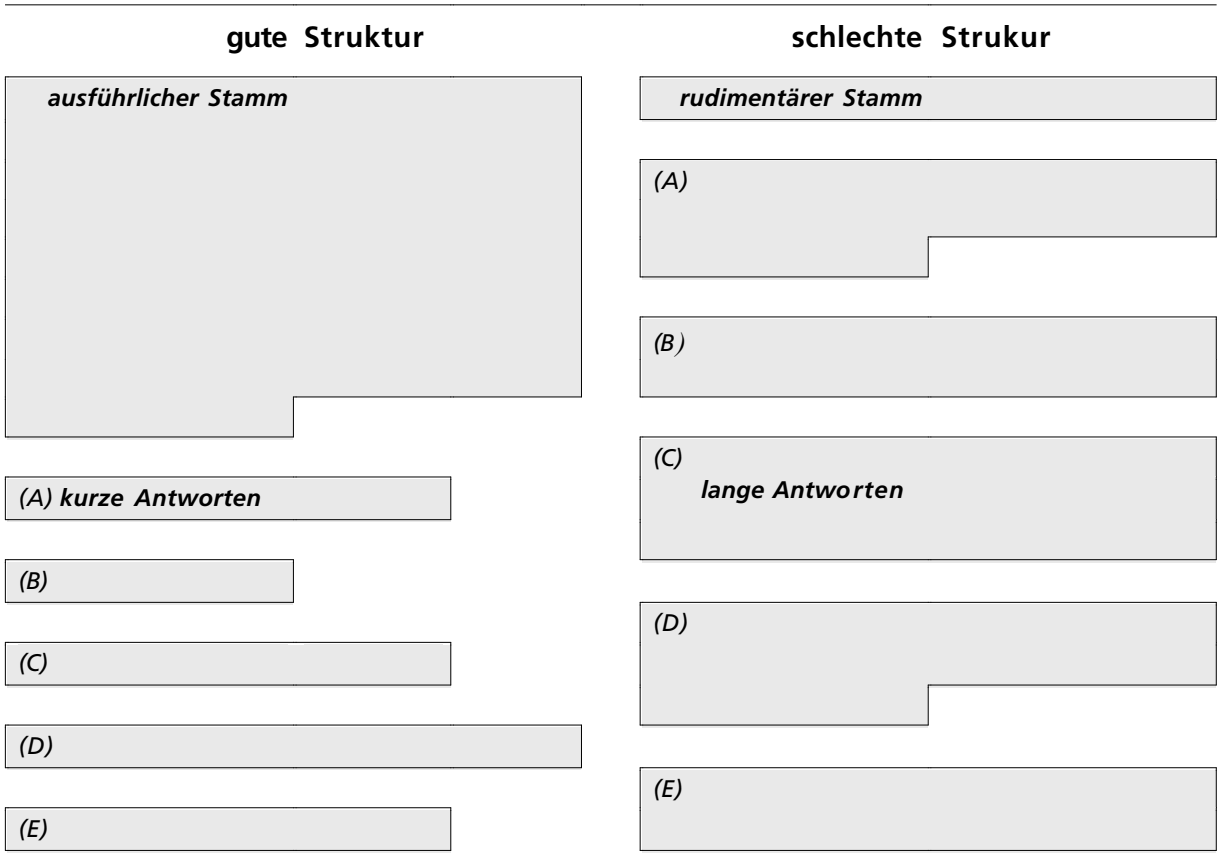
4. Wie werden Fragen formuliert

Drei Grundregeln:

Soll mit MC-Items mehr geprüft werden als Faktenwissen, müssen komplexere Problemsituationen vorgelegt werden, die mehrere zu interpretierende und integrierende Informationen enthalten. Gute Items benötigen deshalb häufig einen **ausführlicheren Itemstamm**.

1. langer Stamm, kurze Antworten

Bei den Best-Antwort-Typen müssen die **Antworten** oftmals gegeneinander abgewogen werden. Dies ist fast nur möglich, wenn diese **kurz und übersichtlich** sind.



2. einfach und klar formulieren

Items sollen **einfach, klar** und in einer allgemein akzeptierten Sprache formuliert werden.

Dazu gehört auch, dass nur sehr geläufige Abkürzungen ohne zusätzliche ausgeschriebene Version in Klammer verwendet werden.

3. keine Fangfragen kreieren

Items sollen **nicht künstlich kompliziert** gemacht **oder** gar als bewusste **Fangfragen** konzipiert werden.

Die Schwierigkeit eines Items soll von der Komplexität des zugrunde liegenden Problems, dem kognitiven Anspruch (Verständnis, Syntheseleistung, Problemlösung) und der Feinheit der erforderlichen Differenzierung (gegenseitige Nähe der Wahlantworten) bestimmt sein.

Es ist unfair und widerspricht mit Sicherheit der Prüfungsabsicht, Items, die triviale Faktenkenntnisse prüfen, durch formale Tricks schwieriger zu machen.

4.1 Was ist bei der Formulierung des Stammes zu beachten?

Formulierungsregeln für den Itemstamm

Bei der Formulierung sollten Sie darauf achten, dass der Stamm

- **alle** für die Beantwortung **erforderlichen Informationen** enthält, so dass in den Wahlantworten keine zusätzlichen solchen gegeben werden müssen.
- in der Regel **keine überflüssigen Informationen** enthält ...

  es sei denn, es soll ausdrücklich die Fähigkeit geprüft werden, relevante Informationen herauszufiltern. Sonst wird mit überflüssigem Text wertvolle Prüfungszeit verschenkt. Hüten Sie sich insbesondere davor, in einer selektiven Prüfung noch Lerninhalte vermitteln zu wollen.
- möglichst **beantwortbar** ist, **ohne die Antworten zu sehen**.

  Wenn dies nicht der Fall ist, fragen Sie sich, ob das Item nicht zu heterogen ist.
- bei Einfachwahlitems möglichst, bei Kprim-Items immer **positiv formuliert** ist.

  Aneg-Items sind unter dem Validitätsgesichtspunkt wenig erwünscht. Bei K'-Items entstehen bei negativer Stammformulierung unweigerlich verwirrliche doppelte Negationen.

4.2 Was ist bei der Formulierung der Antworten zu beachten?

Damit ein Item zur Validität und Reliabilität der Prüfung beiträgt, sollte das Finden der richtigen Antwort möglichst ausschliesslich davon abhängen, ob die Antwortenden über das Wissen verfügen, das mit dem Item geprüft werden soll. Um dies zu erreichen, müssen die Wahlantworten eine ganze Reihe inhaltlicher, formaler und sprachlicher Kriterien erfüllen.

inhaltliche Kriterien

- Alle Wahlantworten sollen in die gleiche Kategorie fallen, also **inhaltlich homogen** sein.

  Diese Regel wird häufig verletzt bei schlecht fokussierten, diffusen Itemthemen.

- **Für Distraktoren** sollen **klare Gründe** bestehen. Es kann sich z.B. um häufige Fehlmeinungen, falsche Konzepte, veraltete Ansichten handeln. Zumindest sollte aber eine klar nachvollziehbare Beziehung zum Itemthema bestehen.

  Unplausible, triviale oder gar völlig unsinnige Distraktoren können auch von leistungsschwachen Kandidaten sofort ausgeschlossen werden, was ihre Chance erhöht, die richtige Antwort zu raten, womit das Item seine Trennschärfe verliert.
  Um die Ratewahrscheinlichkeit klein und konstant zu halten, ist es sinnvoll, möglichst durchgehend fünf Wahlantworten vorzugeben. Falls aber wirklich nur drei sinnvolle Distraktoren gefunden werden können, ist es sinnvoller, nur vier Wahlantworten anzubieten, als etwas Absurdes aufzunehmen.

- Die Distraktoren müssen nicht alle völlig falsch, die **richtige Antwort** muss aber **eindeutig die beste** sein. Je näher richtige Antwort und Distraktoren inhaltlich und hinsichtlich Richtigkeit beieinander liegen, desto schwieriger wird das Item.

- Jede Wahlantwort soll **möglichst kurz** sein und **nur eine Aussage** enthalten.

- **Sich überschneidende Wahlantworten** sollten nur verwendet werden, wenn dies das zugrundeliegende Problem erfordert und nicht, um ein Item künstlich schwieriger zu machen.

- **'Alle der obigen'** oder auch Antworten wie 'Sowohl B und C sind richtig' dürfen **nicht verwendet** werden.

  Falls es sich dabei um die vorgesehene Richtigantwort handelt, gibt es mehr als eine richtige Antwort. Falls es eine Falschantwort ist, ist sie geschenkt, sobald nur eines der enthaltenen Elemente als falsch identifiziert werden kann.

- **'Keines der obigen'** sollte **nur in Ausnahmefällen** als fünfte Wahlantwort verwendet werden. Möglich ist die Aussage, wenn sie relevantes Wissen prüft oder wenn z.B. die richtige Antwort wegen allzu starker (evtl. gar verbaler) Assoziation zum Stamm nicht namentlich erwähnt werden kann. Natürlich müsste 'Keines der obigen' dann auch ab und zu die falsche Antwort sein. Dies ist aber häufig problematisch.

  Wenn 'Keines der obigen' als falsche Antwort vorgesehen ist, darf die richtige Antwort nicht mehr nur die beste unter den aufgeführten sein. Sobald eine noch bessere, z.B. differenziertere Antwort möglich ist, wählen sehr gute Kandidaten 'Keines der obigen', womit das Item seine Trennschärfe verliert.

formale Kriterien

Bei den formalen Kriterien geht es primär darum, **unbeabsichtigte** formale und sprachliche **Lösungshinweise**, sog. Cues zu **vermeiden**. Diese erlauben MC-erfahrenen Kandidaten, auch ohne Fachkenntnis die richtige Antwort zu identifizieren oder doch einzelne falsche Antworten auszuscheiden und damit die Ratechance zu erhöhen.
In zweiter Linie soll der **Einfluss bestimmter Beantwortungstendenzen** von Kandidaten **minimiert** werden.

Prüfen Sie Ihre Testknackerfähigkeit anhand des UTC-Tests.

Zur Illustration der wichtigsten, häufigsten Cues finden Sie auf der folgenden Seite den UTC-Test. Wir empfehlen Ihen, diesen zu bearbeiten, bevor Sie weiterlesen.

UTC-Kurztest

Instruktion: Kreuzen Sie bei den folgenden 7 Fragen je die am wahrscheinlichsten richtige Antwort an. Beantworten Sie alle Fragen der Reihe nach.

Tipp: Lassen Sie sich nicht vom Inhalt irritieren. MC-Fragen lassen sich manchmal auch ohne 'Sach-Kenntnisse' lösen.

1. Anter ist eine
   (A) Legierung
   (B) Konglomerat
   (C) Verbrennungsrückstand
   (D) Spaltprodukt
   (E) chemisches Element

2. Bei der Fermierung von Anter mit saurem Gor
   (A) findet eine Abkühlung statt
   (B) entsteht unter der Bedingung einer leichten Erwärmung Anterit im pH-Bereich 2.8-3.2
   (C) wird OgI_4 freigesetzt
   (D) entsteht Fermantin
   (E) bildet sich Gorantoxol

3. Sie wollen mit Ihrem Ektator Gorantoxin lubrieren.
   Dies funktioniert nur, wenn
   (A) das LTC-Modul des Ektators ausgeschaltet ist
   (B) das LTC-Modul des Ektators eingeschaltet ist
   (C) der Ektator über einen OC-Detektor 2000+ verfügt
   (D) die Ω-Frequenz während der Lubrierung konstant bleibt
   (E) das Gorantoxin vorgängig sterniert wird

4. Die Abkürzung USL heisst ausgeschrieben
   (A) United States Laboratories
   (B) Uniform Source Language
   (C) Uniform Source Locator
   (D) Uniform Starting Label
   (E) Unique Spaceship Locator

5. Welches ist das Hauptmerkmal des KRS (Kognitives Rigiditäts-Syndrom)?
   (A) ein erhöhter Ferminspiegel im Plasma
   (B) zyklische postprandiale Alpträume
   (C) häufige Versteifungen der Nackenmuskulatur
   (D) eine reduzierte Beweglichkeit im kognitiven Bereich
   (E) eine chronische Logo- und Skriptorrhö

6. Warum ist bei trigoten Quergeln die Axosie-Auftretensrate erhöht?
   (A) Trigote Quergel sind nie berop.
   (B) Trigotie führt immer zu Enität.
   (C) Alle trigoten Quergel sind esophym.
   (D) Axosie ist ausschliesslich sequid bedingt.
   (E) Trigote Quergel sind gehäuft susmin.

7. Anter kann nicht mit saurem Gor fermiert werden,
   weil
   zur Bildung von Anterit eine leichte Erwärmung erforderlich ist.
   (A) + weil + (B) +/+ (C) +/- (D) -/+ (E) -/-

des Rätsels Lösung

UTC ist die Abkürzung für 'Use The Cues'. In jeder der sieben vorangehenden Fragen ist ein typischer **formaler oder sprachlicher Lösungshinweis** enthalten. Natürlich springen diese nicht immer gleich ins Auge, sonst würden sie auch den Fragenautoren auffallen und könnten vermieden werden.
Kontrollieren Sie Ihre Lösungen und lernen Sie dabei sieben häufige Cues kennen:

sieben häufige Cues, die Sie vermeiden sollten

1. (A) Nur 'Legierung passt grammatikalisch zur Stammformulierung.

- **Alle Antworten müssen grammatikalisch zum Stamm passen.**

2. (B) Dies ist mit Abstand die längste und differenzierteste Antwort.

- **Distraktoren** sollen möglichst **gleich lang und differenziert** sein **wie die richtige Antwort**.

3. (B) oder (A): zweimal 'LTC-Modul', einmal 'aus' einmal 'ein'

- **Hinweise, welche die Aufmerksamkeit auf 2-3 Antworten einschränken**, sind zu vermeiden.
  Werden Kernelemente der richtigen Antwort noch in einer zweiten Antwort verwendet, wird die Aufmerksamkeit darauf gelenkt. Wird in einer falschen Antwort das Gegenteil der richtigen formuliert, muss logischerweise fast eine der beiden richtig sein.

4. (C) In den Antworten ist 3mal 'Uniform' enthalten, 2mal 'Source' und 2mal 'Locator. (C) enthält alle drei Elemente.

- **Konvergenz-Cues** sind zu vermeiden.
  Die Antwort, welche die grösste Zahl von Elementen mit andern Antworten gemeinsam hat, ist mit erhöhter Wahrscheinlichkeit die richtige (sog. Konvergenzstrategie).

Die Ablenker sind ebenso wichtig wie die richtige Antwort.

Die Cues der Fragen 1-4 entstehen alle dadurch, dass für die Autoren die richtige Antwort stark im Vordergrund steht. Sie schenken dieser deshalb beim Formulieren wesentlich mehr Beachtung als den Ablenkern. Bei der Suche von Falschantworten kreisen sie zudem weiterhin um die richtige Lösung. Vermeiden Sie dies!

5. (D) Das Wort 'kognitiv' taucht im Stamm und Antwort D auf.

- **Verbale Assoziationen zwischen Stamm und richtiger Antwort** sind zu vermeiden.

6. (E) Dies ist die einzige nicht absolute Aussage.

- **Absolute Begriffe** wie 'nie', 'immer', um Aussagen eindeutig falsch zu machen, sind zu vermeiden.

7. (D) Informationen zur richtigen Lösung stehen in Frage 2.

- **Gegenseitige Lösungshinweise** sind zu vermeiden.
  Dieser Cue ist v.a. bei der Prüfungszusammenstellung zu beachten. Seine Gefahr wird aber deutlich reduziert, wenn in den Items möglichst keine überflüssigen Informationen gegeben werden.

Zwei weitere formale Fehler, die vermieden werden sollten, betreffen die Anordnung der Antworten:

Reihungs-Cue vermeiden und …

- Die **Antworten** sollen **möglichst logisch angeordnet** werden.

  Wenn bei den Antworten vom Inhalt her eine logische Hierarchie gegeben ist, sollen sie danach angeordnet werden (z.B. bei Frage nach wahrscheinlichster Problemursache von klein, harmlos zu gross, gravierend, bei Antworten mit Zahlenwerten auf- oder absteigend). Falls keine inhaltliche Logik gegeben ist, empfiehlt sich v.a. bei Einwortantworten am ehesten eine alphabetische Reihung.

  Kandidaten, welche die richtige Antwort nicht kennen, werden auch bei der Reihung nach einer Auffälligkeit suchen, die sie als Lösungshinweis nutzen können.

… dem Einfluss von Beantwortungstendenzen entgegenwirken

- Die **richtige Antwort** soll **nicht überwiegend unter C oder D platziert** werden.

  In der Absicht, die richtige Antwort möglichst 'gut zu verstecken', wurde sie von Autoren früher überdurchschnittlich oft in der Mitte platziert. Dies scheint sich herumgesprochen zu haben. Viele Kandidaten, die raten müssen, bevorzugen bei der Antwortwahl die Mitte. Deshalb platzieren Autoren die richtige Antwort jetzt zunehmend eine Position weiter hinten unter D.
  Damit Positions-Wahltendenzen von Kandidaten keinen Einfluss gewinnen können, sollen die richtigen Antworten über alle Items einer Prüfung möglichst ausbalanciert unter A-E verteilt sein.

5. Wie werden Fragen überprüft?

Der Inhalt und die Form von MC-Items müssen vom Autor und von unabhängigen Experten überprüft werden. Wir empfehlen, ein standardisiertes Itemformular zu verwenden (vgl. Anhang 2). Dies trägt dazu bei, dass alle erforderlichen Angaben gemacht werden, was die Arbeit der Experten, welche neue Items begutachten, erleichtert.

Autoren kontrollieren ...

Als Autor sollten Sie nach dem Verfassen eines Items

- dieses nochmals durchlesen und sich vorstellen, dass Sie es selbst beantworten müssen
- anhand der 'Punkte' im Abschnitt 1 dieser Anleitung überprüfen, ob Sie **alle Grundregeln beachtet** haben
- im Speziellen überprüfen, ob das Item **keine Cues** enthält

... und dokumentieren ihre Items

- angeben, zu welchem **Lernziel oder Inhaltsgebiet** (Blueprintkapitel) das Item gehört und welches die **richtige Lösung** ist
- den inhaltlichen Fokus des Items in Form einiger **Kennworte** festhalten und nach Möglichkeit eine oder mehrere **Literaturreferenzen** angeben, welche die Richtigkeit der "richtigen Lösung" belegen
  Allenfalls ist es sinnvoll auch zu begründen, warum gerade diese Distraktoren gewählt worden sind und warum sie falsch sind.
- Legen Sie das Item dann für eine Woche zur Seite und sehen Sie es danach nochmals durch.

Experten überprüfen deren Inhalt ...

Die Revision neuer Items durch unabhängige Experten sollte sinnvollerweise in mehreren Schritten erfolgen.

- 1-2 Inhaltsexperten sollten unabhängig von der Form die **fachliche Richtigkeit** und die **Zugehörigkeit** des Itemthemas **zum Ausbildungsprogramm** begutachten (vgl. Itemformular in Anhang 2).

... und Form

- Ein MC-Prüfungsexperte sollte die Items ausschliesslich unter den Gesichtspunkten **Form und Sprache** begutachten.

Mit den Anmerkungen dieser Experten werden die Items dem für das Fach repräsentativ zusammengestellten Revisionsgremium zugestellt.

Prüfungskommission entscheidet

In der eigentlichen Revisionssitzung wird über **Annahme oder Rückweisung** und evtl. **Modifikationen** entschieden. Die **Relevanzeinstufung**, welche bei der Prüfungszusammenstellung sehr hilfreich ist, kann als Mittelwert der individuellen Einschätzungen berechnet oder als Konsensurteil ausdiskutiert werden.

6. Wie wird eine Prüfung zusammengestellt?

konstante Gewichtung nach Inhalt und Typen

- Um von Session zu Session vergleichbar zu messen, sollten Prüfungen nicht nur hinsichtlich der inhaltlichen Gewichtung der Teilgebiete (gemäss Blueprint) konstant gehalten werden, sondern möglichst auch bezüglich Verteilung der Itemtypen. Wir empfehlen deshalb, auch für die Itemtypen einen ungefähren Verteilungsschlüssel zu definieren.

standardisierende Beantwortungsanleitung

- Das Prüfungsheft muss eine standardisierende Beantwortungsanleitung enthalten, welche das sicherste und ökonomischste Vorgehen beschreibt. Sie muss die Kandidaten auch darüber informieren, wie die Antworten bewertet werden, insbesondere ob falsche Antworten bestraft werden oder ob sie in Zweifelsfällen raten sollen (vgl. Anhang 3).

Gruppierung nach Typen mit je entsprechender Instruktion

- Im Prüfungsheft sollen die Items nach Typen gruppiert werden. Jede Typengruppe ist durch eine konzise Aufgabenbeschreibung einzuleiten (vgl. Anhang 4).

  Es ist für Kandidaten in einer Prüfungssituation mühsam, stressfördernd und eine vom Fachwissen unabhängige Fehlerquelle, sich auf dauernd wechselnde Itemtypen einstellen zu müssen.

Einstieg mit Eisbrecheritem/s

- Die Prüfung sollte mit 1-2 leichten Items beginnen (sog. Eisbrecheritems).

  Jede Prüfung ist eine Stresssituation. Um wirklich die Wissensleistung messen zu können (und nicht z.B. Stressresistenz) ist ein beruhigender Prüfungseinstieg nützlich.

ausgeglichenes Antwortmuster

- Die Verteilung der richtigen Antworten auf die Positionen A-E sollte über die gesamte Prüfung etwa ausgeglichen sein. Zudem ist es sinnvoll, Folgen von fünf oder mehr Items mit der gleichen Richtigantwort zu vermeiden.

  Die erste Massnahme verhindert den möglichen Einfluss gewisser Antworttendenzen von Kandidaten. (Speziell bekannt ist die Tendenz zur Mitte.) Die zweite Massnahme verhindert, dass Kandidaten angesichts einer langen Folge des gleichen Antwortbuchstabens verunsichert werden, konfuse Überlegungen anzustellen beginnen und evtl. richtige Antworten verschlimmbessern.

Verwendete Quellen

Neben der eigenen langjährigen Erfahrung in der Entwicklung, Revision und Auswertung von MC-Prüfungen im Rahmen der ärztlichen Aus- und Weiterbildung in der Schweiz sowie in weiteren Anwendungsfeldern dieser Prüfungsform basieren die Hinweise in dieser Anleitung auf einer Reihe internationaler Publikationen zum Thema "Entwicklung von MC-Items" ([6,7,8,9,10]). Besonders verweisen möchten wir auf das vorzügliche Manual von Case und Swanson[6], das über das Internet heruntergeladen werden kann:
www.nbme.org/about/itemwriting.asp

Literatur

1 Hubbard JP, Clemans WV. Multiple-Choice Examinations in Medicine. Philadelphia: Lea & Febinger, 1961

2 Krebs R. The Swiss way to score multiple true-false items: theoretical and empirical evidence. In: Scherpbier AJJA, van der Vleuten CPM, Rethans JJ, van der Steeg, eds. Advances in Medical Education. Dordrecht: Kluwer Academic Publishers, 1997:158-61

3 Itten S, Krebs R. Messqualität der verschiedenen MC-Itemtypen in den beiden Vorprüfungen des Medizinstudiums an der Universität Bern. AAE/IML, Forschungsbericht 1997/2, 24 S.

4 Case SM. The use of imprecise terms in examination questions: how frequent is frequently? Acad Med 1994;69(suppl):4-6

5 Holsgrove G, Elzubeir M. Imprecise terms in UK medical multiple-choice questions: what examiners think they mean. Med Educ 1998;32:343-50

6 Case S, Swanson DB. Constructing Written Test Questions For the Basic and Clinical Sciences. 2nd ed. Philadelphia: National Board of Medical Examiners, 1998

7 Gronlund NE. How to construct achievement tests. 4th ed. Englewood Cliffs, NJ: Prentice-Hall, 1987

8 Linn RL, Gronlund NE. Measurement and assessment in teaching. 8th rev. ed. Englewood Cliffs, NJ: Prentice-Hall, 1999

9 Ellsworth AR, Dunnell P, Duell OK. Multiple-choice test items: What are textbook authors telling teachers? J Educ Res 1990;83:289-93

10 Aiken LR. Testing with multiple-choice items. J Res Dev Educ 1987; 20:44-58

Grundlagen für die Herstellung standardisierter Fragen

Prüfziel: .....

.....

Thema: .....

.....

Aussagen

	wichtig	unwichtig
positiv	bedeutsame, grundlegende Aspekte	nebensächliche Aspekte
negativ	gravierende Fehler, häufige Fehlmeinungen	belanglose Fehler

Institut für Dies und Das an der Universität Irgendwo

ANHANG 2

Prüfung: .....

Themenbereich: .....

Revisions–Nr.:

Prüfungsziel der Frage: .....

Quelle, Referenz: .....

richtige Antwort: ..... Fragetyp: .....

Frage:

Formular verkleinert dargestellt

Autor/in: ..... Revisor/in: .....

Formulierungs-kontrolle:	**inhaltliche Vorrevision:**		
	Bezug zur Ausbildung	**fachliche Richtigkeit**	**Relevanz**
☐ gut	☐ ist Ausbildungsinhalt	☐ richtig, eindeutig	☐ essentiell
☐ Korrektur	☐ marginal	☐ fraglich	☐ mittel
☐ ungeeignet	☐ kein Ausbildungsinhalt	☐ falsch	☐ gering

Revisions-entscheide:			
	☐ Frage akzeptiert	☐ zu überarbeiten	☐ abgelehnt
	Relevanz: ☐ essentiell	☐ mittel	☐ gering

Allgemeine Hinweise zur korrekten Beantwortung

Die Kandidaten sollten zu Beginn der Prüfung darauf hingewiesen werden, die folgenden, z.B. auf die Rückseite des Prüfungsheftes gedruckten Hinweise aufmerksam zu lesen. Diese müssen natürlich der konkreten Situation angepasst werden.

1. Lesen Sie immer die ganze Frage und alle Wahlantworten sorgfältig durch.

2. Bezeichnen Sie Ihre Antworten zunächst im Fragenheft. Bei den Fragetypen mit Wahl der besten Antwort umkreisen Sie die eine Wahlantwort, die Sie für die zutreffende halten, bei Typ PickN die angegebene Zahl von Antworten. Bei den Fragen vom Typ K' dagegen ist jede Antwort, die zutrifft, mit (+), jede Antwort, die nicht zutrifft, mit (-) zu bezeichnen.

3. Beantworten Sie alle Fragen. Wenn Sie nicht sicher sind, wählen Sie die für Sie am wahrscheinlichsten richtige Antwort.
   Für jede richtig beantwortete Frage erhalten Sie einen Punkt, bei den K'-Fragen für drei richtige Teilantworten bereits einen halben Punkt. Ebenso können bei den PickN-Fragen mit richtigen Teilantworten halbe Punkte erzielt werden. Falsche Antworten werden nicht bestraft. Jede nicht beantwortete Frage wird wie eine falsch beantwortete mit 0 Punkten bewertet.

4. Übertragen Sie Ihre Antworten erst auf das Auswertungsblatt, nachdem Sie sich bei allen Fragen definitiv für eine Antwort entschieden und diese im Fragenheft bezeichnet haben. Bei den Fragetypen mit Wahl der besten Antwort darf immer nur ein Feld markiert werden, bei Typ PickN genau die verlangte Anzahl Antworten.
   Für die Antworten auf die K'-Fragen sind die mit "K" bezeichneten Felder des Auswertungsblattes auszufüllen. Für alle 4 Teilantworten ist hier je das Plus-Feld (+) oder das Minus-Feld (-) zu markieren.
   Die Antworten auf die R- und PickN-Fragen sind auf der Rückseite des Auswertungsblattes zu markieren.

5. Das Auswertungsblatt darf nur mit dem zur Verfügung gestellten Bleistift ausgefüllt werden. Markieren Sie bitte deutlich und exakt und radieren Sie möglichst wenig. Eine undeutliche Markierung oder zusätzliche Bleistiftstriche ausserhalb der vorgesehenen Felder können zu Fehlern beim optischen Lesen durch den Computer und dadurch zu Falschbewertungen führen.

6. Unterschreiben Sie das Fragenheft und das Auswertungsblatt im bezeichneten Feld. Sie bestätigen damit, dass Sie ohne unzulässige Hilfsmittel gearbeitet haben.

Instruktionen zur Beantwortung der einzelnen Typen

Im Prüfungsheft ist jeder Itemgruppe eines bestimmten Typs die entsprechende Beantwortungsinstruktion voranzustellen.

Typ A **Einfachauswahl**

Bezeichnen Sie nur eine Wahlantwort durch Umkreisen des betreffenden Buchstabens:
- bei positiver Formulierung die einzig richtige, respektive die am meisten zutreffende Antwort
- bei negativer Formulierung die einzige Ausnahme, die einzige falsche Antwort, resp. die Antwort mit dem am wenigsten zutreffenden Inhalt. (Die Negation ist fett gedruckt.)

Typ B **Zuordnungsaufgabe**

Auf fünf mit den Buchstaben (A) bis (E) bezeichnete Wahlantworten folgt eine Gruppe nummerierter Fragen oder Aussagen. Ordnen Sie jeder davon eine Wahlantwort zu, die einzig richtige resp. die am besten passende, und umkreisen Sie den entsprechenden Buchstaben. Ein und dieselbe Antwort kann dabei mehr als einmal die richtige sein.

Typ R **Erweiterte Zuordnung**

Auf eine Liste von maximal 26 Wahlantworten, die alphabetisch geordnet und mit Buchstaben bezeichnet sind, folgt eine Gruppe nummerierter Fragen oder Aussagen. Ordnen Sie jeder davon **eine Wahlantwort** zu, die einzig richtige resp. die am besten passende, und schreiben Sie den entsprechenden Buchstaben unter die Fragennummer. Ein und dieselbe Antwort kann dabei mehr als einmal die richtige sein.

Typ PickN **Wahl einer angegebenen Zahl bester Antworten**

Auf eine Liste von maximal 26 Wahlantworten, die alphabetisch geordnet und mit Buchstaben bezeichnet sind, folgt eine Gruppe nummerierter Fragen oder Aussagen. Ordnen Sie jeder davon **soviele Wahlantworten** zu **wie angegeben**, und schreiben Sie die entsprechenden Buchstaben unter die Fragennummer. Ein und dieselbe Antwort kann dabei mehr als einmal die richtige sein.

Typ Kprim **Vierfache Entscheidung richtig/falsch**

Auf eine Frage oder unvollständige Aussage folgen vier Antworten oder Ergänzungen. Beurteilen Sie bei jeder davon, ob sie richtig oder falsch ist, und bezeichnen Sie sie entsprechend mit (+) oder (-). Unabhängig davon, ob die Frage grammatikalisch im Singular oder Plural formuliert ist, können 1, 2, 3, 4 oder auch gar keine der Antworten richtig sein.
Die korrekte Beurteilung aller 4 Antworten oder Ergänzungen wird mit einem ganzen Punkt honoriert. 3 richtige Beurteilungen erhalten einen halben Punkt.

Typ E **Kausale Verknüpfung**

Zwei Aussagen sind durch das Wort "weil" verknüpft. Es sind zuerst unabhängig voneinander die beiden Aussagen als richtig oder falsch zu beurteilen. Wenn beide richtig sind, ist zusätzlich zu entscheiden, ob die weil-Verknüpfung berechtigt ist. Das ergibt fünf Antwortmöglichkeiten:

(A)	+weil+	Beide Aussagen stimmen, die weil-Verknüpfung ist berechtigt.
(B)	+/+	Beide Aussagen stimmen, deren weil-Verknüpfung ist falsch.
(C)	+/-	Die erste Aussage ist korrekt, die zweite ist falsch.
(D)	-/+	Die erste Aussage ist falsch, die zweite ist korrekt.
(E)	-/-	Beide Aussagen sind falsch.

Glossar prüfungstechnischer Fachbegriffe

Ablenker: Falschantwort in einem MC-Item mit einer besten Antwort. Kandidaten mit fehlender Sachkenntnis sollten die richtige Antwort nicht (z.B. aufgrund von ↗ Cues) von den Ablenkern unterscheiden können. Hingegen darf es nicht darum gehen, Kandidaten durch Ablenker bewusst in Fallen zu locken.

Blueprint: gewichtetes Inhaltsraster der Prüfungsinhalte, nach dem alle Prüfungen zusammengesetzt werden. Kann eine oder auch mehrere Dimensionen enthalten.

Cue: „Wink" oder „Fingerzeig". Von einem Cue spricht man, wenn durch die Art der Fragestellung oder die Formulierung der Antworten ein Hinweis auf die richtige Antwort gegeben wird bzw. diese mehr hervorgehoben wird.
Im Interesse der ↗ Reliabilität sollte dies vermieden werden.

Distraktor: s. Ablenker

Evaluation, formative vs summative: Dienen Prüfungsergebnisse zur Orientierung über den Zwischenstand des Lernverlaufs und haben für den Geprüften keine zwingenden Konsequenzen, spricht man von formativer Evaluation. Als summative Evaluation wird dagegen eine sanktionierende Beurteilung am Schluss eines Bildungsabschnitts bezeichnet.

Item: einzelnes Element. Im Zusammenhang mit Prüfungen einzelne Frage, Aufgabe, Beobachtungs- oder Beurteilungseinheit

Item-Analyse: Beurteilung der Messeigenschaften eines Items. Man beurteilt primär die ↗ Item-Schwierigkeit und die ↗ Item-Trennschärfe. Bei MC-Items wird zudem überprüft, ob die einzelnen falschen Antworten wunschgemäss vorwiegend schwache Kandidaten von der richtigen Antwort ablenken (sog. Distraktorenfunktion).

Item-Schwierigkeit: Verhältnis der von einer Kandidatengruppe bei einem Item erreichten zur maximal möglichen Punktzahl. Sie wird entweder als Prozentwert (P) oder als Wahrscheinlichkeitswert (p) angegeben.

Zu bedenken ist, dass bei 5 Wahlantworten die blosse Ratewahrscheinlichkeit 20 % (P=20) beträgt. Für eine gute Leistungsdifferenzierung eignen sich Items im Bereich von P 40-90.

Item-Trennschärfe: Fähigkeit eines Items, Kandidaten mit guter und schlechter Leistung in der Gesamtprüfung zu trennen. Sie wird berechnet als ↗ Korrelationskoeffizient (R) zwischen der erreichten Punktzahl in diesem Item und der Gesamtpunktzahl in der Prüfung ohne dieses Item.

Für eine zuverlässige Leistungsdifferenzierung sind Items mit klar positiver Trennschärfe erforderlich, wenn möglich $\geq .20$, sicher aber $\geq .10$. Items mit Trennschärfen um 0 tragen nichts zur Differenzierung bei, solche mit negativer Trennschärfe laufen der Differenzierungsabsicht zuwider und sollten aus messtechnischer Perspektive eliminiert werden.

Korrelationskoeffizient: statistisches Mass des Zusammenhangs zwischen zwei variablen Grössen, das Werte von –1 bis +1 annehmen kann. Bei Null besteht kein Zusammenhang, Eins ist der bestmögliche positive Zusammenhang, bei Minuswerten besteht ein inverser Zusammenhang.

Objektivität: Im Zusammenhang mit Prüfungen wird unter Objektivität die Unabhängigkeit der Prüfungsergebnisse von den Untersuchern verstanden. Es wird weiter differenziert zwischen Durchführungs-, Auswertungs- und Interpretationsobjektivität. Ermittelt wird die Objektivität meist als statistische Übereinstimmung zwischen verschiedenen Untersuchern. Der dabei verwendete Begriff Interrater-Reliabilität weist auf den Zusammenhang mit der ↗Reliabilität hin.

Reliabilität: Zuverlässigkeit. Prüfungsqualitätskriterium, das danach fragt, wie genau ein Merkmal gemessen wird, gleichgültig, ob dieses Merkmal auch zu messen beansprucht wird (vgl. Validität). Der Reliabilitätskoeffizient schwankt zwischen 0 und dem Maximalwert 1. Fehlereinflüsse, welche eine Messung trüben können, sind etwa mangelnde ↗ Objektivität, Rateeinflüsse, zu kleine, nicht repräsentative Itemauswahl, Zufälligkeiten. Im Vordergrund steht heute die Reliabilität des Prüfungsinstrumentes, meist erfasst in Form des Koeffizienten alpha von Cronbach. Dieser gibt Auskunft darüber, wie stark die Ergebnisse von der spezifischen Itemauswahl abhängen, resp. wie gut eine alternative Prüfung mit einer gleichen Zahl von Items, die nach gleichen Kriterien aus dem gleichen Inhaltsbereich gezogen würden, zur gleichen Rangierung der Kandidaten führen würde.

Für Entscheide mit einschneidender Konsequenz für die betroffenen Personen sollte die Reliabilität mindestens 0.8 betragen; als optimal wird 0.9 erachtet.

Validität: Gültigkeit. Prüfungsqualitätskriterium, das danach fragt, ob das betreffende Verfahren wirklich das misst, was beabsichtigt ist. Bezogen auf eine Prüfung am Ende einer Ausbildung ist es die Frage, ob die für die Berufsaufgaben erforderlichen Kompetenzen gemessen werden. Es werden folgende Aspekte unterschieden:

Inhaltsvalidität fragt danach, wie repräsentativ die ausgewählten Prüfungsitems für den Inhalt bzw. den Umfang des zu prüfenden Kompetenzbereichs oder Stoffgebiets sind. Die Beurteilung sollte durch Fachexperten erfolgen.

Kriteriumsvalidität fragt danach, wie gut die Prüfungsergebnisse mit Leistungen ausserhalb der Prüfungssituation, z.B. in der weiteren Ausbildung oder im Berufsalltag übereinstimmen. Sie wird meist durch Korrelationsstudien zu klären versucht.

Konstruktvalidität fragt danach, ob Hypothesen, die aus einer Theorie über das zu messen beabsichtigte Konstrukt (z.B. 'Problemlösefähigkeit') abgeleitet sind, durch Befunde im Zusammenhang mit den Prüfungsergebnissen gestützt werden.

Augenscheinvalidität (Face validity) ist - wie der Name sagt - eine Scheinvalidität und damit kein Gütekriterium wie die drei vorangehenden. Eine Prüfung, die Augenscheinvalidität besitzt, erweckt beim Betrachter den Eindruck das zu erfassen, was sie zu erfassen beansprucht. Dies kann erstrebenswert sein, weil sie damit von den Kandidaten und Entscheidungsträgern besser akzeptiert wird.