Empfehlung zum Einsatz von Multiple-Choice-Prüfungen

TUM Center for Study and Teaching

Qualitätsmanagement

Stand: Oktober 2022

Vorwort

Sehr geehrte Damen und Herren,
liebe Kolleginnen und Kollegen,

die vorliegende Empfehlung zum Einsatz von Multiple-Choice Prüfungen soll Sie bei der Konstruktion, der Testauswertung sowie bei der Qualitätssicherung Ihrer Multiple-Choice-Prüfungen unterstützen. Die Empfehlung ist daher bewusst umfangreich gestaltet und soll Ihnen zu jedem Themenbereich ausführliche Hilfestellung geben.

Das Study and Teaching Board hat eine AG Multiple-Choice mit der Aufgabe eingesetzt, Empfehlungen für den Einsatz von Multiple-Choice Prüfungen an der TU München mit Anregungen und Beispielen aus der Praxis zusammenzustellen.

Für die Mitarbeit bedanken sich die Autorinnen bei den Mitgliedern Nicolas André, Katharina Eben, Thomas Stolte und Endres Volkmann.

Ihr TUM Center for Study and Teaching – Qualitätsmanagement

Simone Gruber und Manuela Avallone

Inhaltsverzeichnis

1. Einleitung: Zielsetzung der Empfehlungen

Hochschulprüfungen spielen für Studierende und Lehrende gleichermaßen eine große Rolle. Eine wesentliche Funktion von Prüfungen liegt darin, als „diagnostische" Leistungsmessung Lehrenden und auch Studierenden Rückmeldungen über den Leistungsstand zu geben und damit Hinweise für künftige Steuerung von Lehr- und Lernprozessen zu liefern (vgl. Müller/Bayer 2007: 225).

Wenn bei großem Prüfungsvolumen eine hohe Auswertungseffizienz gefordert ist, können Multiple-Choice (MC) Tests die Prüfungsformen der Wahl sein. MC-Tests sind allerdings nur dann effiziente, zuverlässige, objektive und gegenüber den Studierenden auch „gerechte" Prüfungsformen, wenn sie „richtig" gemacht werden. Damit kompetenzorientiertes Prüfen eine „möglichst genaue und fehlerfreie Abbildung der Fähigkeiten von Studierenden in einem bestimmten Stoffgebiet zu einem bestimmten Zeitpunkt" (Brauns/Schubert 2008: 93) ermöglicht, gelten für MC-Prüfungen – wie bei jeder Prüfung – ähnliche Anforderungen, die auch an andere Messinstrumente in der empirischen Forschung angelegt werden.

Mit diesen Empfehlungen sollen daher Standards gesetzt werden, um für MC-Prüfungen an der TUM ein einheitliches und möglichst hohes Qualitätsniveau sicherstellen zu können. Hierzu gilt es Prüfer für die Schwierigkeiten „guter" Fragenkonstruktion zu sensibilisieren und anhand von Beispielen Lösungsmöglichkeiten zur Vermeidung von Fehlerquellen aufzuzeigen.

Dazu werden im Kapitel 3 Empfehlungen zum Einsatz und zur Konstruktion von MC-Tests gegeben, wobei auf Einsatzbereiche von MC-Tests und die dabei zu beachtenden Vorteile und Nachteile bei unterschiedlichen Fragetypen eingegangen wird (vgl. 4.1), bevor die Besonderheiten der Fragebogenkonstruktion (vgl. 4.2) und Testauswertung (vgl. 4.3) erläutert werden. In Kapitel 5 werden Empfehlungen zur Organisation der MC-Tests innerhalb der Departments aufgeführt, die neben Fragen der Itemanalyse und -statistik auch Prozesse der Qualitätssicherung umfassen. Einen Überblick über die wichtigsten Empfehlungen ermöglicht die folgende Zusammenfassung. Ein Glossar (vgl. Kapitel 6) und Hinweise auf weiterführende Literatur (vgl. Kapitel 7) schließen die Empfehlungen ab.

2. Zusammenfassung der wichtigsten Empfehlungen

A. Empfehlungen zum Einsatz von MC-Tests	
Einsatzgebiete und Formen	▪ Multiple-Choice (MC) Tests eignen sich bei großem Prüfungsvolumen, die eine hohe Auswertungseffizienz fordern. ▪ MC-Prüfungen sind allerdings nur dann effiziente, zuverlässige, objektive und gegenüber den Studierenden auch „gerechte“ Prüfungsformen, wenn sie „richtig“ gemacht werden. ▪ MC-Tests sollten nur eingesetzt werden, wenn die mit einem Modul angestrebten Lernergebnisse auf dem Wissens- bzw. Erkenntnisniveau des Erinnerns, Verstehens, Anwendens und Analysierens liegen. (vgl. „Wegweiser zur Erstellung von Modulbeschreibungen“) (vgl. 4.1, S.7f). ▪ Die Bandbreite mit MC-Tests prüfbarer Lernergebnisse reicht prinzipiell von der Reproduktion bzw. dem Erinnern von Faktenwissen bis zur Entwicklung neuer Problemlösungen unter Anwendung von Problemlösungsstrategien. Zu beachten ist allerdings, dass die Entwicklung der Prüfungsfragen mit zunehmender Erkenntnisstufe schwieriger wird (vgl. 4.1, S. 8f). ▪ Die zu messende Fähigkeit sollte die Wahl des Messinstruments bestimmen, d.h. vorab sollte klar sein, was geprüft werden soll, um zu entscheiden, welche Frageform sinnvoll ist (vgl. 4.1, S.9f).
B. Empfehlungen zur Konstruktion von MC-Tests	
Fragebogenkonstruktion	▪ Bei der Fragebogenkonstruktion ist darauf zu achten, dass die Prüfungsergebnisse gültige und zuverlässige Schlüsse auf die Leistung insgesamt ermöglichen, also auf diejenige Leistung, die über das Lösen der konkreten Prüfungsfragen hinausreicht (vgl. 4.2, S.13). ▪ Vor der Fragebogenkonstruktion sollten alle relevanten Prüfungsinhalte in einem Themenraster zusammengestellt und nach Relevanz gewichtet werden, um das angestrebte Wissensniveau zu bestimmen (vgl. 4.2, S.13f).
Formulierungsregeln	▪ Der Aufgabenstamm (d.h. die eigentliche MC-Frage) kann als Frage formuliert werden, so dass die Antwortalternativen die Antwortmöglichkeiten auf diese Frage darstellen. Der Aufgabenstamm kann auch als unvollständiger Satz formuliert sein, der dann von den Antwortoptionen in verschiedener Weise vervollständigt wird. Zudem sollte darauf geachtet werden, dass der Aufgabenstamm einfach, klar und positiv formuliert ist und alle für die Beantwortung der Frage erforderlichen Informationen enthält. Idealerweise lässt sich die Frage beantworten auch ohne die Antwortoptionen zu lesen (vgl. 4.2, S.14). ▪ Die Schwierigkeit einer Frage sollte sich aus der Komplexität des Aufgabeninhalts ergeben und nicht aufgrund einer künstlichen Verkomplizierung durch Schachtelsätze, doppelte Verneinungen, überflüssige Informationen

	o.ä. (vgl. 4.2, S.15). Falls der Aufgabenstamm eine Verneinung enthält, sollte diese kenntlich gemacht werden. ▪ Auf die Wahl und Formulierung der falschen Antwortalternativen (Ablenker/Distraktoren) sollte besondere Aufmerksamkeit gelegt werden, denn die falschen Antwortoptionen bestimmen die Schwierigkeit der Frage. Eine Frage wird umso schwieriger, je näher richtige und falsche Antworten beieinander liegen (vgl. 4.2, S.16). ▪ Alle Antwortoptionen sollten aus demselben Themenbereich stammen, d.h. inhaltlich homogen sein, plausibel auf die Fragestellung bezogen sein und nur eine inhaltliche Aussage enthalten (vgl. 4.2, S.16f). ▪ Jede Antwortalternative sollte klar unterscheidbar sein und es sollte möglichst vermieden werden, sich überschneidende Antwortoptionen zu formulieren. Zudem sollten nur so viele Antwortoptionen formuliert werden, wie sich plausible Ablenker bzw. Distraktoren finden lassen (vgl. 4.2, S.18f). ▪ Alle Antwortoptionen sollten grammatikalisch zum Aufgabenstamm passen und sich in Länge und Differenzierungsgrad nicht unterscheiden (vgl. 4.2, S.20). ▪ Ähnlichkeiten von Wörtern (verbale Assoziationen bzw. Wortwiederholungen) im Aufgabenstamm und in der korrekten Lösung sollten vermieden werden (vgl. 4.2, S.20). ▪ Lehrbuchformulierungen sollten vermieden werden (vgl. 4.2, S.21). ▪ „Absolute“ sprich uneingeschränkte Begriffe wie beispielsweise „niemals“, „immer“, „alle“, „kein“, „nur“ usw., in der Formulierung lassen Ablenker/Distraktoren vermuten, während moderate Begriffe wie beispielsweise „manchmal“, „möglicherweise“, „gewöhnlich“, auf die richtige Antwort schließen lassen (vgl. 4.2, S.21). ▪ Ebenso sollten sog. „Konvergenz-Cues“ (Hinweis auf die richtige Antwort) vermieden werden. Diejenige Antwort, die die größte Zahl an gemeinsamen Elementen, Begriffen mit den anderen Antwortalternativen gemeinsam hat, ist mit hoher Wahrscheinlichkeit die richtige Antwort (vgl. 4.2, S.23). ▪ Antwortoptionen sind möglichst in einer logischen Reihenfolge (aufsteigend, absteigend, alphabetisch) anzuordnen. Die Platzierung der richtigen Antwort sollte nach dem Zufallsprinzip erfolgen (vgl. 4.2, S.23).
C. Empfehlungen zur <u>Auswertung</u> von MC-Tests	
Prüfungsaus-wertung	▪ Bei der Wahl der richtigen Antwort wird das Bonussysteme eingesetzt. Hierbei wird für die richtig angekreuzte Antwort ein Punkt oder mehr - je nach Gewichtung der Frage – gutgeschrieben, ansonsten werden keine Punkte vergeben.

	▪ Bei der Erstellung des Note-Punkte Schlüssels sollte die Ratewahrscheinlichkeit berücksichtigt werden (vgl. 4.3, S.24f).
D. Empfehlungen zur <u>Qualitätssicherung</u> bei MC-Fragen	
Qualitäts-sicherung auf Fragenebene	▪ Nach dem Einsatz einer Frage sollten der Item-Schwierigkeitsgrad (erfasst den Anteil der Prüfungsteilnehmer, die eine Frage richtig beantwortet haben, an der Gesamtzahl der Teilnehmer (P=x/n; 0-1)) und die Item-Trennschärfe (gibt an wie gut das Item zwischen den leistungsstarken und leistungsschwachen Studierenden differenzieren kann) ermittelt werden (vgl. 5.1, S.26).
Qualitätssicherung auf Ebene des MC-Tests	▪ Vor dem Einsatz eines Fragebogens sollte ein Pretest (Probeprüfung des Messinstruments) durchgeführt werden, mit dem die Validität (Gültigkeit des Messinstruments) und Reliabilität (Zuverlässigkeit des Messinstruments) geprüft werden (vgl. 5.2, S.26f).
Qualitätssicherung auf Department-Ebene	▪ Es ist empfehlenswert an einer School Itempools mit bewährten Fragen anzulegen sowie regelmäßig Itemstatistiken durchzuführen. Ebenso ist es ratsam, dass neue MC-Prüfer Schulungen erhalten (vgl. 5.3, S.27). Die Entwicklung qualitativ hochwertiger MC-Tests verlangt von den Prüfenden umfassende Kenntnisse im Bereich der Testkonstruktion und- auswertung.

3. Rechtliche Regelungen

Die TUM hat durch die Änderungssatzung vom Oktober 2012 unter § 12 a (Nr. 4) Regelungen zu MC-Test in die Allgemeinen Prüfungs- und Studienordnung für Bachelor- und Masterstudiengänge (APSO) aufgenommen. § 12 a (Nr. 4) der APSO ersetzt damit die FPSO-Musterregelung zu Multiple-Choice-Aufgaben.

Die bisherige Multiple-Choice-Regelung der FPSO-Musterregelung wurde auf Veranlassung des Vizepräsidenten für Studium und Lehre, Professor Gritzmann, mit der Maßgabe geändert, die Ratewahrscheinlichkeit auf unter 3 Prozent zu reduzieren, so dass in Abs. 1 Satz 1 die Begrenzung auf Einzelfälle entfallen kann.

Aufgrund aktueller Rechtsprechung, welche die Vergabe von Maluspunkten im Rahmen von Mehrfachauswahlaufgaben (x aus n) für unzulässig erachtet, wurden die Mehrfachauswahlaufgaben auf Beschluss des Study and Teaching Board gestrichen. Somit sieht Abs. 2 Satz 1 vor, dass Multiple-Choice-Aufgaben nur in Form von Einfachauswahlaufgaben (1 aus n) gestellt werden dürfen.

Um eine Ratewahrscheinlichkeit von weniger als drei Prozent sicherzustellen, regeln Abs. 2 Sätze 2 und 3, dass bei jeder Prüfungsaufgabe mindestens drei Antwortvorschläge zur Auswahl stehen müssen und jede Multiple-Choice-Aufgabe mindestens 35 Prüfungsaufgaben umfassen muss.

Abs. 3 regelt, dass in Fällen, in denen Prüfungen nur teilweise in Form des Multiple-Choice-Verfahrens abgenommen werden, die Vorgaben des Absatzes 2 nur gelten, sofern der Prüfungsteil, der in Form des Multiple-Choice-Verfahrens abgenommen wird, 20 Prozent übersteigt. Bei nur einem geringen Prüfungsanteil von weniger als 20 Prozent wird eine gegebenenfalls höhere Ratewahrscheinlichkeit durch den überwiegenden Prosaanteil der Klausur ausgeglichen.

4. Empfehlungen zum Einsatz und zur Konstruktion von Multiple-Choice-Tests

Unter „Multiple-Choice-Tests“ werden im engeren Sinne alle Formen schriftlicher Prüfungsformate zusammengefasst, die sich aus mehreren Fragen zusammensetzen und aus denen die Prüfungskandidaten eine richtige Antwort (Einfachauswahl bzw. Single Choice) oder mehrere richtige Antworten (Mehrfachauswahl bzw. Multiple Choice) aus vorgegebenen Antwortmöglichkeiten auswählen. Mit dem Beschluss vom Oktober 2012 sind an der TUM Mehrfachauswahlaufgaben (x aus n) nicht mehr als Prüfungsformate zu wählen. Im weiteren Sinne sollen an der TUM zudem alle Prüfungsformate unter „Multiple-Choice-Tests“ subsumiert werden, die als Antwortmöglichkeiten keine frei formulierten Texte enthalten. Darunter zählen neben den bereits oben beschriebenen „Multiple-Choice-Tests“ auch Ergänzungsaufgaben. Ergänzungsaufgaben bestehen aus einem Satz, in dem bestimmte wichtige Wörter, Zahlen etc. ausgelassen sind. An ihrer Stelle befindet sich eine Lücke, die auszufüllen ist. Davon zu unterscheiden sind Prüfungsformate, bei denen Prüfungskandidaten die Antworten selbst und frei formulieren müssen (vgl. Brauns/Schubert 2008: 93f.).

4.1 Einsatzgebiete und Formen von MC-Tests

Vor- und Nachteile von MC-Tests

Die Vorteile von MC-Tests liegen vor allem in der hohen Kosteneffizienz bei der Prüfung großer Studierendenzahlen und den günstigen Eigenschaften von MC-Tests als Messinstrument von Lernleistungen[1]. Zudem lassen sich Lernleistungen unterschiedlicher Komplexitätsniveaus mit MC-Tests prüfen – von der Abfrage von Faktenwissen bis zum Wissenstransfer (vgl. Nitko 1983: 193).

Vorteile:

- Kosteneffizienz
- Prüfung unterschiedlicher Komplexitätsgrade möglich

Damit diese Vorteile realisiert werden können und MC-Tests gültige und zuverlässige Ergebnisse über die Leistung der Studierenden liefern, sind qualitativ hochwertige MC-Fragen und deren sorgfältige Zusammenstellung zu einem MC-Test notwendig (vgl. Brauns/Schubert 2008: 94). Als Nachteile sind daher der hohe Zeitaufwand für die Prüfungserstellung und die relativ schwierige Konstruktion von Prüfungsfragen – gerade für komplexere Wissensniveaus – zu beachten (vgl. Brauns/Schubert 2008: 94f.). Zudem wird kritisiert, dass eine Beschränkung auf MC-Tests den Studierenden wenig Möglichkeit zum Erwerb schriftlicher Ausdrucksfähigkeit bietet.

Nachteile:

- Hoher Vorbereitungsaufwand
- Umfassende Kenntnisse der Fragebogenkonstruktion nötig
- Wenig Möglichkeit zur schriftlichen Ausdrucksfähigkeit

[1] vgl. hohe Durchführungs- und Auswertungsobjektivität, Validität und Reliabilität sind überprüfbar.

Die Entwicklung qualitativ hochwertiger MC-Tests verlangt daher von den Prüfenden umfassende Kenntnisse im Bereich der Testkonstruktion und-auswertung, die nicht einfach vorausgesetzt werden können. Empfehlungen wie diese können allerdings nur auf zentrale Problempunkte in diesem Bereich hinweisen. Sie können den Besuch von Schulungsangeboten nicht ersetzen.

➔ *Die Carl von Linde-Akademie bietet eine Vielfalt an Kursen an, unter anderem den vierstündigen Workshop „Kompetent prüfen mit Multiple-Choice-Aufgaben“.*

Anwendungsbereiche

Prüfungsformat anhand der Lernergebnisse wählen

Grundsätzlich gilt, dass die Wahl des Prüfungsformats von den zu messenden Lernergebnissen eines Moduls bestimmt sein sollte und nicht von der Verfügbarkeit und Kosteneffizienz des jeweiligen Prüfungsinstruments (vgl. Brauns/Schubert 2008: 95).

Unter **Lernergebnissen** wird eine Aussage darüber verstanden, was eine Lernende/ein Lernender nach dem Abschluss eines bestimmten Lernprozesses weiß, versteht und tun kann. Sie werden als Kompetenzen unter Einschluss von Kenntnissen und Fertigkeiten definiert; Lernergebnisse können beschrieben werden als Wissen (die Studierende/der Studierende kennt…), Fertigkeiten (die Studierende/der Studierende beherrscht die Methode x), Qualifikationen (die Person ist befähigt, eine bestimmte Position einzunehmen oder Tätigkeit auszuüben) (vgl. Schermutzki o.J., 5).

Lernziele mittels Lernergebnis-Taxonomien bestimmen

Um die Einsatzmöglichkeiten von MC-Prüfungen beurteilen zu können, ist es notwendig, die Lernergebnisse einer Lehrveranstaltung nach Niveaugraden zu unterscheiden. Ein sinnvolles Hilfsmittel dazu ist die Taxonomie-Tabelle der modifizierten Bloom´schen **Lernergebnis-Taxonomie** von Anderson/Krathwohl (vgl. Anderson/Krathwohl 2001). Die Tabelle differenziert Lernergebnisse nach sechs Erkenntnisdimensionen. Sie ermöglicht die Zuordnung von Lehrveranstaltungen nach ihren Lernergebnissen zu verschiedenen Wissens- und Erkenntnisniveaus. Mit den dargestellten Kategorien auf Ebene des Wissens und des Erkenntnisprozesses lassen sich die geforderten kognitiven Leistungen einordnen und damit die Anforderungen an die Studierenden transparent machen.

➔ *Eine ausführliche Beschreibung mit konkreten Beispielen zum Thema Lernergebnisse und Taxonomien finden Sie in der Handreichung „Wegweiser zur Erstellung von Modulbeschreibungen“.*

Kognitive Prozessdimensionen – Arten der Wissensanwendung					
Erinnern Relevantes Wissen aus dem Langzeitgedächtnis abrufen	**Verstehen** Bedeutung und Relevanz von Wissen erkennen und herstellen, indem bspw. neues und altes Wissen verknüpft wird	**Anwenden** Bestimmte Verfahren in bestimmten Situationen verwenden	**Analysieren** Gliederung eines Materials in seine konstituierenden Teile und Bestimmung ihrer Interrelation und/oder Relation zu einer übergeordneten Struktur	**Bewerten** Urteile anhand von Kriterien und Standards fällen	**Entwickeln** Elemente zu einem neuen, kohärenten, funktionierenden Ganzen zusammenführen/reorganisieren
Bspw. 1. und 2. Hauptsatz der Thermodynamik wiederaufrufen	Bspw. Zusammenhang zwischen den Hauptsätzen der Thermodynamik und unterschiedlichen Wärme-Kraft-Maschinen erläutern	Bspw. den 1. und 2. Hauptsatz der Thermodynamik auf den Dieselmotor anwenden	Bspw. Einzelne Elemente einer Wärme-Kraft-Maschine unterscheiden und die Beziehung der Elemente untereinander erkennen	Bspw. Unterschiedliche Arten von Wärmeabfuhr in Bezug auf ihre Nutzleistung untersuchen und vergleichen	Bspw. Eine Wärme-Kraft-Maschine bzgl. Abwärmenutzun-gen in Produktionsanlagen optimieren
MC-Tests prinzipiell möglich					
MC-Tests-empfehlenswert				**MC-Tests nicht zu empfehlen**	

Die Bandbreite mit MC-Tests prüfbarer Lernergebnisse reicht prinzipiell von der Reproduktion bzw. dem Erinnern von Faktenwissen bis zur Entwicklung neuer Problemlösungen unter Anwendung von Problemlösungsstrategien. Zu beachten ist allerdings, dass die Entwicklung der Prüfungsfragen mit zunehmenden Komplexitätsgrad des Wissens und der kognitiven Prozesse schwieriger wird. Für die Anwendung von MC-Prüfungen ist daher eine **Beschränkung auf die kognitiven Prozesse des Erinnerns, Verstehens, Anwendens und Analysierens** empfehlenswert (vgl. Williams/Haladyna 1982: 166ff.).

MC-Tests bei Lernzielen auf Ebene des Erinnerns, Verstehens, Anwendens und Analysierens einsetzen.

Fragetypen

Abhängig vom angestrebten Lernergebnis können **unterschiedliche Fragetypen** sinnvoll sein.

Folgende Tabelle stellt die gebräuchlichsten Typen und deren Eignung vor (vgl. Krebs 2002: 192):

Fragetypen		
1. Beste-Antwort-Typen: **Zu jeder Frage wird die Anzahl der auszuwählenden Antworten angegeben**		**Eignung:**
Typ A (positiv)	Auf eine Frage folgt eine bestimmte Anzahl an Wahlantworten, aus denen die einzig richtige oder beste Antwort auszuwählen ist.	Ist der Standardtyp bei MC-Tests und hat sich international auch unter messtechnischen Gesichtspunkten bewährt.
Typ A (negativ)	Auf eine Frage folgt eine bestimmte Anzahl an Wahlantworten, aus denen die Ausnahme bzw. die am wenigsten zutreffende Antwort zu wählen ist.	Ist in den (eher seltenen) Fällen geeignet, in denen die Kenntnis einer wichtigen Ausnahme entscheidend ist.
Typ B (Zuordnung)	Es wird eine bestimmte Anzahl von Wahlantworten vorgegeben. Dann folgen mehrere Aufgaben, zu denen jeweils die passenden Lösungen aus den Wahlantworten angegeben werden müssen.	Es sollte von der Breite und Bedeutung des Themas her entschieden werden, ob dieser Typ B notwendig ist, oder Typ A ausreichend wäre.
Typ Kprim (Mehrfachauswahl[2])	Es wird eine Anzahl an Wahlantworten vorgegeben, aus denen die besten Antworten auszuwählen sind, wobei nicht angegeben wird, wie viele Lösungen richtig sind.	Der Typ ist für Problemstellungen geeignet, bei denen es mehrere wichtige Optionen gibt, die sich deutlich von anderen abheben.
2. Richtig/Falsch-Typen: **Für jede einzelne Antwort auf eine Frage muss eine ja/nein-Entscheidung getroffen werden**		**Eignung:**
Typ Kprim/K´ (Entscheidung richtig/falsch)	Auf eine Frage wird eine bestimmte Zahl an Wahlantworten angegeben, wobei für jede entschieden werden muss, ob sie richtig oder falsch ist.	Der Typ eignet sich für Sachverhalte, bei denen mehrere Aspekte bedeutsam sein können, bzw. für ein Problem, zu dessen Lösung mehrere Elemente gehören können. Messtechnisch beurteilt, ist dieser Typ häufig problematisch, da eine einzige nicht-funktionierende Teilantwort die gesamte Frage zu Fall bringen kann.
Typ E (Kausale Verknüpfung)	Zwei Aussagen sind durch eine Weil-Verknüpfung verbunden. Zunächst sind die Aussagen unabhängig voneinander als richtig oder falsch zu beurteilen. Wenn beide richtig sind, ist die Weil-Verknüpfung zu beurteilen.	Der Typ eignet sich für Wissensgebiete, in denen kausale Zusammenhänge bedeutsam sind. Messtechnisch ist dieser Typ allerdings insofern problematisch, als Kausalitäten selten eindeutig als richtig/falsch zu beurteilen sind.

[2] Der Fragetyp Kprim (Mehrfachauswahl) ist an dieser Stelle der Vollständigkeit halber angeführt, ist aber als Prüfungsformat an der TUM nicht mehr zulässig.

➔ Beispiele zur Formulierung von Beste-Antwort-Typen:

Beispiel Typ A (positiv):

Auf wie viele Arten kann ein König auf einem 8x8-Schachbrett von der linken unteren Ecke in die rechte obere Ecke ziehen, wenn er dabei pro Zug entweder ein Feld nach rechts, ein Feld nach oben oder ein Feld (diagonal) nach rechts-oben ziehen darf?

1) 48639
2) 48640
3) 48641
4) 48642

Beispiel Typ A (negativ):

Welche der genannten Krankheiten gehört nicht zu den Hauptgruppen psychischer Störungen nach dem ICD-10?

1) Depression
2) schizoaffektive Psychose
3) Annorexie/Bulimie
4) Verhaltensauffälligkeiten mit körperlichen Störungen
5) Intelligenzminderung

Beispiel Typ B (Zuordnung):

Lesen Sie die folgenden Interventionstechniken in Gesprächen und ordnen Sie sie den unten genannten Situationen zu. Welche Interventionstechniken im Gespräch sind am angemessensten in folgenden Situationen?

1) Aktiv zuhören
2) Rapport herstellen
3) Zusammenfassen
4) Verbalisieren emotionaler Erlebnisinhalte
5) Konfrontieren mit Diskrepanzen in Klientenaussagen

a) ***Klientin berichtet ausführlich und offen über ihre Problemsituation***
1) 2) 3) 4) 5)

b) ***Klientin wirkt im Gespräch ängstlich und scheu***
1) 2) 3) 4) 5)

c) ***Klientin berichtet über starke Ambivalenzen***
1) 2) 3) 4) 5)

Beispiel Typ Kprim (Mehrfachauswahl):

Welche Aussage(n) über die Eindeutigkeit von Werkstoffwerten treffen zu?

1) Versuche in denen Beanspruchungszeiten ermittelt werden, werden als Dauer- und Langzeitversuche ausgeführt.
2) Versuche in denen Spannungs-Dehnungs-Spielzahlen ermittelt werden, werden als Kurzzeitversuche durchgeführt.
3) Versuchsbasierte Werkstoffkennwerte unterliegen keiner Streuung.
4) Werkstoffe lassen sich mit modernen Produktionsverfahren mit identischer Qualität herstellen.
5) Faktoren wie Korngröße, Risse und Verunreinigungen spielen eine große Rolle.

➔ Beispiele zur Formulierung von <u>Richtig-Falsch-Typen</u>:

Beispiel Typ Kprim/K´ (Vierfache Entscheidung richtig/falsch):

Treffen die folgenden Aussagen zum Produktentwicklungsprozess zu?

1)	*Der Entwicklungsprozess umspannt alle Aufgaben, ausgehend von der Entwicklungsaufgabe bis zum Produkt (Baustruktur).*	*ja (trifft zu) / nein (trifft nicht zu)*
2)	*Nach der Festlegung des Funktionsmodells können Prinziplösungen generiert werden.*	*ja (trifft zu) / nein (trifft nicht zu)*
3)	*Nach der Festlegung des Funktionsmodells können Gestaltlösungen generiert werden.*	*ja (trifft zu) / nein (trifft nicht zu)*
4)	*Baumodelle bilden die Grundlage für fertigungs- und montagetechnische Lösungen.*	*ja (trifft zu) / nein (trifft nicht zu)*

Beispiel Typ E (Kausale Verknüpfung):

a) Ein erfolgreicher Lehrabschluss erhöht die Chancen auf dem Arbeitsmarkt,

<u>weil</u>

b) die Unternehmen bei der Stellenbesetzung Wert auf eine gute Ausbildung legen.

1) a) gilt, <u>weil</u> b) gilt
2) a) gilt und b) gelten
3) a) gilt, b) gilt nicht
4) a) gilt nicht, b) gilt
5) a) und b) gelten nicht

Da die Gruppe der Richtig/Falsch-Typen Wahlantworten verlangt, die eindeutig als richtig bzw. falsch identifizierbar sind, werden diese Fragetypen bei komplexeren Lernergebnissen messtechnisch problematisch und eignen sich daher überwiegend für die Prüfung auf einem niedrigen kognitiven Niveau (bspw. Erinnern von Faktenwissen). Aus diesem Grund sind Beste-Antwort-Typen bei der Konstruktion von MC-Prüfungen eher zu empfehlen (vgl. Krebs 2002: 5).

Richtig/Falsch-Typen für die Prüfung von Faktenwissen

Beste-Antwort-Typen bei komplexeren Lernergebnissen einsetzen

➔ *Die zu messende Fähigkeit sollte die Wahl des Messinstruments bestimmen, d.h. vorab sollte klar sein, was geprüft werden soll, um zu entscheiden, welche Frageform sinnvoll ist. Dies spielt für die Fragebogenkonstruktion insgesamt eine wesentliche Rolle.*

4.2 Fragebogenkonstruktion

Bei der Fragebogenkonstruktion geht es zum einen um die Fragenentwicklung, d.h. Formulierung der einzelnen MC-Fragen und zum anderen auch um die Fragebogengestaltung, d.h. die Auswahl und Zusammenstellung der Einzelfragen zum Gesamttest.

Entwicklung von Blueprints:

Bei der Fragebogengestaltung für eine gültige Leistungsmessung ist folgendes zu beachten: Sollen die Ergebnisse eines Tests einen Nutzen für den Prüfer und auch den Studierenden zur Einschätzung der generellen Studienleistung haben, so müssen die Prüfungsergebnisse gültige und zuverlässige Schlüsse auf die Leistung insgesamt ermöglichen. Wenn die einzelnen Prüfungsfragen das eigentlich angestrebte Lernergebnis nicht hinreichend gut repräsentieren, wären Fehlschlüsse die Folge (vgl. Krebs 2008: 2).

Blueprints vor der Fragenkonstruktion entwickeln

Um sicherzustellen, dass die Zusammenstellung der Prüfungsfragen eines MC-Tests den zu prüfenden Lerngegenstand gut widerspiegelt, empfiehlt sich die Entwicklung von **Blueprints**. Es handelt sich hierbei um ein Themenraster mit allen relevanten Prüfungsinhalten in Form einer zweidimensionalen Matrix, bei der zu jedem Thema bspw. der Prüfungsinhalt das angestrebte Wissensniveau bestimmt und die Prüfungsthemen nach Relevanz gewichtet werden (vgl. Krebs 2008: 4).

Beispiel:
Blueprint MW erstes Semester Bachelor

	Dimension 2: Lernniveau (d.h. angestrebtes Wissensniveau)						**Anzahl Fragen/ Gewich-tung**
	Erinnern	Verstehen	Anwen-den	Analysieren	Bewer-ten	Entwi-ckeln	
Dimension 1: Prüfungs-inhalt 1. System-denken	***4***	***3***					***9,9%***
2. Werkstoff-eigenschaf-ten	***4***	***3***					***10,9%***
3. Ferti-gungs-ver-fahren	***14***	***10***					***34,8%***
4. Bauteilge-staltung	***14***	***11***					***36,2%***
5. Festig-keitsberech-nung	***3***	***3***					***8,2%***
Anzahl Fra-gen	***39***	***30***					**100%**

Ein Blueprint sollte in folgenden drei Schritten entwickelt werden (vgl. Flateby o.J.: 12):

1) Auflistung aller in einer Lehrveranstaltung relevanter, d.h. zu vermittelnder Themen bzw. Lerninhalte.
2) Bestimmung des Lernniveaus, das die Studierenden für jedes der relevanten Themen erreichen sollen (vgl. „Wegweiser zur Erstellung von Modulbeschreibungen“).
3) Gewichtung der Themen nach ihrer Bedeutung, indem bestimmt wird, welchen %-Anteil ein Thema am gesamten Prüfungsumfang einnehmen soll. Abhängig von der geplanten Gesamtzahl an Fragen ergibt sich somit die Anzahl an Fragen, die zu einem Themenbereich gestellt werden sollen.

Sind die Lerninhalte und Lernniveaus einer Lehrveranstaltung mithilfe eines Blueprints angemessen erfasst, ist eine wesentliche Voraussetzung für eine gültige Leistungsmessung geschaffen. Zudem lassen sich durch die Konstruktion „guter" Antwortalternativen (Distraktoren) auch innerhalb der einzelnen Lernniveaus die Schwierigkeitsstufen differenzieren. Wesentlich ist allerdings auch eine an diesen Festlegungen orientierte Fragenentwicklung.

Formulierung von Aufgabenstamm und Antwortoptionen:

Ziel der Fragenformulierung ist es, die Frage in einer Weise zu stellen, dass sie von „wissenden" Studierenden ohne große Schwierigkeiten beantwortet werden kann und dabei die Wahrscheinlichkeit möglichst gering ist, dass die „unwissenden" Studierenden die richtige Antwort durch bloßes Erraten finden können.

Die Form von MC-Fragen folgt einer einheitlichen Grundstruktur. So setzen sich MC-Fragen aus einem **Aufgabenstamm**, in dem die Aufgabe erläutert bzw. die Frage gestellt wird, und mehrere **Antwortoptionen** zusammen, von denen im Fall von Single-Choice nur eine richtig ist und einige falsch sind. Die falschen Antwortoptionen werden als Distraktoren bezeichnet.

Abb. : Grundstruktur einer MC-Frage

Aufgabenstamm	Problembeschreibung und Fragestellung
Antwortoptionen	Alternative 1
	Alternative 2
	…
	Alternative n

Um die oben genannte Zielsetzung zu erreichen, ist die Beachtung folgender Formulierungsregeln für Aufgabenstamm und Antwortoptionen zu empfehlen.[3]

Empfehlungen für den Aufgabenstamm:

Für den Aufgabenstamm sind zwei Formulierungsvarianten möglich: Der Aufgabenstamm kann als Frage formuliert werden, so dass die Alternativen die Antworten auf diese Frage darstellen. Der Aufgabenstamm kann auch als unvollständiger Satz formuliert werden, der dann von den Alternativen in verschiedener Weise vervollständigt wird (vgl. Jakobs 2005: 3).

[3] Für einen Überblick und die empirische Prüfung zu Regeln der MC-Formulierung vgl. Haladyna/Downing 1989 und Haladyna/Downing/Rodriguez 2002.

Beispiel 1:
Was ist die Hauptfunktion des M. glutaeus maximus?

1) Streckung des Oberschenkels
2) Innenrotation des Oberschenkels
3) Neigung des Beckens zur Spielbeinseite
4) Neigung des Beckens zur Standbeinseite

Beispiel 2:
Wenn eine Gruppe ein Brainstorming durchführt, muss sie besonders darauf achten, …

1) … die Gedanken und Vorschläge in geordnete Bahnen zu lenken
2) … die vorgebrachten Ideen sofort nach Wichtigkeit zu ordnen
3) … das Aufkommen von Leistungsdruck und Kritik zu vermeiden
4) … sich voll auf den Weg zum Ziel zu konzentrieren

Für beide Varianten ist wesentlich, dass in der Formulierung der Aufgaben- bzw. Fragestellung, alle zur Beantwortung relevanten Informationen enthalten sind und sich nicht erst in den Antwortoptionen wichtige Angaben finden.
Um die Lesezeit einer Frage zugunsten der Denkzeit möglichst kurz zu halten, empfiehlt es sich in diesem Zusammenhang auch, den Aufgabenstamm möglichst ausführlich, die Antworten dagegen aber kurz und übersichtlich zu gestalten.

Alle relevanten Informationen in der Aufgabenformulierung nicht in den Antwortoptionen

Zudem sollte sich die Schwierigkeit einer Frage aus der Komplexität des Aufgabeninhalts ergeben und nicht aus einer künstlichen Verkomplizierung durch Schachtelsätze, doppelte Verneinungen, überflüssige Informationen o.ä..

Durch einfache, klare und positive Formulierungen und die Beschränkung auf gebräuchliche Fremdwörter, Fachausdrücke und Abkürzungen kann sichergestellt werden, dass mit der Frage das Fachwissen der Studierenden geprüft wird und nicht deren Textverständnis (vgl. Schmidts/Lischka 2001: 7).

Einfache, klare und positive Frageformulierungen verwenden

Idealerweise sollte die sog. „cover-the-options-rule“ gelten, nach der eine Frage dann gut formuliert ist, wenn sie auch ohne Antwortvorgaben beantwortet werden kann (Nitko 1983: 197).

Frage formulieren, die auch ohne Antwortoptionen beantwortet werden können.

Negativbeispiel:
Innovationen sind für die Unternehmung sehr wesentlich.

1) Was neu ist für die Unternehmung, ist immer neu für den Markt!
2) Nur Pionierunternehmungen sind erfolgreich!
3) Marktführende Unternehmungen tätigen mehr Innovationen als Außenseiter!
4) keine der vorstehenden Antworten ist richtig
5) alle der vorstehenden Antworten sind richtig

(Aufgabenstamm enthält keine eindeutige Frageformulierung)

Im Aufgabenstamm ist zudem anzugeben, welche Bedingungen die richtige Antwort erfüllen muss. So können sich richtige Antworten von den Antwortalternativen (Distraktoren) auf zwei Weisen unterscheiden (vgl. Jacobs 2004: 3):

- Die richtige Antwort ist wahr, die Distraktoren sind falsch („true o. correct answer form“)
- Die richtige Antwort ist die beste Auswahl unter allen Alternativen („best answer form“)

Zusammenfassend ist die Einhaltung folgender Formulierungsregeln zu empfehlen.

➔ *Der Aufgabenstamm sollte einfach, klar und positiv formuliert sein, alle für die Beantwortung der Frage erforderlichen Informationen enthalten und in einer Weise formuliert sein, dass die Frage beantwortet werden kann ohne die Antwortoptionen lesen zu müssen.*

Empfehlungen zu Antwortoptionen:

Im Rahmen von MC-Tests werden den Studierenden eine Reihe von Antwortoptionen als Lösungen für die im Aufgabenstamm dargestellte Problemstellung angeboten. Neben im Fall von Single-Choice nur einer richtigen bzw. besten Antwort gehören dazu immer auch falsche Antwortoptionen, die sog. Ablenker bzw. Distraktoren. Für die Schwierigkeit einer Frage sind vor allem diese Ablenker entscheidend. Sie sollen den unwissenden Studierenden plausibel erscheinen, damit sie von der richtigen Antwort abgelenkt werden, dürfen die wissenden Studierenden aber nicht verwirren (vgl. Jakobs 2005: 9).

Die falschen Antwortoptionen sind für die Schwierigkeit wichtiger als die richtigen.

Damit das Finden der richtigen Antwort ausschließlich vom Wissen des Studierenden abhängt und die Ratewahrscheinlichkeit möglichst minimiert wird, ist die Einhaltung einer Reihe von inhaltlichen sowie formalen und sprachlichen Regeln empfehlenswert.

(1) Inhaltliche Kriterien

- Die Antwortoptionen sollten **inhaltlich homogen** sein, d.h. aus demselben Gegenstandsbereich/Themenbereich stammen. Eine von den übrigen Antwortkategorien deutlich abweichende Antwortvorgabe ist auch für den unwissenden Studierenden als falsche Antwort leicht identifizierbar. Je homogener die Antworten, desto schwieriger die Frage (vgl. Krebs 2002: 16; Nitko 1983: 203).

Alle Antwortoptionen aus demselben Themenbereich wählen.

Negativbeispiel:
Sie arbeiten unter Windows. Beim Ausdrucken einer Seite mit einer komplexen Grafik auf einem Laserdrucker erscheint nur der obere Teil der Figur. Welches ist die wahrscheinlichste Ursache?

1) *Die Auflösung des Laserdruckers ist zu klein.*
2) *Die interne Speicherkarte der Nikon D3 ist zu klein.*
3) *Linux ist nicht in der Lage, komplexe Figuren zu drucken.*
4) *Laserdrucker sind allgemein für das Drucken komplexer Figuren ungeeignet.*
5) *Der PC befindet sich in der Küche.*

(Antwortoptionen sehr heterogen)

- Alle Antwortalternativen – die Distraktoren - sollten sich **plausibel auf die Fragestellung beziehen** bzw. in einem klar nachvollziehbaren Verhältnis zur Fragestellung stehen („functional alternatives“). Hinsichtlich der Distraktoren bietet es sich an, häufige Fehlmeinungen, falsche Konzepte oder veraltete Ansichten zu verwenden. Distraktoren, die unplausibel, trivial bzw. völlig unsinnig sind, können auch von unwissenden Studierenden als Antwort ausgeschlossen werden. Folge ist, dass trotz vieler Antwortalternativen die Ratewahrscheinlichkeit steigt. Um plausible Distraktoren für eine Frage zu ermitteln, kann diese bspw. im Rahmen einer Veranstaltung als offene Frage gestellt werden. Häufige Fehler können dann als plausible Distraktoren für den MC-Test verwendet werden (vgl. Krebs 2002: 16; Jacobs 2005: 9).

Alle Antwortoptionen in plausiblen Bezug zur Fragestellung setzen.

Negativbeispiel:
Von der Führungskonzeption „management by objectives“ ausgehend, bestehen die wesentlichen Unterschiede zu der Führungskonzeption „management by exception“:

1) *Das Prinzip der "Führung durch Zielvereinbarung" berücksichtigt in gleicher Weise ökonomische Belange des Betriebes wie humane Interessen der Mitarbeiter. Es betont die Zusammenarbeit im Team. Der Vorgesetzte beteiligt seine Mitarbeiter am Prozeß der Zielbildung und der Planung. Auch die Analyse der Abweichungen des Ist vom Soll führt der Vorgesetzte mit den Mitarbeitern gemeinsam durch (Selbstkontrolle).*
2) *Führung durch Ausnahmeentscheidungen. Eingriff im Ausnahmefall. Eine in den Betrieben verwendete Methode. Beispiele finden sich in vielen Richtlinien und Arbeitsanweisungen in der Form, dass die Entscheidungen und Abläufe programmiert und in die Verantwortlichkeit der sachbearbeitenden Stelle gegeben werden. Ausnahmesituationen bleiben Leistungseinheiten vorbehalten.*
3) *Keine der obigen Antworten (1+2) ist richtig.*
4) *Alle der obigen Antworten (1+2) sind richtig.*

(Antwort 3 und 4 schließen die anderen Alternativen aus)

- Bei der Auswahl der Antwortalternativen ist darauf zu achten, dass im Fall von Single-Choice nur eine richtige Antwort gegenüber vorhandenen Distraktoren **eindeutig die beste Antwort** ist. Eine Frage wird umso schwieriger, je näher richtige und falsche Antworten beieinander liegen (vgl. Krebs 2002: 17).

Die richtige Antwort müssen gegenüber den Distraktoren eindeutig die beste sein.

Negativbeispiel:
1) A und B unterscheiden sich wenig
2) A und B sind in ihrer Beschaffenheit ähnlich
3) A und B unterscheiden sich kaum

(Antwortoptionen liegen zu nah beieinander)

- Antwortalternativen sollten möglichst kurz sein und **nur eine inhaltliche Aussage** (homogen gestaltet bzw. in die gleiche Kategorie gehörend) enthalten (vgl. Krebs 2002: 17).

Nur eine inhaltliche Aussage pro Antwort

Positivbeispiel:
Es sei M die Menge aller U-Bahnhaltestellen in München. Wir definieren eine Relation R auf M durch: (a,b) R : b kann von a aus in 15 Minuten mit der U-Bahn erreicht werden. Dabei wird die Wartezeit in Punkt a nicht berücksichtigt.
1) R ist eine reflexive Relation
2) R ist eine symmetrische Relation
3) R ist eine transitive Relation

(Antwortalternativen homogen, beziehen sich alle auf den gleichen Sachverhalt)

- Jede Antwortalternative sollte **klar unterscheidbar** sein und es sollte möglichst vermieden werden, sich überschneidende Antwortoptionen zu formulieren. Ist eine Antwortalternative ein Teilbereich einer anderen Option, so gibt dies Hinweise auf die richtige Antwort. So müssen zwei ähnliche Alternativen falsch sein, wenn es nur eine richtige Antwort gibt (vgl. Jacobs 2005: 8f.).

Sich inhaltlich überschneidende Antwortoptionen vermeiden

Negativbeispiel:
Warum muß man „wirtschaften“ ?
1) Es ist gesetzlich vorgeschrieben, dass alle Wirtschaftssubjekte einer Tätigkeit in der Wirtschaft nachgehen müssen.
2) Die Notwendigkeit des Wirtschaftens ergibt sich aus den unbegrenzten Bedürfnissen der Menschen und der Knappheit der zur Bedürfnisbefriedigung benötigten Mittel.
3) Gäbe es alle zur Bedürfnisbefriedigung geeigneten Mittel im Überfluss, brauchte nicht gewirtschaftet zu werden. Doch die vorhandenen Mittel sind relativ knapp und sie können jeweils zur Befriedigung verschiedener Bedürfnisse eingesetzt werden. Man muss deshalb auswählen, für

welchen Zweck die begrenzt zur Verfügung stehenden Mittel verwendet werden, um ein möglichst günstiges Ergebnis zu erreichen.

(die dritte Antwortalternative 3 enthält Antwortalternative 2)

- Von der Verwendung der Optionen **„alle genannten Alternativen" bzw. „keine der genannten Alternativen"** und der Art **„sowohl A als auch C"** ist abzuraten. Sind diese Alternativen als Distraktoren gedacht, dann lassen sie sich leicht identifizieren, wenn nur eine Alternative als falsch bzw. richtig erkannt wird. Die Option „keine der genannten Antworten" ist dagegen nur bei der „true/correct answer form" sinnvoll (vgl. Halaayna et al. 2002: 319f.).

Antwortoptionen, die sich auf mehrere Alternativen beziehen, vermeiden.

Negativbeispiel:

1) a und b, aber nicht c
2) nur a, sofern c
3) a oder b, wenn nicht c
4) …

- Von Alternativen, die **komplexe Entscheidungen** verlangen und ein hohes Maß an Logik voraussetzen – bspw. „nur a, sofern c"; „a oder b, wenn nicht c, doppelte Verneinungen" - ist ebenfalls abzuraten. Auf diese Weise lässt sich der Schwierigkeitsgrad der Frage zwar erhöhen, es reduziert sich allerdings die Validität der Lernergebniserfassung (vgl. Jacobs 2005: 10).

Antwortoptionen die komplexe Entscheidungen verlangen vermeiden.

Negativbeispiel:

Welche Aussage trifft <u>nicht</u> zu?
1) MC-Prüfungsfragen können kein Anwendungswissen prüfen
2) MC-Prüfungsfragen können kein praktisches Wissen prüfen
3) MC-Prüfungsfragen können kein …
4) MC-Prüfungsfragen sind nicht geeignet für die Abfrage von …

(komplexe Entscheidung durch doppelte Verneinung)

(2) Formale und sprachliche Kriterien

Die Einhaltung der folgenden formalen und sprachlichen Kriterien ist zur Vermeidung von unbeabsichtigten Lösungshinweisen (cues) empfehlenswert. Hierbei handelt es sich um formale, logische oder sprachliche Hinweise, die es MC-Test-erfahrenen Studierenden erlauben, auch ohne Fachwissen die richtige Antwort zu identifizieren bzw. Antworten als distraktorenverdächtig auszuschließen und damit die Ratewahrscheinlichkeit zu erhöhen. Ursache dafür ist zumeist, dass die Prüfer der Formulierung der richtigen Antwortoptionen mehr Aufmerksamkeit als der Bestimmung und Formulierung von Distraktoren schenken (vgl. Krebs 2002: 17, 20; Nitko 1983: 207f.).

- Alle Antwortoptionen müssen **grammatikalisch** zum Aufgabenstamm passen. Häufig wird dies nur bei der richtigen Antwort beachtet, so dass sich die falschen Antworten leicht ausschließen lassen (vgl. Nitko 1093: 205).

Alle Antwortoptionen müssen grammatikalisch zum Aufgabenstamm passen.

Negativbeispiel:
Anter ist eine …
1) …Legierung
2) …Konglomerat
3) …Verbrennungsrückstand
4) …Spaltprodukt

(nur Antwort Nr. 1 passt grammatikalisch zum Aufgabenstamm)

- Die Distraktoren sollten in ihrer **Länge** und in ihrem **Differenzierungsgrad** der richtigen Antwort entsprechen. Meist wird die richtige Antwort umfangreicher ausformuliert als die falschen Antworten (vgl. Jacobs 2005: 8; Nitko 1093: 211).

Alle Antwortoptionen in gleicher Länge und Differenzierungsgrad formulieren.

Negativbeispiel:
Bei der Fermierung von Anter mit saurem Gor …
1) findet eine Abkühlung statt
2) entsteht unter der Bedingung einer leichten Erwärmung Anterit im pH-Bereich 2.8-3.2
3) wird Ogl4 freigesetzt
4) entsteht Fermatin

(Antwortalternative Nr. 2 ist die umfangreichste)

- **Ähnlichkeiten von Wörtern** (verbale Assoziationen bzw. Wortwiederholungen) im Aufgabenstamm und in der korrekten Lösung sollten vermieden werden (vgl. Krebs 2002: 19).

Ähnlichkeiten bei Aufgabenstamm und richtiger Antwort vermeiden.

Negativbeispiel:
Welches ist das Hauptmerkmal des <u>Kognitives</u> Rigiditäts-Syndrom?
1) ein erhöhter Ferminspiegel im Plasma
2) zyklische postprandiale Alpträume
3) häufige Versteifungen der Nackenmuskulatur
4) eine reduzierte Beweglichkeit im <u>kognitiven</u> Bereich

(Wortwiederholung im Fragestamm und der Antwortalternativen)

- Eine Formulierung der korrekten Antwort ähnlich der **Lehrbuchformulierungen** ist zu vermeiden, da diese Studierenden bekannt vorkommen und somit als richtige Antworten erscheinen, auch wenn der Sachverhalt nicht verstanden wurde (vgl. Jacobs 2005: 8; Nitko 1093: 210).

Keine Lehrbuchformulierungen verwenden

Negativbeispiel:
Eine 55-jährige Patientin stellt sich wegen Schmerzen und einer Hauteinziehung im oberen äußeren Quadranten der linken Brust vor. Unter der Haut tastet man eine derbe Verdickung. Welches weitere Vorgehen ist nun angezeigt ?
1) Kontrollmammographie in 6 Monaten
2) Extripation der Verdickung <u>und</u> histologische Untersuchung
3) Lokale Behandlung mit Rotlichtbestrahlung
4) Antibiose
5) Östrogengabe

(„und“ als Hinweis einer Lehrbuchformulierung)

- **„Absolute“ Begriffe** sprich uneingeschränkte Begriffe wie beispielsweise „niemals“, „immer“, „alle“, „kein“, „nur“ usw., in der Formulierung lassen Ablenker/Distraktoren vermuten, während moderate Begriffe („manchmal“, „möglicherweise“, „gewöhnlich“) auf die richtige Antwort schließen lassen (vgl. Jacobs 2005: 8).

„Absolute“ Begriffe vermeiden

Negativbeispiel:
Warum ist bei trigoten Quergeln die Axosie-Auftretensrate erhöht?
1) Trigote Quergeln sind <u>nie</u> berop
2) Trigotie führt <u>immer</u> zu Enität
3) Trigote Quergeln sind gehäuft susmin
4) Axosie ist <u>ausschließlich</u> sequid bedingt

(„Absolute“ Begriffe wie immer, nie, ausschließlich usw. vermeiden)

- Auch sollten sog. „**Konvergenz-Cues**“, d. h. Elemente die bereits auf die richtige Antwort hinweisen, vermieden werden. Diejenige Antwort, die die größte Anzahl an gemeinsamen Elementen oder Begriffen mit den anderen Antwortalternativen gemeinsam hat, ist mit hoher Wahrscheinlichkeit die richtige Antwort (vgl. Krebs 2002: 19).

Gemeinsame Elemente und Begriffe bei Antwortoptionen vermeiden.

Negativbeispiel:
Die Abkürzung USL heißt ausgeschrieben?
1) United States Laboratories
2) Uniform Source Language
3) Uniform Source Locator
4) Uniform Starting Label
5) Unique Spaceship Locator

(3xUniform, 2xSource, 2xLocator)

- Damit die **Anordnung der Antwortoptionen** keine Hinweise auf die richtige Lösung gibt, sollten die Alternativen - soweit möglich – nach dem Zufallsprinzip in eine Reihenfolge gebracht werden (aufsteigend, absteigend, alphabetisch). Beispielsweise werden richtige Antworten überdurchschnittlich häufig von den Prüfern in der Mitte der Antwortalternativen platziert. Um diese Tendenz zu vermeiden, sollte die Position der richtigen Lösung und die der Distraktoren entsprechend dem Zufallsprinzip angeordnet werden (vgl. Krebs 2002: 20).

Antwortoptionen nach dem Zufallsprinzip anordnen.

Häufig werden die richtigen Antworten an die Position C oder D gesetzt
(siehe oben)

- Hinsichtlich der **Anzahl der Antwortoptionen** gilt zwar theoretisch, dass die Ratewahrscheinlichkeit mit steigender Zahl der Antwortalternativen sinkt, aber nur wenn die Distraktoren plausibel und gut formuliert sind. Deshalb ist zu empfehlen, immer nur so viele Alternativen zu formulieren, wie sich plausible Distraktoren finden lassen. In der Praxis haben sich 4 Antwortoptionen als brauchbar erwiesen. Studien zeigen allerdings, dass bei nur 3 Alternativen zwar die sog. „Item-Schwierigkeit“ (Verhältnis der von einer Kandidatengruppe bei einem Item erreichten zur maximal möglichen Punktzahl) sinkt, die Trennschärfe der Frage und die Reliabilität des Tests dadurch nicht beeinträchtigt werden (vgl. Brauns/Schubert 2008: 99).

Nur so viele Antworten anbieten, wie sich plausible Optionen finden lassen.

Beispiel:

Sie sind Abteilungsleiter in einem mittleren Unternehmen. Seit einiger Zeit klagen zwei ihrer Gruppenleiter häufig über Arbeitsüberlastung. Welche der nachstehenden Maßnahmen würden Sie als Vorgesetzter bevorzugen?

1) Die Klagen am besten überhören
2) Die Gruppen verkleinern und neue Gruppen bilden
3) In jeder Gruppe einen zusätzlichen Mitarbeiter „für besondere Aufgaben" einstellen
4) Den Gruppenleitern vorschlagen zu prüfen, ob den Mitarbeitern mehr Entscheidungskompetenz übertragen werden kann
5) Keine der Maßnahmen (1-4)

(Antwortalternativen Nr. 1 und 5 sind keine plausiblen Distraktoren)

- Bei der **Anordnung der Fragen** sollten die allgemeinen Regeln der Fragebogengestaltung gelten, nach denen eine Gliederung von einfachen zu schwierigen Fragen erfolgt, Seitenumbrüche innerhalb der Antwortalternativen vermieden und Erläuterungen zur Beantwortung einer Frage gemacht werden sollten. Zudem ist eine übersichtliche grafische Aufbereitung des Fragebogens empfehlenswert. Auch sollte die Fragenreihenfolge keine Lösungshinweise geben, indem eine vorangegangene Frage Informationen zur Lösung der folgenden Frage enthält (vgl. Zimmerman et al. 1990: 5).

Frageanordnung von einfachen zu schwierigen Fragen

4.3 Testauswertung/Prüfungsauswertung/Gesamtauswertung

Für die Auswertung von Multiple-Choice-Tests gibt es mehrere Möglichkeiten die, eingesetzt werden können.

- Bei **Einfachauswahl/Single-Choice** sind **Bonussysteme** üblich. Da nur eine Antwort richtig ist, wird **ein Punkt oder mehr - je nach Gewichtung der Frage - gutgeschrieben**, wenn diese angekreuzt wurde. Ansonsten werden keine Punkte vergeben.

- Der Noten-Punkte-Schlüssel wird durch die rechtlichen Vorgaben bestimmt, welche die **Ratewahrscheinlichkeit** berücksichtigen. Insbesondere wird beachtet, wie viele Punkte ein Studierender durch zufälliges Ankreuzen erhalten kann (Erwartungswert bei Raten).
  Für die beschriebene **Einfachauswahl/ Single Choice** ist die Ratewahrscheinlichkeit abhängig von der **Anzahl n** der angebotenen Antworten und somit **1/n**. Der Erwartungswert ergibt sich dann aus der Punktezahl der Aufgabe geteilt durch Anzahl n der Antworten.

- Ein weiterer wichtiger Aspekt bei der Erstellung von Single-Choice Klausuren ist die **Anzahl der Aufgaben**. Folgende Tabelle zeigt den Zusammenhang zwischen Anzahl der Antworten, Anzahl der Aufgaben und der Wahrscheinlichkeit (in Prozent), dass ein Studierender trotz Raten, die Klausur besteht.

	Single-Choice		
Anzahl Antworten	*zwei*	*drei*	*vier*
5 Aufgaben	50 %	21 %	10,4 %
10 Aufgaben	62,3 %	21,3 %	7,8 %
20 Aufgaben	58,8 %	9,2 %	1,4 %
30 Aufgaben	57,2 %	4,4 %	0,3 %

Im Folgenden wird davon ausgegangen, dass ein Studierender mit 50% der Punkte bestanden hat. Dies ist z.B. der Fall, wenn alle Studierenden einer Klausur bei Aufgaben raten, deren Antwort sie nicht wissen.

- Bei Single Choice-Klausuren mit nur zwei Antwortmöglichkeiten je Aufgabe liegt die Wahrscheinlichkeit, dass ein Studierender besteht (Bestehwahrscheinlichkeit), unabhängig von der Aufgabenanzahl bei mindestens 50%. Das bedeutet, dass nach drei Klausurversuchen 87,5% aller Studierenden bestanden haben. Daher **müssen mindestens drei Antwortmöglichkeiten** angeboten werden.
- Bei Single Choice-Klausuren mit drei Antwortmöglichkeiten je Aufgabe ist die Bestehwahrscheinlichkeit erst mit 35 Aufgaben bei 2%.

5. Empfehlungen zur Organisation und Qualitätssicherung von Multiple-Choice-Tests an den Departments

Um mittels MC-Prüfungen eine möglichst genaue und fehlerfreie Abbildung der Fähigkeiten von Studierenden auf einem Stoffgebiet zu erhalten, sind Maßnahmen der Qualitätssicherung sowohl auf Ebene der einzelnen Fragen als auch auf Ebene des gesamten MC-Tests sinnvoll (Brauns/Schubert 2008: 95). Darüber hinaus sind Verfahren der Qualitätssicherung auch auf Ebene des Departments zu empfehlen.

5.1 Qualitätssicherung auf Fragenebene

Neben den oben genannten Regeln der Fragenformulierung ist nach dem Einsatz der jeweiligen Frage in einem MC-Test eine sog. Itemanalyse sinnvoll. Es handelt sich dabei um die Beurteilung der Messeigenschaften der Frage (Item) anhand der Berechnung zweier Gütekriterien: Item-Schwierigkeit und Item–Trennschärfe (Brauns/Schubert 2008: 95, 101).

Nach dem Einsatz einer Frage deren Schwierigkeit und Trennschärfe ermitteln.

- **Item-Schwierigkeit** erfasst den Anteil der Prüfungsteilnehmer, die eine Frage richtig beantwortet haben, an der Gesamtzahl der Teilnehmer (P=x/n; 0-1). Je höher die „Schwierigkeit“, desto leichter die Aufgabe. Fragen, die alle bzw. kein Studierender beantworten kann, haben keine Trennschärfe. Es ist zu empfehlen, dass bei Prüfungen die Schwierigkeit der Fragen einer Normalverteilung entspricht.
- **Item-Trennschärfe** gibt an, wie gut das gesamte Testergebnis aufgrund der Beantwortung eines einzelnen Items vorhersagbar ist, wie gut also das Item zwischen den leistungsstarken und leistungsschwachen Studierenden differenzieren kann. Die Berechnung erfolgt über Korrelationskoeffizienten zwischen der erreichten Punktzahl des Items und der Gesamtpunktzahl der Prüfung ohne dieses Item (unter 0,3 niedrig, 0,3-0,5 mittel, über 0,5 hoch).

5.2 Qualitätssicherung auf Ebene der MC-Tests

Zur Qualitätssicherung des MC-Tests insgesamt sollte vor dem Einsatz als Prüfungsinstrument ein sog. **Pretest** durchgeführt werden. Dies bedeutet, dass mit dem Messinstrument vorab eine Probeprüfung durchgeführt wird, um folgende Gütekriterien zu prüfen: Validität und Reliabilität (vgl. Krebs 2002: 29).

Vor dem Einsatz des Fragebogens Pretest durchführen.

- **Validität** gibt die Gültigkeit des Messinstruments an, d.h. inwiefern die Prüfung tatsächlich misst, was sie messen soll – ein Learning-Outcome und nicht nur das Textverständnis o.ä..
- Es lassen sich unterschiedliche Aspekte der Validität unterscheiden: Inhaltsvalidität, Kriteriumsvalidität, Konstruktvalidität. Zu empfehlen ist v.a.

die Prüfung der Inhaltsvalidität, d.h. ob die Prüfungsfragen für den zu prüfenden Kompetenzbereich geeignet sind. Hierzu wäre die Beurteilung von Fachexperten einzuholen.

- **Reliabilität** erfasst die Zuverlässigkeit des Messinstruments bzw. Wiederholbarkeit des Messergebnisses. Hierbei ist zu prüfen, inwiefern die Ergebnisse von der spezifischen Auswahl der Fragen abhängen bzw. inwiefern eine alternative Prüfung mit gleicher Anzahl an Fragen, die nach gleichen Kriterien aus dem gleichen Inhaltsbereich formuliert wurden, zu denselben Ergebnissen führt.
- Zur Berechnung werden verschiedene Verfahren eingesetzt: Retest-Methode, Parallel-Test-Methoden oder Konsistenzanalysen, bei denen die mittlere Interkorrelation aller Fragen berechnet wird (vgl. Crombachs alpha oder Kruder-Richardson-Formel).

5.3 Qualitätssicherung auf Department-Ebene

Itempools mit bewährten Fragen anlegen.

Auf Department-Ebene ist die Einrichtung eines **Itempools mit Itemstatistik** zu empfehlen. Haben sich Items als gute Messinstrumente bewährt, sollten sie wieder eingesetzt werden können. Dazu können Itempools in Form von Datenbanken aufgebaut werden, mittels derer Itemstatistiken zu jeder Frage mit der Zugehörigkeit zu Themenbereich und Kompetenzniveau, der Item-Schwierigkeit und -Trennschärfe sowie dem Einsatzdatum usw. geführt werden können. Der wiederholte Einsatz von Fragen stellt insofern eine Sicherung der Fragenqualität dar, als deren Schwierigkeit und Trennschärfe besser beurteilt werden kann.

Um die Schwierigkeit und Trennschärfe von Fragen nicht zu beeinträchtigten, wäre zwar eine Geheimhaltung der Fragen am besten geeignet. Besteht der Fragenpool allerdings aus mehreren tausend Fragen, so ist es für die Studierenden wenig sinnvoll, Fragen anstelle des Prüfungsinhalts zu lernen (vgl. Brauns/Schubert 2008: 99).

Als weiteres wesentliches Element der Qualitätssicherung sollten gerade neue MC-Prüfer vorab Schulungen (Carl von Linde Akademie/ProLehre) erhalten, die eine Prüfungserstellung an konkreten Beispielen erläutern.

➔ *Damit Studierende nicht erst in der Prüfungssituation mit Multiple-Choice-Aufgaben konfrontiert und gegebenenfalls unnötig überfordert werden, ist es zu empfehlen, bereits während dem Semester in Übungen und Vorlesungen ein solches Prüfungsformat anzuwenden und zu üben.*

6. Glossar

Themenfeld Multiple Choice	
Ablenker:	Falschantwort bei einer MC-Frage. Kandidaten mit fehlender Sachkenntnis sollten die richtige Antwort nicht (z.B. aufgrund von Cues) von den Ablenkern unterscheiden können.
Blueprint:	Gewichtetes Inhaltsraster der Prüfungsinhalte, nach dem alle Prüfungen zusammengesetzt werden. Kann eine oder auch mehrere Dimensionen enthalten.
Cue:	Von einem Cue spricht man, wenn in der Art der Frage-oder Aufgabenstellung bereits ein Hinweis auf die richtige Antwort enthalten ist bzw. diese stärker hervorgehoben wird.
Distraktor:	s. Ablenker
Inhaltsvalidität:	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.
Item:	Einzelheit. 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 Wahrscheinlichkeit (p) angegeben.
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.
Konstruktvalidität:	Konstruktvalidität fragt danach, ob Hypothesen, die aus einer Theorie über das zu messen beabsichtigte Konstrukt abgeleitet sind, durch Befunde im Zusammenhang mit den Prüfungsergebnissen gestützt werden.
Korrelations-koeffizient:	Eine statistische Kenngröße, die die Stärke und die Richtung des Zusammenhangs zweier Variablen ausdrückt. Ein Korrelationskoeffizient von +1 drückt eine vollständige direkte Beziehung aus,

	bei 1 ist die Richtung entgegengesetzt und bei 0 besteht keinerlei Beziehung.
Kriteriumsvalidität:	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.
Multiple Choice:	Auswahl einer oder mehrerer korrekter Antworten aus einer vorgegebenen Liste von Antwortoptionen.
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.
Pretest:	Pretest bezeichnet die Qualitätsverbesserung eines Fragebogens vor der Durchführung einer Befragung mittels der Durchführung von Tests.
Ratewahrscheinlichkeit:	Durch zufälliges Ankreuzen erreichbarer Wert, beispielsweise eine bestimmte Punkteanzahl (Erwartungswert bei Raten)
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. In der Empfehlung steht die Reliabilität des Prüfungsinstruments im Vordergrund, meist erfasst in Form des Koeffizienten alpha von Cronbach. Dies gibt Auskunft darüber, wie stark die Ergebnisse von der spezifischen Itemauswahl abhängen
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 Inhaltsvalidität, Kriteriumsvalidität und Konstruktvalidität unterschieden.

7. Literaturempfehlungen

Anderson, L.W. & Krathwohl, D.R. (Eds.) (2001): A Taxonomy of Learning, Teaching and Assessing. A Revision of Bloom's Taxonomy of Educational Objectives. Addison Wesley Longman.

BLK-Projekt Leistungspunktesystem Universität Hannover (2004): Erläuterungen zur Beschreibung und Abstrahierung von intendierten Lernzielen.

Bloom, B.S., et al. (1976) Taxonomie von Lernzielen im kognitiven Bereich. Weinheim und Basel: Beltz, 5. Auflage (engl. Originalausgabe 1956).

Brauns, K. und Schubert S. (2008): Qualitätssicherung von Multiple-Choice-Prüfungen. In: Dany, Sigrid; Szczyrba, Birgit und Johannes Wildt (Hrsg.): Prüfungen auf die Agenda! Hochschuldidaktische Perspektiven auf Reformen im Prüfungswesen. Bielefeld: 92-102.

Bühler, C. (1980): Zweidimensionale Taxonomie von Lernzielen und Inhalten im kognitiven Bereich. Weil.

Burton, S. J. et al. (1991): How to Prepare Better Multiple-Choice Test Items: Guide-lines for University Faculty.

Ebel, R. L., & Frisbie, D. A. (1986): Essentials of educational measurement (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Flateby, T. L. (o.J.): A Guide for Writing and Improving Achievement Tests.

Gronlund, N. E (1998): Assessment of Students Achievement. 6th edition, Boston.

Gronlund, N. E. (1993): How to Make Achievement Tests and Assessments, 5th ed., Allyn and Bacon, Needham Heights, MA.

Haladyna, T.M. & Downing, S.M. (1989): A taxnonomy of multiple-choice item-writing rules. In: Applied Measurement in Education: 37-50.

Haladyna, T.M. & Downing, S.M. (1989b): Vadility of a taxnonomy of multiple-choice item-writing rules. In: Applied Measurement in Education. 2(1) 1989: 51-78.

Haladyna, T.M. & Downing, S.M (1989c): Multiple-Choice Item-Writing Guidelines/Rules/Suggestions/Advice As Derived From 46 Authoritative Textbooks.

Haladyna, T.M.; Downing, S M. & Rodriguez M.C. (2002): A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. In: Applied Measurement in Education. 15(3) 2002: 309-334.

Jacobs, B. (2005): Richtlinien zur Erstellung von einfachen Multiple-Choice-Aufgaben nach Gronlund. http://www.phil.unisb.de/mz/verweise/psych/aufgaben/mcguideline.html.

Krebs, R. (2002):Anleitung zur Herstellung von MC-Fragen und MC-Prüfungen. Institut für Aus-, Weiter- und Fortbildung IAWF; Abt. für Ausbildungs- und Examensforschung AAE. Bern.

Krebs, R. (2008): Multiple Choice Fragen? Ja, aber richtig. Medizinische Fakultät; Institut für Medizinische Lehre IML; Abteilung für Assessment- und Evaluation AAE. Bern.

Müller, F. und Bayer C. (2007): Prüfungen: Vorbereitung – Durchführung – Bewertung. In: Harelka, B.; M. Hammerl und H. Gruber (Hrsg.): Förderung von Kompetenzen in der Hochschullehre. Kröning: 223-238.

Nitko, A. J. (1983): Educational tests and measurement: An introduction. New York

Popham, W. J. (1999): Modern educational measurement: practical guidelines for educational leaders. 3, überarbeitete Auflage.

Rodriguez, M.C. (2005): Three Options Are Optimal for Multiple-Choice Items : A Meta-Analysis of 80 Years of Research. In: Educational Measurement: Issues and Practice. Summer (2005): 3-13.

Roid, G. H., & Haladyna, T. M. (1982): A technology for test-item writing. New York: Academic Press.

Schermutzki, M. (2008): Learning Outcomes – Lernergebnisse: Begriffe, Zusammenhänge, Umsetzung und Erfolgsermittlung. Lernergebnisse und Kompetenzvermittlung als elementare Orientierung des Bologna-Prozesses.

Schmidts, M. & Lischka M. (2001): Prüfungsfragen für Multiple-Choice Tests erstellen. Kurzanleitung mit Beispielen.

Schulze, J. et al. (2005): Einfluss des Fragenformates in Multiple-Choice-Prüfungen auf die Antwortwahrscheinlichkeit. Eine Untersuchung am Beispiel mikrobiologischer Fragen. In: GMS Zeitschrift für Medizinische Ausbildung: 22(4) 2005.

Universität Hannover/elsa (o.J.): Erstellen und Bewerten von Multiple-Choice-Aufgaben.

Williams, R.G. & Haladyna T.M. (1982): Logical Operations for Generating Intended Questions (LOGIQ): A Typology for Higher Level Test Items. In: Roid, G. H., & Haladyna, T. M. (1982). A technology for test-item writing. New York: 161-186.

Zimmerman, B. B., Sudweeks, R. R., Shelley, M.F. & B. Wood (1990). How to Prepare Better Tests: Guidelines for University Faculty. Provo, UT: Brigham Young University Testing Services.