UNIVERSITÄT BERN

# Multiple Choice Prüfung

[Robert Hilbe & Silke Wehr Rappo]

Bei dieser schriftlichen Prüfungsform werden kurze Fragen gestellt und verschiedene Antworten vorgegeben. Die richtige Antwort muss dabei von den Studierenden angekreuzt werden. Ein Vorteil dieser verbreiteten Prüfungsform ist, dass die Prüfungen elektronisch ausgewertet werden können. Die Antwortbogen werden dabei eingescannt und automatisch ausgewertet. Es ist allerdings relativ aufwendig und anspruchsvoll, gute MC-Fragen zu entwickeln. Es sind formale und sprachliche Hinweise (cues) auf die richtige Antwort zu vermeiden (Krebs 2004), die es erfahrenen Studierenden erlauben, die richtigen Antworten zu erkennen. Ein Nachteil dieser Methode ist auch, dass nur bestimmte Fragetypen gestellt werden können. Bei Prüfungswiederholungen kann es schwierig sein, MC-Fragen mit dem gleichen Schwierigkeitsgrad zu entwickeln.

## Vorteile von MC-Prüfungen

Multiple-Choice-Aufgaben (MC) kommen im universitären Kontext häufig zum Einsatz, da sie gerade bei grossen Studierendenzahlen eine ökonomische Messung von Leistungen der Studierenden ermöglichen. Sowohl die Prüfungsdurchführung als auch die -auswertung ist weniger aufwendig als bei anderen Prüfungsformen, da dieser Aufgabentyp auch eine maschinelle Auswertung erlaubt. Neben diesen ökonomischen Vorteilen sprechen auch testtheoretische Gründe für MC-Aufgaben: Sie weisen eine hohe Durchführungs- sowie Auswertungsobjektivität sowie eine gute Reliabilität auf, wenn die Testaufgaben gut konstruiert sind.

## Nachteile von MC-Prüfungen

Trotz dieser Vorteile haben MC-Prüfungen einen schlechten Ruf, der sich durch die Nachteile dieses Aufgabenformats erklären lässt:

- Es ist anspruchsvoll und aufwendig, geeignete MC-Aufgaben zu entwerfen, die nicht nur Wissen abfragen, sondern Problemlösefähigkeiten und die Anwendung von Wissen fokussieren (höhere Taxonomiestufen nach der Bloomschen Taxonomie der kognitiven Lernziele: Anderson & Krathwohl 2001);
- die richtigen Antworten können erraten werden;
- geeignete Falschantworten zu finden, ist anspruchsvoll;
- komplizierte Formulierungen und Antwortoptionen in schlechten MC-Prüfungen testen eher das Textverständnis und das logische Denken als die zu messenden Kompetenzen bzw. Kenntnisse der Studierenden.

Das Erstellen guter MC-Aufgaben setzt Erfahrung und Wissen in der Formulierung von MC-Fragen

voraus:

Die Crux von MC-Aufgaben liegt jedoch nicht im Format per se, sondern vielmehr in der Qualität der Aufgaben. Das Erstellen guter MC-Aufgaben ist anspruchsvoll und zeitintensiv und setzt Erfahrung in der Generierung von Prüfungsfragen voraus. Das schlechte Image von MC-Prüfungen rührt daher von der mangelnden Qualität der Fragen.

Besonders augenfällig wird eine mangelnde Qualität von MC-Fragen, wenn die Falschantworten (so genannte Distraktoren) genauer betrachtet werden. Diese sollten «für den Nichtkönner plausibel und attraktiv sein, dem Könner hingegen in ihrer Falschheit evident sein» (Jacobs 2000). Vermieden werden sollten daher inhaltlich unplausible und abwegige Distraktoren. Neben inhaltlich unplausiblen Antwortoptionen können es aber auch formale Hinweise sein, welche den Studierenden die richtigen bzw. die falschen Antwortoptionen verraten. Solche Cues können beispielsweise absolute Begriffe wie «immer», «nie», «alle», «kein», «nur» usw. in Distraktoren sein.

In MC-Aufgaben mit schlechter Qualität führen solche Cues zu einer mangelnden Reliabilität der Leistungsmessung. Gemessen werden dann nicht die zu messenden Kompetenzen, sondern die Testkompetenz bzw. «Testschlauheit» der Studierenden.

# Testtheoretische Grundlagen zur Planung, Durchführung und Auswertung von MC-Prüfungen

Prüfungen können die von ihnen zu erfassenden Leistungsmerkmale nicht mit absoluter Exaktheit messen, da sich ein Testergebnis gemäss dem ersten Axiom der klassischen Testtheorie stets aus dem «wahren Wert» und einem Messfehler zusammensetzt (vgl. Eugster & Lutz 2004, S. 16). Durch sorgfältige Planung, Durchführung und Auswertung von Prüfungen soll dieser Messfehler möglichst klein gehalten werden, wobei die klassischen Gütekriterien Objektivität, Reliabilität und Validität zu berücksichtigen sind (vgl. dazu Eugster & Lutz 2004; Metzger & Nüesch 2004; Woschnack & Schatz 2011). Zusätzlich sollten Prüfungen als Nebenkriterien normiert, vergleichbar und ökonomisch sein (vgl. Lienert & Raatz  1998, S. 7ff.).

## Objektivität

Ein Test ist dann objektiv, wenn seine Ergebnisse unabhängig von der durchführenden Person sind. Eine Prüfung weist eine hohe Reliabilität auf, wenn unterschiedliche Prüfer unter gleichen Bedingungen zu denselben Beurteilungen gelangen. Da der Einfluss des Prüfenden in verschiedenen Phasen wirksam werden kann, werden folgende drei Aspekte der Objektivität unterschieden (vgl. Eugster & Lutz 2004, S. 16ff.).

Durchführungsobjektivität

Eine hohe Durchführungsobjektivität ist dann gegeben, wenn das Verhalten der prüfenden Person

während des Tests das Ergebnis nicht beeinflusst. Bei schriftlichen Prüfungen sind dafür genaue Instruktionen für die Aufsichtspersonen notwendig, die ein möglichst vergleichbares Verhalten der Prüfenden während der Prüfung gewährleisten.

Die Durchführungsobjektivität kann beispielsweise dadurch gewährleistet werden, dass mehrere Personen die Prüfung gemeinsam in einem grossen Hörsaal durchführen. Der Ablauf der Prüfung sollte vorgängig verbindlich festgelegt werden (Sitzanordnung, Instruktion vor der Prüfung, Verhalten der Prüfenden bei Täuschungsversuchen der Studierenden). Auch können zwei Prüfungsversionen mit gleicher Schwierigkeit verwendet werden und jeweils ein Platz zwischen zwei Studierenden frei gelassen werden, um die Möglichkeit des Abschauens zu reduzieren. Die Verwendung unerlaubter Hilfsmittel sollte erschwert werden, indem nur Schreibzeug am Platz erlaubt ist. Die geprüften Studierenden können auf dem Deckblatt der Prüfung über die Art und Anzahl der Fragen (z.B. 24 Multiple Choice + 3 offene Fragen), die zur Verfügung stehende Zeit, unerlaubte Hilfsmittel, Konsequenzen von Täuschungsversuchen und über das Vorgehen der Bewertung informiert werden.

## Auswertungsobjektivität

Bei der Korrektur der Prüfung muss die Zuteilung von Punktewerten zu den gegebenen Antworten unabhängig von der prüfenden Person erfolgen. Für schriftliche Prüfungen empfiehlt es sich, dazu Musterlösungen und Korrekturanleitungen zu entwickeln, damit eine einheitliche Punktevergabe gesichert ist. Empfehlenswert ist (falls möglich) auch eine Zweitkorrektur durch eine andere Person.

Geschlossene MC-Fragen können maschinell beispielsweise mit der Software EvaExam ausgewertet werden. Mehrdeutige oder unklare Markierungen der Ankreuzfelder können der kontrollierenden Person zur Verifizierung angezeigt. Die Auswertungsobjektivität dieser geschlossenen Fragen wird dadurch gesichert.

## Interpretationsobjektivität

Ziel der Interpretationsobjektivität ist die Unabhängigkeit der Notengebung aufgrund bestimmter Punktewerte von den prüfenden Personen. Als Interpretationsnorm gilt meist eine von der Prüfungsordnung vorgegebene Notenskala. Dabei ist vor allem innerhalb eines Studiengangs auf eine möglichst einheitliche Zuteilung von Punktewerten zu Noten zu achten.

Die Notenskala kann unter Berücksichtigung bereits vergangener Prüfungen des gleichen Lernmoduls in vergangenen Semestern, der tatsächlichen Punkteverteilung der Antworten und der im Rahmen von Bologna empfohlenen prozentualen Verteilung der Noten erfolgen.

# Reliabilität

Reliabilität (dt.: Zuverlässigkeit) bezeichnet die Genauigkeit der Messung. Ein Test ist dann reliabel,

wenn er möglichst ohne Messfehler misst. Eine hohe Reliabilität führt dazu, dass bei einer Wiederholung der Messung unter gleichen Rahmenbedingungen das gleiche Messergebnis erzielt wird (Replizierbakeit von Ergebnissen). Reliabilität lässt sich daher auch als Anteil der wahren Varianz an der beobachteten Varianz, die zusätzlich durch Zufallsfehler verursacht wird, bestimmen. Um die Reliabilität eines Tests empirisch abzuschätzen, werden vier Techniken eingesetzt:

## Testhalbierungsmethode

Bei diesem auch als Split-half-Reliabilität bezeichneten Vorgehen wird der Test in zwei Hälften unterteilt, wobei die Mittelwerte und die statistischen Verteilungskennwerte der beiden Hälften bei genügend grosser Studierendenzahl möglichst gleich sein sollten.

## Testwiederholung

Der gleiche Test wird bei den gleichen Studierenden zu verschiedenen Zeitpunkten eingesetzt, um die Konstanz der Ergebnisse zu prüfen (Retest-Reliabilität). Für viele Tests stellt diese Form der Reliabilitätsüberprüfung keine praktikable Möglichkeit dar, da durch Erinnerungs-, Lern- und Übungseffekte eine Beeinflussung der wiederholten Testdurchführung stattfindet.

Eine abgeschwächte Form der Retest-Reliabilität kann sein, dass eine bereits bestehende Prüfung eines vergangenen Semesters verwendet und überarbeitet wird. Ein Vergleich der Lösungswahrscheinlichkeiten und der Punkteverteilung der Prüfungswiederholung dient der Überprüfung der Retest-Reliabilität, mit der Einschränkung, dass es sich nicht auf die gleiche Stichprobe an Studierenden bezieht.

## Paralleltest-Reliabilität

Bei dieser Form der Reliabilität werden den Kandidaten zwei verschiedene Prüfungen vorgelegt, welche beide dieselben Lernziele bzw. Kompetenzen messen. Führen beide Tests zu vergleichbaren Ergebnissen, ist die Paralleltest-Reliabilität hoch.

Bei MC-Prüfungen können, wie oben erwähnt, zwei Versionen des Tests eingesetzt werden, um die Gefahr des Abschauens zu minimieren. Durch die Kontrolle der Mittelwerte und Verteilungskennwerte nach der Korrektur der Versionen A und B kann die Paralleltest-Reliabilität überprüft werden.

## Innere Konsistenz

Bei der inneren Konsistenz wird geprüft, wie stark die richtige Beantwortung einer Prüfungsfrage mit der richtigen Beantwortung anderer Prüfungsfragen einher geht. Sie stellt damit eine interne Form der Qualitätsmessung dar, sofern die Möglichkeit eines Retests oder Paralleltests nicht zur Verfügung

steht. Jedes Prüfungsitem wird dabei mit jedem anderen Prüfungsitem korreliert (Interkorrelationsmatrix).

Eine Form der Überprüfung der inneren Konsistenz kann darin bestehen, dass die Lösungswahrscheinlichkeiten (bzw. der Schwierigkeitsindex) der einzelnen Prüfungsfragen bei der Korrektur analysiert werden. Fragen, die von sehr vielen Studierenden beantwortet bzw. nicht richtig beantwortet werden konnten, schmälern die innere Konsistenz und sollten daher nur sparsam eingesetzt werden (vgl. Lienert & Raatz 1998, S. 213).

# Validität

Ein Test ist dann valide, wenn er das misst, was er zu messen vorgibt. Eine hohe Objektivität bzw. Reliabilität stellen dabei notwendige, aber noch nicht hinreichende Bedingungen für die Validität dar. Ein Test, der zwar objektiv und reliabel das Wissen über ein Stoffgebiet misst, kann nicht valide sein, wenn dieses Stoffgebiet nicht Inhalt der Veranstaltung war.

Es gibt drei Techniken, um die Validität eines Tests zu bestimmen: Inhaltsvalidität, Kriteriumsvalidität und Konstruktvalidität (vgl. Eugster & Lutz 2004, S. 18f.).

## Inhaltsvalidität

Eine Prüfung ist dann valide, wenn sie die Fülle der zu prüfenden Lernziele repräsentativ abdeckt. Dabei ist einerseits zu berücksichtigen, dass die thematische Abdeckung des Prüfungsstoffs gewährleistet ist, andererseits muss beachtet werden, welche kognitiven Prozesse zur Erreichung der Lernziele nötig sind. Hierbei hat sich die Einteilung anhand der kognitiven Lernzieltaxonomie nach Bloom (Anderson & Krathwohl 2001, vereinfacht dargestellt in Metzger & Nüesch 2004, S. 14ff.) durchgesetzt: Wissen, Verstehen, Anwenden, Analysieren, Synthese erstellen und Beurteilen.

Eugster und Lutz (2004, S. 19) empfehlen, dass Prüfungsaufgaben …

• eine präzise Aufgabenstellung beinhalten, die verschiedene Teilaufgaben deutlich voneinander abgrenzt,
• «klare Angaben über die Form der Ergebnisdarstellung» beinhalten, d.h. für den Prüfling ersichtlich ist, in welcher Form die Antwort ausfallen soll (z.B. «in Stichworten» oder «in Form einer Tabelle mit folgenden Elementen …»),
• Angaben über den Beurteilungsmassstab (erreichbare Punkte) enthalten sollen.

Dazu empfiehlt es sich, bereits vor der Korrektur der Prüfung einen Lösungsschlüssel mit den richtigen Antworten in der verlangten Darstellungsform zu erstellen.

## Kriteriumsvalidität

Bei der Kriteriumsvalidität geht es darum, das Ergebnis eines Tests anhand eines Aussenkriteriums zu prüfen. Man unterscheidet zusätzlich danach, wann das gewählte Aussenkriterium erhoben wird. Bei einer gleichzeitigen Messung von Test und Aussenkriterium spricht man von Übereinstimmungsvalidität (concurrent validity). Diese liegt etwa dann vor, wenn die Ergebnisse einer Prüfung mit anderen, vergleichbaren Leistungskontrollen übereinstimmen. Wird das Aussenkriterium zu einem späteren Zeitpunkt gemessen, so ist von Prognosevalidität (prognostic validity) die Rede. Beispielweise kann das Prüfungsergebnis zur Vorhersage künftiger Leistungen im selben oder in benachbarten Fächern verwendet werden.

## Konstruktvalidität

Tests oder Prüfungen dienen zur Messung von Kompetenzen, die als Lernziele definiert sind. Diese Kompetenzen sind jedoch oft nicht direkt beobachtbar und somit auch nicht direkt überprüfbar. Um Kompetenzen prüfbar zu machen, müssen sie daher operationalisiert, d.h. in beobachtbares Verhalten im Sinne von Wissen, Anwenden, Analysieren usw. übersetzt werden. «Die Konstruktvalidität eines Tests bezieht sich auf diesen Gesichtspunkt, indem sie Auskunft darüber gibt, ob mit Hilfe der Prüfungsaufgaben gültige Rückschlüsse auf den Ausprägungsgrad der im Lernziel beschriebenen Kompetenzen möglich sind. Mit anderen Worten: Eine Prüfung ist konstruktvalide, wenn sie Hypothesen bestätigt, die sich aus den Lernzielen ableiten lassen» (Eugster & Lutz 2004, S. 19). Prüfungen sind daher so zu planen, dass sie möglichst ein facettenreiches und differenziertes Bild des von den für die angestrebten Lernziele erforderlichen Fähigkeiten liefern. Die als Lernziel formulierten Kompetenzen müssen dazu in konkret messbare Fähigkeiten übersetzbar sein.

# Nebengütekriterien

Zusätzlich zu den klassischen Gütekriterien werden von Eugster und Lutz (2004) die Nebengütekriterien Normierung, Vergleichbarkeit und Ökonomie genannt. An anderer Stelle finden sich weitere, zusätzliche Nebenkriterien: Skalierung, Testfairness, Nützlichkeit, Zumutbarkeit, Vergleichbarkeit, Unverfälschbarkeit, Transparenz, Akzeptanz, äussere Gestaltung. Diese sollen hier nicht explizit erklärt werden, da sie teils selbsterklärend sind oder sich stark mit den hier aufgeführten Kriterien überschneiden.

## Normierung

Mittels Normierung soll ein Bezugssystem geschaffen werden, das es erlaubt, ein individuelles Ergebnis vergleichend zu anderen Ergebnissen einer Population einzuordnen. Im Falle von Noten dient die Notenskala von 1 bis 6 als solches Bezugssystem. Die Notenskala ist jedoch streng genommen nur eine Ordinalskala, d.h. sie bringt lediglich eine Rangfolge zum Ausdruck. Mathematische Operationen, wie sie bei Intervall- oder gar bei Verhältnisskalen möglich sind, dürfen damit streng genommen mit Notenwerten nicht vorgenommen werden (vgl. Eugster & Lutz 2004, S. 20). Eine Studentin mit einer 6 (beste Note in der Schweiz) ist damit nicht doppelt so gut ist wie eine Studentin mit einer 3 (ungenügend in der Schweiz).

## Vergleichbarkeit

Zur Vergleichbarkeit trägt bei, wenn von einem Test mehrere Parallelformen vorliegen. Für Prüfungen empfiehlt es sich daher, verschiedene Varianten auszuarbeiten, die bei auftretenden Güteproblemen eine überprüfbare Verbesserung der Tests ermöglichten.

## Ökonomie

Ein Test sollte möglichst ressourcenschonend (z.B. Zeit und Geld) durchführbar sein. Eine Prüfung sollte daher für die Studierenden in einer verhältnismässigen Dauer durchführbar sein, nicht zu viel Material (z.B. Papier für Kopien) beanspruchen, für die Kandidaten einfach zu handhaben und für die Durchführenden einfach auszuwerten sein.

Die mechanische Auswertung der geschlossenen Fragen erlaubt eine zeiteffiziente Korrektur – ein Argument, das bei über 100 Studierenden ebenfalls wichtig ist.

# Formale und sprachliche Lösungshinweise (Cues) erhöhen richtige Antworten durch Raten

Cues erlauben es, erfahrenen Studierenden unabhängig von ihrem Wissen die richtige Antwort zu identifizieren bzw. einzelne falsche Antworten auszuschliessen und damit die Ratewahrscheinlichkeit für die richtige Antwort zu erhöhen. Damit führen sie dazu, dass die Reliabilität und letztlich auch die Validität einer Prüfung reduziert wird, da nicht das inhaltliche Wissen gemessen wird, sondern «Testkompetenz». Cues stellen damit – testtheoretisch formuliert – Störeinflüsse dar, die den Anteil der Fehlervarianz im Vergleich zur tatsächlichen Varianz erhöhen.

# Formale und sprachliche Cues, die bei der Erstellung von MC-Fragen vermieden werden sollten – Varianten

Krebs (2004) führt folgende Arten von Cues auf, die jeweils mit einem Beispiel illustriert werden:

## Grammatikalisch unpassende Anschlüsse

Alle Antworten müssen grammatikalisch zum Stamm passen. In folgendem Beispiel passt nur Antwort A. Oft wird jedoch bei komplexeren Anschlüssen und der Formulierung von Distraktoren die Prüfung der grammatikalischen Korrektheit vernachlässigt.

Anter ist eine …

(A) Legierung
(B) Konglomerat
(C) Verbrennungsrückstand
(D) Spaltprodukt
(E) chemisches  Element

## Formal auffällige richtige bzw. falsche Antworten

Distraktoren sollen möglichst gleich lang und ausdifferenziert sein wie die richtige Antwort. In diesem Beispiel ist Antwort B mit Abstand die längste und differenzierteste Antwort, was darauf hindeutet, dass diese Antwort richtig ist.

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 OgI4 freigesetzt
(D) entsteht Fermantin
(E) bildet sich Gorantoxol

## Mehrere ähnliche bzw. sich ausschliessende Antworten

Auch Hinweise, welche die Aufmerksamkeit auf zwei bis drei 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» (Krebs 2004, S. 20).

In folgendem Beispiel wird die Auswahl der richtigen Antwort auf diese Weise auf Option A oder B eingeschränkt.

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 ?-Frequenz während der Lubrierung konstant bleibt
(E) das Gorantoxin vorgängig sterniert wird

## Konvergenz-Cues

«Die Antwort, welche die grösste Zahl von Elementen mit andern Antworten gemeinsam hat, ist mit erhöhter Wahrscheinlichkeit die richtige (sog. Konvergenzstrategie)» (ebd., S. 20).

In diesem Beispiel deutet die mehrfache Verwendung von «Uniform», «Source» und «Locator» darauf hin, dass Option C richtig sein muss.

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

## Verbale Assoziationen zwischen Stamm und richtiger Antwort

Verbale Assoziationen zwischen Stamm und richtiger Antwort sind zu vermeiden. Im Beispiel ist sowohl im Stamm auch als in Antwort D das Wort «kognitiv» enthalten.

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

## Verwendung absoluter Begriffe als Indikator für falsche Antworten

Absolute Begriffe wie «nie», «immer» usw., die in Distraktoren verwendet werden, um Aussagen eindeutig falsch zu machen, sind zu vermeiden. Im Beispiel ist nur Option E keine absolute Formulierung.

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.

## Gegenseitige Lösungshinweise

Wird die Prüfung am Schluss nicht mehr als Ganzes geprüft, kann es vorkommen, dass Antworten auf eine Frage Hinweise auf die richtige Lösung einer anderen Frage geben. Dies ist daher v.a. bei der Zusammenstellung der Prüfungsfragen zu beachten. Die Gefahr eines Lösungshinweises kann dadurch minimiert werden, dass möglichst keine überflüssigen Informationen in den Items gegeben werden.

Neben diesen Cues kann auch die Anordnung der Antworten Hinweise auf die richtige bzw. falsche Lösung geben. Daher sind zusätzlich folgende formale Fehler zu vermeiden:

## Unlogische Anordnung der Antworten

Die Antworten sollen möglichst logisch angeordnet werden. Jene Kandidaten, die  die Antwort nicht kennen, werden nämlich auch bei der Reihung nach einer Auffälligkeit suchen. Krebs (2004) schlägt daher vor, logische Hierarchien einzuhalten, falls diese gegeben sind (Anordnung von klein zu gross, Zahlen auf- bzw. absteigend). Bei Einwortantworten empfiehlt sich eine alphabetische Reihung.

## Platzierung der richtigen Antwort überwiegend unter C oder D

Positions-Wahltendenzen von Studierenden können die Reliabilität der Prüfung minimieren, wenn die richtige Antwort häufig an bestimmten Positionen plaziert wird. In der Absicht, die richtige Antwort möglichst gut zu verstecken, haben viele Prüfer die Tendenz, diese häufig unter C oder D zu platzieren. Die richtige Antwort sollte jedoch möglichst ausbalanciert unter allen möglichen Antwortoptionen verteilt werden.

# Weiterlesen

Anderson, L.W. & Krathwohl D.R. (eds.) (2001). A Taxonomy for Learning, Teaching and Assessing. New York: Longman.

Eugster, B.; Lutz, L. (2004). Leitfaden für das Planen, Durchführen und Auswerten von Prüfungen an der ETHZ. Zürich: ETH, Didaktikzentrum.

Krebs, R. (2004). Anleitung zur Herstellung von MC-Fragen und MC-Prüfungen für die ärztliche Ausbildung. Bern: Institut für Medizinische Lehre IML, Abteilung für Assessment und Evaluation AAE. Verfügbar unter: *Anleitung zur Herstellung von MC-Fragen und MC-Prüfungen für die ärztliche*

*Ausbildung* [12.03.2020].

Lienert, G. A.; Raatz, U. (1998). Testaufbau und Testanalyse (6. Aufl). Weinheim: Beltz Psychologie VerlagsUnion.

Metzger, C.; Nüesch, C. (2004). Fair prüfen: ein Qualitätsleitfaden für Prüfende an Hochschulen. St. Gallen: Universität, Institut für Wirtschaftspädagogik.