

JOHANNES GUTENBERG
UNIVERSITÄT MAINZ



Zentrum für
Qualitätssicherung
und -entwicklung

Handreichung
der Prüfungswerkstatt

Qualitätskriterien für Prüfungen

André Beem, M.A.

LOB LEHREN – ORGANISIEREN – BERATEN



Bundesministerium
für Bildung
und Forschung



INHALT DIESER HANDREICHUNG

Definition und Bedeutung von Qualitätskriterien im Prüfungsprozess	3
Eigenschaften und Erfolgsbedingungen von Qualitätskriterien	3
Prüfungsqualität im Kontext der Kompetenzorientierung	4
Literaturverzeichnis	6

DEFINITION UND BEDEUTUNG VON QUALITÄTSKRITERIEN IM PRÜFUNGSPROZESS

Wissenschaftlich fundierte Testverfahren (z. B. Abschlussprüfungen) zeichnen sich durch die Erfüllung von Güte-/ Qualitätskriterien aus, die empirisch überprüfbar sind. Diese Kriterien stellen ein Instrument der Qualitätsbeurteilung psychologischer Tests dar und gehen i.d.R. auf die klassische Testtheorie (KTT) zurück (Moosbrugger & Kelava 2008: S. 8; Rost 2004: S. 17). Allgemein gilt, dass ein Test bzw. eine Prüfung als Messung einer gezeigten Performanz verstanden wird, die Rückschlüsse auf die zugrundeliegende Kompetenz ermöglichen kann. Ziel ist die möglichst genaue und fehlerfreie Abbildung der Fähigkeiten von Studierenden in einem bestimmten Stoffgebiet zu einem bestimmten Zeitpunkt. Grundsätzlich gelten für Prüfungen an Hochschulen die gleichen Anforderungen wie für andere (psychologische) Messinstrumente (z. B. Intelligenz- oder Persönlichkeits-tests), die nachfolgend im Einzelnen erläutert werden (Brauns & Schubert 2008: S. 93).

EIGENSCHAFTEN UND ERFOLGSBEDINGUNGEN VON QUALITÄTSKRITERIEN

- **Objektivität** wird in die Bereiche Durchführungs-, Auswertungs- und Interpretationsobjektivität gegliedert. Durchführungsobjektivität ist gegeben, wenn die Bedingungen des Tests (z. B. Bearbeitungszeit) für alle Testteilnehmer/innen gleich sind. Auswertungsobjektivität betrifft die einheitliche Bewertung der Testergebnisse. Es muss also klar formulierte Vorgaben (z. B. in Form eines Kriterienrasters) für die Punktvergabe bei einzelnen Fragen geben. Bei offenen Fragen können bspw. vorab Schlagwörter definiert werden, die die Testteilnehmer/innen in der Antwort nennen müssen. Interpretationsobjektivität zielt auf eine möglichst konsistente Beurteilung der Testergebnisse. Demnach sollte möglichst jede/r Auswerter/in z. B. durch Verwen-

dung eines Notenschemas zur gleichen Beurteilung der Testergebnisse kommen (Brauns & Schubert 2008: S. 96; Bühner 2011: S. 58ff.). Durch eine Standardisierung der Prüfungsdurchführung (z. B. Bearbeitungszeit) und Bewertungsverfahren (z. B. Kriterienraster) lässt sich die Objektivität einer Prüfung steigern.

- **Reliabilität** beschreibt den Grad der Genauigkeit, mit der ein Test ein bestimmtes Merkmal misst bzw. wie hoch die Reproduzierbarkeit der Ergebnisse ist. Ein Test sollte demnach bei jedem Durchlauf genau dasselbe Merkmal prüfen. Durch eine Erhöhung der Anzahl von Fragen kann die Ratewahrscheinlichkeit verringert und somit eine Steigerung der Reliabilität erreicht werden. Allerdings muss dabei beachtet werden, dass eine zu lange Prüfung zu Ermüdungseffekten und damit zu einer Erhöhung der inhaltsirrelevanten Varianz führen kann (Brauns & Schubert 2008: S. 96; Rodriguez 2005: S. 10; Lee et al. 2011: S. 132).
- **Validität** gibt an, ob der Test auch tatsächlich das misst, was er messen soll. Dementsprechend sollten Testfragen die in der zugehörigen Lehrveranstaltung vermittelten Inhalte repräsentativ widerspiegeln. Ein Faktor, der die Validität negativ beeinflussen kann, ist bspw. die Verständlichkeit/Lesbarkeit des Aufgabentextes. Sind die Fragen nur schwer verständlich, wird demnach die Lesekompetenz und nicht das eigentlich intendierte Lernziel geprüft. Insbesondere bei geschlossenen Fragetypen kann auch die Ratewahrscheinlichkeit die Validität beeinflussen. Um eine möglichst hohe Validität zu erreichen, sollte jede Prüfungsfrage auf Grundlage eines verbindlichen und transparenten Lernziels formuliert sein. Die Relevanz und Repräsentativität der Frage für das betreffende Lernziel kann zusätzlich noch von Experten/innen (z. B. Fachkollegen/innen) geprüft werden (Brauns

& Schubert 2008: S. 96ff.; Downing & Haladyna 2004: S. 329).

- **Skalierbarkeit** beschreibt die Bildung eines Testwerts durch eine gültige Verrechnungsregel. Demnach müssen die Testergebnisse die empirische Merkmalsrelation adäquat abbilden. Die Umsetzbarkeit dieses Gütekriteriums hängt insbesondere vom Skalenniveau des Mesinstrumentes ab. In der Regel reicht eine Messung des Merkmals auf Nominalskalenniveau nicht aus, um die größer/kleiner Relation zwischen den Testpersonen zu beschreiben. Damit eine leistungsfähigere Testperson einen besseren Testwert als eine leistungsschwächere erhält, muss zumindest eine Messung auf Ordinalskalenniveau (die Ausprägungen lassen sich in eine Rangfolge bringen; z. B. Schulnoten) erfolgen. Eine Messung auf Intervallskalenniveau (die Abstände der Ausprägungen lassen sich exakt bestimmen; z. B. Temperatur) erlaubt darüber hinaus eine Beurteilung der Größe inter- und intraindividuelle Differenzen (Bühner 2011: S. 67 f.; Moosbrugger 2008: S. 18-19).
- **Zumutbarkeit** beschreibt die Durchführung des Tests aus Sicht der Prüflinge und ist dann gegeben, wenn diese im Verhältnis zum diagnostischen Erkenntnisgewinn der Prüfung in zeitlicher, psychischer und körperlicher Hinsicht nicht über Gebühr belastet werden. So steigt bspw. die Anforderung an die Konzentrationsfähigkeit des Prüflings, je länger ein Test dauert und je schwerer die Aufgaben zu lösen sind. Eine Vielzahl unlösbarer Aufgaben kann zudem zu Frustration führen und das Testergebnis zusätzlich verzerren (Eid & Schmidt, 2014: S. 68; Moosbrugger 2008: S. 22).
- **Testfairness** ist dann gegeben, wenn Personen z. B. aufgrund ihrer Zugehörigkeit zu unterschiedlichen Gruppen oder aufgrund ihres soziokulturellen Hinter-

grunds oder ihres Geschlechts keine systematischen Vor- oder Nachteile bei der Beantwortung einer Aufgabe haben (Eid & Schmidt 2014: S. 68 f.). Verständlichkeit und Transparenz des Arbeitsauftrags sowie einheitliche Bedingungen bei der Prüfungsvorbereitung (z. B. Bereitstellung von Lernmaterialien) steigern die Testfairness.

- **Testökonomie** spielt bei der Prüfungsgestaltung ebenfalls eine Rolle. Der diagnostische Erkenntnisgewinn einer Prüfung sollte in einem sinnvollem Maß zum zeitlichen, personellen und finanziellen Aufwand hinsichtlich Vorbereitung, Durchführung und Auswertung des Tests stehen (Eid & Schmidt 2014: S. 68 f.; Moosbrugger 2008, S. 21).

Zur Qualitätssicherung sollten Prüfungsfragen nach jedem Einsatz einer Itemanalyse unterzogen werden, um u. a. deren Schwierigkeit und Trennschärfe zu ermitteln. Auf diese Weise können „auffällige“ Fragen (z. B. keine/wenige korrekte Antworten) identifiziert werden und ggf. herausgenommen oder nachbewertet werden (Bauns & Schubert 2008, S. 101).

PRÜFUNGSQUALITÄT IM KONTEXT DER KOMPETENZORIENTIERUNG

Das Leitprinzip der Kompetenzorientierung in der Hochschullehre setzt für Prüfungen folgende Aspekte zur Qualitätssicherung voraus (Hauer, 2011; Schaper & Hilkenmeier, 2013; Zlatkin-Troitschanskaya et. al., 2013):

- Prüfungen sollten einen hohen Handlungsbezug haben und neben deklarativen auch andere Wissensdimensionen (z.B. Prozeduren) erfassen.
- Konstruktives und kriterienbasiertes Feedback sollte in den gesamten Lernprozess und damit auch die Prüfung selbst eingebettet werden.

Gute Prüfungsfragen sollten daher (Auspurg et al., 2009; Downing & Haladyna, 2004; Hauser 2011; Zlatkin-Troitschanskaia et al., 2013)...

- ...eine praxisnahe Problemstellung behandeln. Dies kann bspw. durch die Formulierung hypothetischer Objekt- oder Fallbeschreibungen, sog. Vignetten, erreicht werden.
- ...auf Grundlage eines spezifischen Lernziels formuliert werden.
- ...einfach sowie unmissverständlich und grammatikalisch korrekt formuliert sein, damit die Lesekompetenz des einzelnen Prüflings nicht zu einem Störfaktor wird.
- ...in Schwierigkeit und Struktur variieren.
- ...transparent hinsichtlich Arbeitsauftrag (z. B. erlaubte Hilfsmittel) und Bewertung (z. B. maximal erreichbare Punkte) sein.

LITERATURVERZEICHNIS

- Brauns, Katrin; Schubert, Sebastian (2008): Qualitätssicherung von Multiple-Choice-Prüfungen. In: Sigrid Dany (Hg.): Prüfungen auf die Agenda! Hochschuldidaktische Perspektiven auf Reformen im Prüfungswesen. Bielefeld: Bertelsmann (Blickpunkt Hochschuldidaktik, 118), S. 92–102.
- Bühner, Markus (2011): Einführung in die Test- und Fragebogenkonstruktion. 3., aktualisierte und erw. Aufl. München: Pearson Studium (Psychologie).
- Downing, Steven M.; Haladyna, Thomas M. (2004): Validity threats. Overcoming interference with proposed interpretations of assessment data. In: *Med Educ* 38 (3), S. 327–333. DOI: 10.1046/j.1365-2923.2004.01777.x.
- Eid, Michael; Schmidt, Katharina (2014): Testtheorie und Testkonstruktion. Göttingen: Hogrefe (Bachelorstudium Psychologie Psychlehrbuch plus).
- Eugster, Balthasar; Lutz, Leonhard (2003): Leitfaden für das Planen, Durchführen und Auswerten von Prüfungen an der ETHZ. Eidgenössische Technische Hochschule Zürich. Zürich. Online verfügbar unter http://www.let.ethz.ch/pruefungen/Leitfaden_PDA_Pruefungen_DiZ-2003.pdf, zuletzt geprüft am 11.08.2014.
- Hauer, Erich (2011): Wird dumm geprüft, wird dumm gelernt. Plädoyer für den Einsatz anwendungsorientierter Prüfungsaufgaben im Hochschulbereich. In: Elke Gruber (Hg.): Qualität ist kein Zufall. MAGAZIN erwachsenenbildung.at, 12/2011. Norderstedt: Books on Demand, S. 10.1-10.10. Online verfügbar unter http://www.pedocs.de/volltexte/2013/7418/pdf/Erwachsenenbildung_12_2011_Hauer_Wird_dumm_geprueft.pdf, zuletzt geprüft am 11.08.2014.
- Lee, Hee-Sun; Liu, Ou Lydia; Linn, Marcia C. (2011): Validating Measurement of Knowledge Integration in Science Using Multiple-Choice and Explanation Items. In: *Applied Measurement in Education* 24 (2), S. 115–136. DOI: 10.1080/08957347.2011.554604.
- Lienert, Gustav A.; Ratz, Ulrich (1998): Testaufbau und Testanalyse. 6. Aufl., Studienausg. Weinheim: Beltz.
- Moosbrugger, H., Kelava, A. (2008). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In *Testtheorie und Fragebogenkonstruktion* (pp. 7-26). Springer Berlin Heidelberg.
- Moosbrugger, H. (2012). Item-Response-Theorie (IRT). In *Testtheorie und Fragebogenkonstruktion* (pp. 227-274). Springer Berlin Heidelberg.
- Rost, Jürgen (2004): Lehrbuch Testtheorie – Testkonstruktion. 2. Aufl. Bern: Huber.
- Rodriguez, Michael C. (2005): Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research. In: *Educational Measurement issues and practice*, 24, S. 3–13.
- Schaper, Niclas; Hilkenmeier, Frederic (2013): Umsetzungshilfen für kompetenzorientiertes Prüfen.
- Wannemacher, Klaus (2007): Computerbasierte Prüfungen. Zwischen Self-Assessment und Abschlussklausuren. In: Michael H. Breitner, Beate Bruns und Franz Lehner: *Neue Trends im E-Learning [Elektronische Ressource]. Aspekte der Betriebswirtschaftslehre und Informatik*. Hg. v. Michael H. Breitner. Heidelberg: Physica-Verlag HD, S. 427–440.
- Zlatkin-Troitschanskaia, Olga; Förster, Manuel; Brückner, Sebastian; Hansen, Manuela; Happ, Roland: Modellierung und Erfassung der wirtschaftswissenschaftlichen Fachkompetenz bei Studierenden im deutschen Hochschulbereich. In: *Lehrerbildung auf dem Prüfstand 2013, Sonderheft*, S. 108–133.