Universität Zürich[UZH]

Department of Economics

Universität Zürich
ECON Teaching Center
Zürichbergstrasse 14
CH-8032 Zürich

Qualitätsanalyse von Prüfungen

Für die Prüfungsanalyse wird unterschieden zwischen Kennzahlen auf Auswahl-, Item oder Prüfungsebene[1]. In diesem Dokument werden Ihnen zuerst für die Kennzahlen auf Auswahlebene die Erläuterung, Berechnung und Interpretation vorgestellt. Danach finden Sie dasselbe für die Kennzahlen auf Itemebene (S. 3) und zum Schluss für die Kennzahl auf Prüfungsebene (S. 12).

Erläuterung, Berechnung und Interpretation Kennzahlen auf Auswahlebene

Ratewahrscheinlichkeit
Erläuterung
Die Ratewahrscheinlichkeit gibt an, wie gross die theoretische Wahrscheinlichkeit ist, dass Studierende eine Aufgabe durch Raten richtig lösen. Bei einer Single Choice-Aufgabe mit drei Auswahlmöglichkeiten beträgt die Ratewahrscheinlichkeit 1/3. Das heisst, durch reines Raten wählt eine Studierende zu 33% die korrekte Antwortmöglichkeit aus. Bei Aufgaben mit offenem Antwortformat beträgt die Ratewahrscheinlichkeit null; es ist nicht möglich durch reines Raten die korrekte Lösung auszuwählen, da die Lösung selbst entwickelt werden muss
Berechnung
$$Ratewahrscheinlichkeit = \frac{100}{Anzahl\ Auswahlmöglichkeiten}$$ Grenzwerte: Müssen selbst definiert werden. Für mögliche Vorschläge siehe Dokument Interpretation der Kennzahlen.

1 Auwahlebene: Kennzahlen die sich auf die Auswahl der einzelnen Antwortmöglichkeiten beziehen.
Itemebene: Kennzahlen die sich auf die einzelnen Items (Fragen) beziehen.
Prüfungsebene: Kennzahlen die sich die gesamte Prüfung beziehen.

Interpretation
Die Ratewahrscheinlichkeit wird beim Erstellen der Prüfung durch Wahl des Frageformats bestimmt. Ziel ist eine tiefe Ratewahrscheinlichkeit: je mehr plausible Antwortmöglichkeiten bei Single Choice- oder Multiple Choice-Aufgaben zur Verfügung stehen, desto kleiner wird die Wahrscheinlichkeit, dass durch reines Raten die korrekte Möglichkeit ausgewählt wird. Die Ratewahrscheinlichkeit benutzen wir für die Berechnung der korrigierten Antwortwahrscheinlichkeit.
Korrigierte Antwortwahrscheinlichkeit
Erläuterung
Die korrigierte Antwortwahrscheinlichkeit setzt die durchschnittlich erreichte Punktzahl der Aufgabe ins Verhältnis zur Ratewahrscheinlichkeit. Sie gibt somit an, wieviel Prozent der Studierenden die richtige Antwort gewählt haben, korrigiert um die Ratewahrscheinlichkeit.
Berechnung
$$Korrigierte\ Antwortmöglichkeit = \frac{ø\ erreichte\ Punktzahl\ pro\ Aufgabe\ in\ \%}{Ratewahrscheinlichkeit\ in\ \%}$$ Grenzwerte: Müssen selbst definiert werden. Für mögliche Vorschläge siehe Dokument Interpretation der Kennzahlen.
Interpretation
Diese Kennzahl zeigt, wie hoch im Durchschnitt die Wahrscheinlichkeit war, dass eine Antwortmöglichkeit durch «Überlegen», anstatt «Raten» gewählt wurde. Die korrigierte Antwortwahrscheinlichkeit ist damit eine Erweiterung der Ratewahrscheinlichkeit um das tatsächlich beobachtete Verhalten der Studierenden. Sie eignet sich insbesondere zur Analyse der Prüfung vor Notensetzung, da diese Kennzahl das Auffinden falsch hinterlegter Lösungen oder falsch gestellter Fragen, durch Grenzwertevaluation erlaubt.

Erläuterung, Berechnung und Interpretation Kennzahlen auf Itemebene

Absolutes und effektives Gewicht

Erläuterung

Das absolute Gewicht einer Aufgabe wird bereits beim Erstellen der Prüfung durch Zuweisung der maximal pro Aufgabe zu erreichenden Punktzahl festgelegt. Je mehr Punkte pro Aufgabe erreicht werden können, desto grösser ist das Gewicht dieser Aufgabe in der Prüfung. Wünscht man dementsprechend ein hohes Gewicht einer Aufgabe, definiert man eine hohe Maximalpunktzahl für diese Aufgabe.

Dieses im vorhinein festgelegte Gewicht kann sich jedoch vom effektiven Gewicht unterscheiden. Das effektive Gewicht ist abhängig von der tatsächlich durchschnittlich erreichten Punktzahl pro Aufgabe. Das heisst, das effektive Gewicht kann erst ermittelt werden, nachdem die Prüfung geschrieben worden ist.

Berechnung

$$Absolutes\ Gewicht = \frac{Maximalpunktzahl\ pro\ Aufgabe}{Gesamtpunktzahl}$$

$$Effektives\ Gewicht = \frac{ø\ pro\ Aufgabe\ erreichte\ Punktzahl}{ø\ erreichte\ Gesamtpunktzahl}$$

Grenzwerte: Müssen selbst definiert werden. Für mögliche Vorschläge siehe Dokument Interpretation der Kennzahlen.

Interpretation

Ziel ist es, dass absolutes und effektives Gewicht nahe beieinander liegen (oder sogar gleich sind). Dann hat jede Aufgabe auch tatsächlich das beabsichtigte Gewicht in der Prüfung. Liegen beide Werte stark auseinander, bedeutet das, dass nicht die gewünschte Gewichtung erreicht wurde. Beispielsweise wurde die Aufgabe zu schwer oder zu leicht gestellt.

Schwierigkeit

Erläuterung

Die Aufgabenschwierigkeit entspricht der Wahrscheinlichkeit, dass eine Studierende eine Aufgabe richtig löst. Sie ist normiert auf Werte zwischen null und eins. Je grösser der Wert dieser Kennzahl, desto grösser ist die Wahrscheinlichkeit einer korrekten Lösung und desto leichter ist die Aufgabe. Je schwieriger eine Aufgabe, desto kleiner die Wahrscheinlichkeit, dass diese korrekt gelöst wird und desto kleiner dann auch der durch diese Kennzahl ermittelte Wert.

Berechnung

$$Schwierigkeit = \frac{ø\ erreichte\ Punktzahl\ pro\ Aufgabe}{Maximalpunktzahl\ pro\ Aufgabe}$$

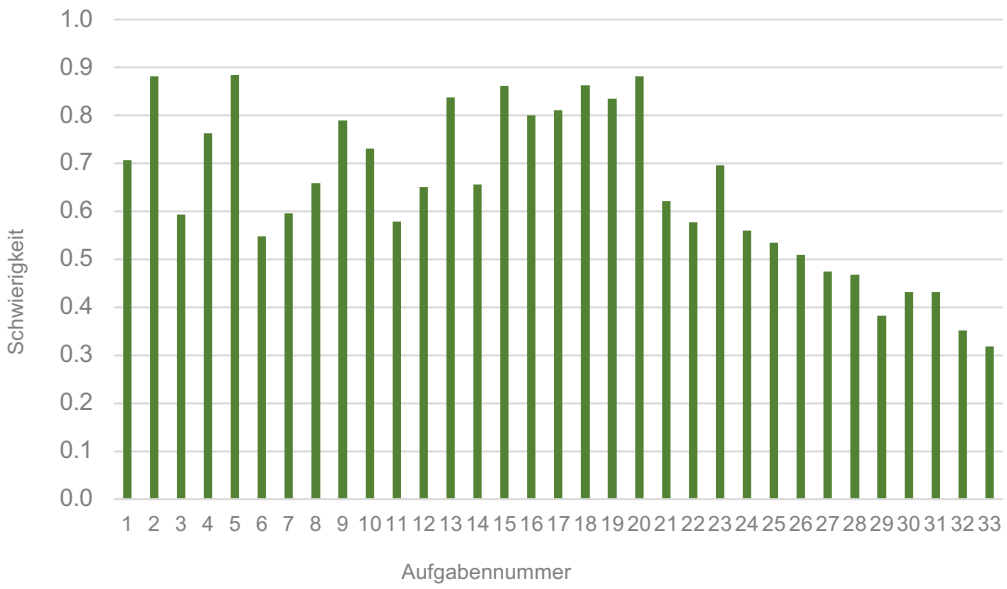
Grenzwerte:

Schwierigkeit > 0.8	leichte Aufgabe
Schwierigkeit ∈[0.2, 0.8]	mittelschwierige Aufgabe
Schwierigkeit < 0.2	schwierige Aufgabe

Interpretation

Diese Kennzahl ermöglicht es zu ermitteln, wie schwierig die einzelnen Aufgaben (und damit die Gesamtprüfung) tatsächlich war. In der folgenden Grafik sieht man die Schwierigkeit einzelner Aufgaben für eine analysierte Prüfung. Es ist ersichtlich, dass die Prüfung mit eher leichten Aufgaben begann (beispielsweise betrug bei Aufgabe 1 die Wahrscheinlichkeit, dass die Aufgabe richtig gelöst wurde, 70%) und gegen Ende schwieriger wurde (so wurde Aufgabe 33 nur noch mit einer Wahrscheinlichkeit von ca. 30% korrekt gelöst).[2]

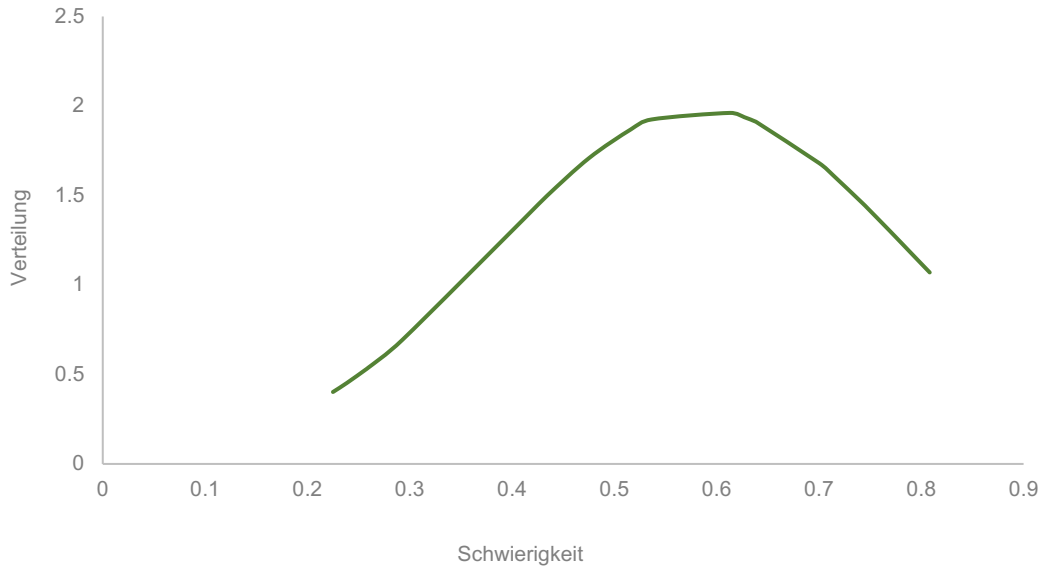
Schwierigkeit:



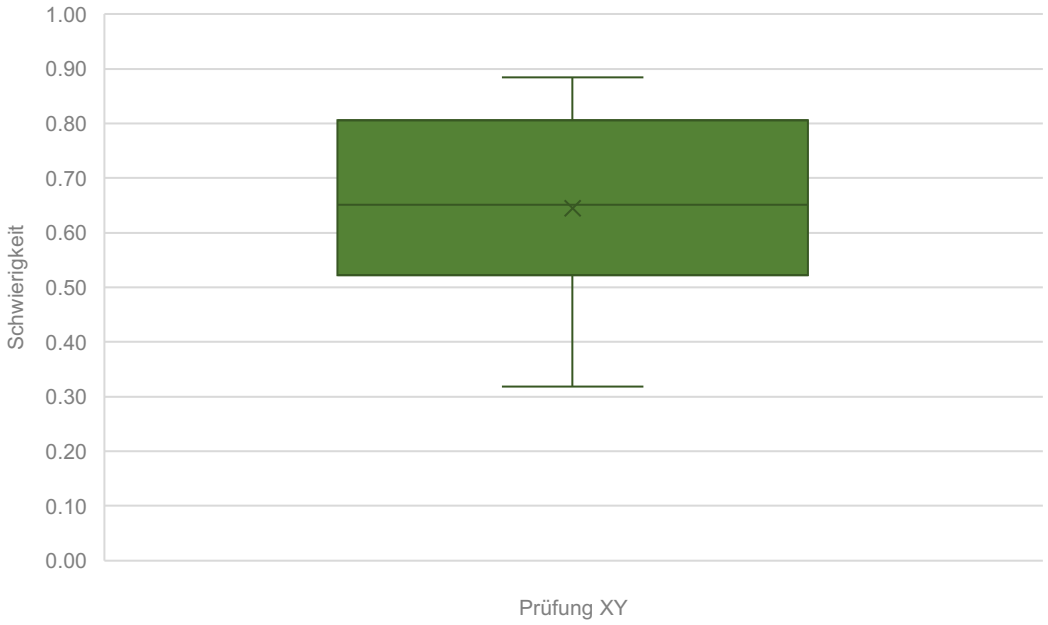
Ziel ist es, dass die Schwierigkeit einer Prüfung normalverteilt ist. Im folgenden Grafikbeispiel einer Prüfung ist ersichtlich, dass diese leicht linksschief ist. Die Prüfung besteht damit aus relativ vielen leichten Aufgaben (grosser Wert = leichte Aufgabe). Diese Erkenntnis zeigt auf, wie wichtig die Evaluation der Kennzahlen nach der Prüfung ist. Vor allem wenn es um die Qualitätssicherung der nächsten zu erstellenden Prüfung geht. Auch in einem Boxplot lässt sich diese Verteilung sehr gut ablesen und interpretieren.

[2] Bezüglich des Aufbaus der Prüfung ist die Wahl einer leichten Aufgabe für den Anfang der Prüfung oft am besten, damit die Prüflinge den Einstieg ins Thema einfacher schaffen (sogenannte Eisbrecher-Fragen).

Schwierigkeit Verteilung:



Schwierigkeit Boxplot:



Standardabweichung

Erläuterung

Die Standardabweichung gibt die durchschnittliche Streuung der Punktzahlen um ihren Mittelwert an und gibt erste Hinweise darauf, wie differenziert die jeweilige Aufgabe die Leistung der Studierenden misst, wie stark sie also die gewählte Skala ausschöpft.

Berechnung

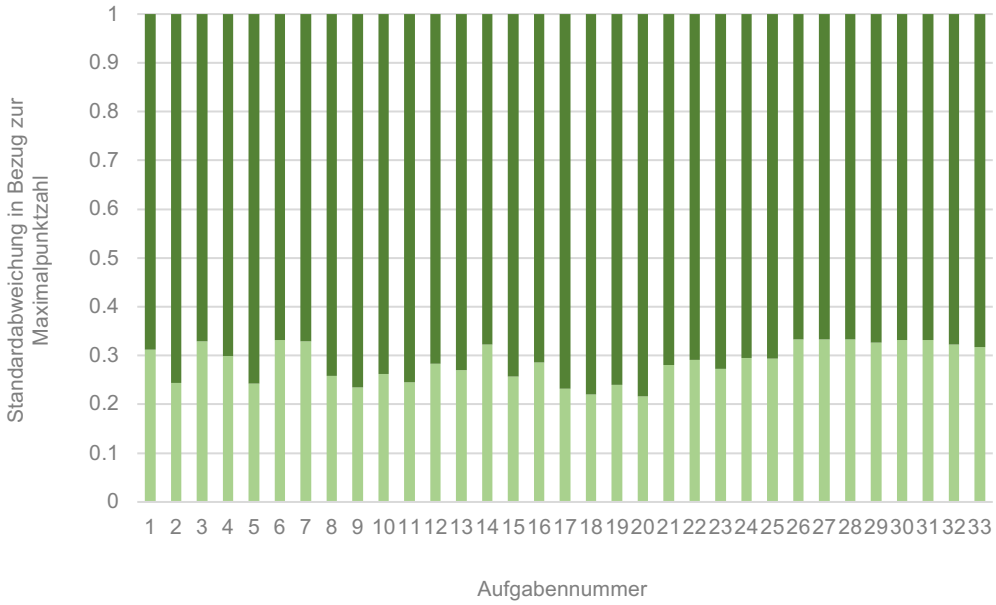
$$Standardabweichung = \frac{\Sigma\,(erreichte\ Punkte\ pro\ Aufgabe\ je\ Studierende - ø\ erreichte\ Punktzahl\ pro\ Aufgabe)^2}{Maximalpunktzahl\ pro\ Aufgabe - 1}$$

Grenzwerte: Müssen selbst definiert werden. Für mögliche Vorschläge siehe Dokument Interpretation der Kennzahlen.

Interpretation

Ziel ist eine relativ hohe Standardabweichung. Diese hohe Streuung würde die unterschiedlich gute Vorbereitung und somit unterschiedlich erreichten Punkte der Studierenden, wie auch beabsichtigt, widerspiegeln. Zeigt eine Aufgabe eine sehr kleine Streuung, kann dies im Umkehrschluss bedeuten, dass Studierende, die sehr viel gelernt haben, im Durchschnitt gleich viele Punkte erzielt haben, wie schlecht vorbereitete Studierende. Diese Aufgabe würde den Qualitätsstandards somit nicht entsprechen.
Die Interpretation der Standardabweichung hängt jedoch stark von Rahmenbedingungen ab: So sollte bei einer Prüfung mit nur wenig Studierenden, welche alle sehr gut auf die Prüfung gelernt haben, die Standardabweichung ebenfalls sehr klein sein. Dies aber nicht im negativen Sinne. Deshalb sollte diese Kennzahl darum immer zusammen mit der Trennschärfe analysiert werden.

Standardabweichung:



Trennschärfe

Erläuterung

Die Trennschärfe gibt an, wie gut eine Frage zwischen leistungsstarken und -schwachen Studierenden unterscheidet. Sie ist ein Mass für die Korrelation einer Aufgabe mit dem Gesamtergebnis. Eine hohe Trennschärfe liegt dann vor, wenn Studierende mit einer hohen Gesamtpunktzahl auch bei der jeweiligen Aufgabe eine hohe Punktzahl erzielt haben. Maximale Trennschärfe bedeutet, dass die Frage von keinem der leistungsschwachen, aber von allen leistungsstarken Studierenden gelöst werden konnte.

Berechnung

$$Trennschärfe = Korrelation\ ø\ erreichte\ Punktzahl\ pro\ Aufgabe\ mit\ ø\ Gesamtpunktzahl$$

Grenzwerte:

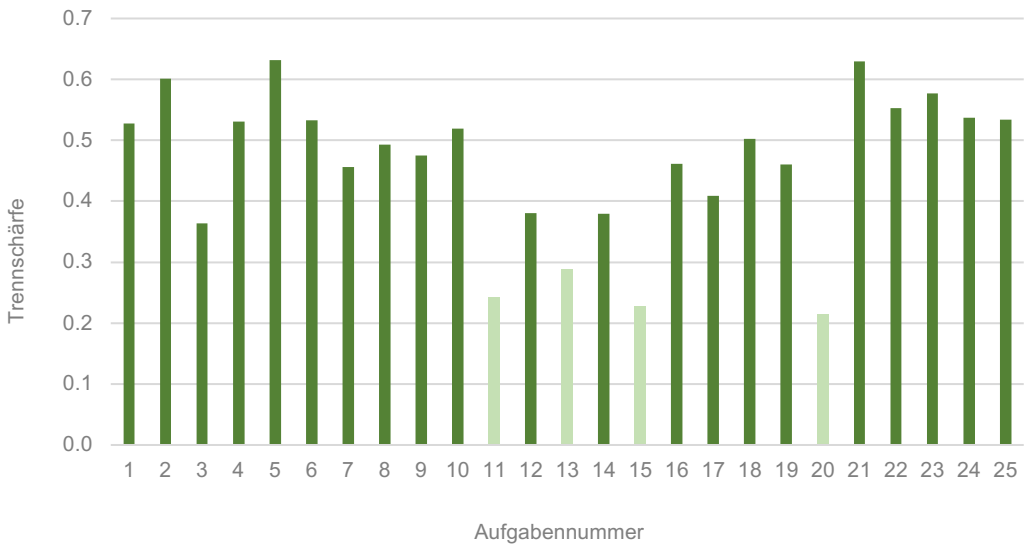
Trennschärfe > 0.5	sehr gut
Trennschärfe ∈[0.3, 0.5]	akzeptabel
Trennschärfe < 0.3	needs work

Interpretation

Die Trennschärfe setzt die erreichte Punktzahl pro Aufgabe ins Verhältnis zu der je Studierenden erreichten Gesamtpunktzahl. Es herrscht die Grundannahme, dass die erreichte Gesamtpunktzahl die Leistungsstärke der Studierenden repräsentiert. Wünschenswert sind Aufgaben, die von Studierenden mit einer hohen Gesamtpunktzahl (gut vorbereitet) korrekt, von Studierenden mit einer tiefen Gesamtpunktzahl (schlecht vorbereitet) hingegen schlecht gelöst werden. Die Trennschärfe hilft damit auch zu beurteilen, wie fair eine Prüfung war.

Ziel ist es, eine möglichst hohe Trennschärfe für alle Aufgaben zu erhalten. Aufgaben mit einer tiefen Trennschärfe sollten unbedingt genauer analysiert werden. Beispielsweise könnte die Aufgabe irreführend formuliert oder die falsche Lösung hinterlegt worden sein. Auch ein Abgleich mit der Kennzahl der Schwierigkeit könnte Aufschluss über das dahinterliegende Problem geben. In der folgenden Grafik sieht man beispielsweise, dass Aufgaben 11, 13, 15 und 20 genauer analysiert werden sollten.

Trennschärfe:



P-Wert

Erläuterung

Der P-Wert ist der Anteil der richtigen Antworten der Studierenden einer Frage. Er lässt sich mit der Schwierigkeit vergleichen. In Abhängigkeit der Anzahl Antwortmöglichkeiten einer Aufgabe sollte der P-Wert innerhalb eines Intervalls liegen. Je nach Anzahl Antwortmöglichkeiten ist also auch ein anderer P-Wert wünschenswert. Während die Kennzahl der Schwierigkeit angibt, wie schwierig eine Prüfung oder eine Aufgabe war, gibt der P-Wert ergänzend dazu an, ob dieser Wert der Schwierigkeit auch optimal ist.

Berechnung

$$P - Wert = \frac{Anzahl\ Studierende, die\ die\ Frage\ richtig\ beantwortet\ haben}{Anzahl\ Studierende\ insgesamt}$$

Grenzwerte:

2 Antwort-möglichkeiten	0.61 — 0.75 — 0.9
3 Antwort-möglichkeiten	0.5 — 0.67 — 0.9
4 Antwort-möglichkeiten	0.44 — 0.62 — 0.9
Offene Frage	0.25 — 0.5 — 0.9

Interpretation

Der P-Wert kann als eine Mischung zwischen Schwierigkeit und korrigierter Antwortwahrscheinlichkeit gesehen werden: Er gibt an, ob im Verhältnis zur Anzahl Auswahlmöglichkeiten die «Schwierigkeit» im optimalen Intervall liegt.

Erläuterung, Berechnung und Interpretation Kennzahlen auf Prüfungsebene

Cronbach's Alpha

Erläuterung

Cronbach's Alpha gibt an, inwieweit eine bestimmte Aufgabe eine Gruppe von Aufgaben (oder die Gesamtprüfung) repräsentiert. Damit lässt sich erkennen, wie gut die Aufgaben zusammenpassen. Ferner ist das *Cronbach's Alpha* die Untergrenze der Reliabilität. Die Reliabilität einer Prüfung gibt an, wie gut eine Prüfung den Leistungsstand eines Studierenden misst. Die Reliabilität ist tendenziell höher, je mehr Aufgaben Teil einer Prüfung sind. Dies zeigt sich auch im Cronbach's Alpha: Diese Kennzahl kann durch Hinzufügen von weiteren Aufgaben erhöht werden.

Berechnung

$$Cronbach's\ Alpha = \frac{Anzahl\ Aufgaben}{Anzahl\ Aufgaben - 1} \times \frac{1 - Summe\ der\ Varianz\ der\ Items}{Varianz\ der\ Summe\ der\ Items}$$

Grenzwerte:

Cronbach's Alpha ≥ 0.8	gut
Cronbach's Alpha < 0.8	needs work

Interpretation

Diese Kennzahl gibt Auskunft über die Gesamtprüfung. Sie misst den Grad der Übereinstimmung zwischen den Fragen der Prüfung. Ist das *Cronbach's Alpha* sehr klein, sind die Prüfungsaufgaben nicht kohärent. Das bedeutet, es kann beispielsweise sein, dass gewisse Fragen nicht das eigentliche Thema der Prüfung abfragen.

Quellen

Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion* (2., aktualisierte und erw. Aufl.). München: Pearson Studium.

Dany, S., Szczyrba, B. & Wildt, J. (2008). *Prüfungen auf die Agenda!: Hochschuldidaktische Perspektiven auf Reformen im Prüfungswesen (Blickpunkt Hochschuldidaktik)* (1. Aufl.). wbv Media.

Krebs, R. (2004). *Anleitung zur Herstellung von MC-Fragen und MC-Prüfungen für die ärztliche Ausbildung.* Bern: Institut für Medizinische Lehre IML, Abteilung für Ausbildungs-und Examensforschung AAE. Abrufbar auf: www.iml.unibe.ch

Möltner, A., Schellberg, D., & Jünger, J. (2006). *Grundlegende quantitative Analysen medizinischer Prüfungen.* GMS Z Med Ausbild, *23*(3), 2006-23.

Schlomske-Bodenstein, N., Strasser, A., Schindler, C., & Schulz, F. (999). *Handreichungen zum kompetenzorientierten Prüfen.* ProLehre, TU München. Abrufbar auf: https://www.prolehre.tum.de/materialien-und-tools/handreichungen/

Stichting Cito, Instituut voor toets- en examenontwikkeling (2018). De waarde van de ritwaarde. Arnhem.

van Berkel, H., Bax, A., & Joosten-ten Brinke, D. (2017). *Toetsen in het hoger onderwijs*. Springer. Doi: 10.1007/978-90-368-1679-3