Klausurauswertungen an der Medizinischen Fakultät der Julius-Maximilians-Universität Würzburg: Ein Praxisbeispiel

Stand: 11.05.2026 – Version für Website

Autor: Dr. rer. nat. Tobias Leutritz

Inhalt

Einführung

Derzeit erfolgt die Klausurauswertung über ein server-basiertes System[1].

Als *Eingabe*dateien werden die CaseTrain-Pakete (enthalten Aufgabenstellung und Metadaten zur Klausur), sowie die Excel-Auswertungsdatei von CaseTrain zur Verfügung gestellt.

Zur *Ausgabe* stehen zurzeit folgende Versionen zur Verfügung, die im Anschluss an die Durchführung einer Klausur vom Prüfungsteam erstellt und an die Prüfungsverantwortlichen versandt werden):

Der Kurzbericht umfasst

- Kenndaten der Leistungskontrolle
- Kennwerte der Aufgaben (summarisch; am Ende tabellarisch einzeln aufgelistet)
- Notenverteilung (Punkteverteilung als *Tooltip* in o. g. Tabelle)
- Bestehensgrenzen mit Hinweisen zum Notenshift
- summarische Darstellung zu Kommentaren (Auffälligkeiten bei vielen Kommentaren)
- grafische Darstellung der Kennwerte Trennschärfe und Schwierigkeit (als Histogramme im *Tooltip* der o. g. Tabelle)
- Richtwerte für Schwierigkeit, Trennschärfe und Diskriminationsindex
- Aufgabenstellung mit tabellarischer Angabe der korrekten Lösung(en) als *Tooltip* bei Nennung der Aufgabennummern bzw. in der Übersichtstabelle

Die ausführliche Darstellung enthält

- die Aufgabenstellung mitsamt Bildmaterial und Antwortmöglichkeiten
- zusätzliche Angaben (grafisch/tabellarisch) zu den ausgewählten Antwortoptionen
- sowie weitere grafische Darstellungen der Kennwerte (als Balkendiagramme und Histogramme) und der Punkteverteilung
- optional: zusätzliche Grafik je Aufgabe zur Trennschärfe

Im Folgenden soll nun anhand einer Medizindidaktik-Klausur aus dem Sommersemester 2024 auf Einzelheiten der Berichte eingegangen und Besonderheiten von Ergebnissen diskutiert werden.

[1] T. Leutritz, A. Hörnlein, A. Weingart, M. Appel, A. Entwistle, S. König: Mehr Überblick, weniger Aufwand: Dynamische Prüfungsberichte auf Knopfdruck. Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA). 08.-10.09.2025 in Düsseldorf. doi:10.3205/25gma252

Kurzbericht

Download des Berichts

Kenndaten der Leistungskontrolle

Kennzahlen[1]	Werte	Grafik
Anzahl der Aufgaben, Σ (ohne Wertung)	49 (0)	
max. erreichbare Punktzahl	49	
erreichte Punkte [min. - max.]	17 - 46	
Punktzahl, Ø (SD)	34 (8,9)	
Prüfungsdauer in Minuten	11010	
Bearbeitungszeit in Minuten, Ø (SD)	98 (57)	
Schwierigkeit ❓ [%], Ø (SD) [%]	70 (16)	
Trennschärfe ❓, Ø (SD)	0,43 (0,22)	
Diskriminationsindex ❓, Ø (SD)	0,36 (0,21)	
Cronbachs alpha ❓	0,92	

[1] Σ = Gesamt, Ø = Mittelwert, SD = Standardabweichung

Die obenstehende Tabelle stellt den ersten Anhaltspunkt im Kurzbericht dar und listet die wichtigsten Kennzahlen der Klausur auf. Den *Tooltip*s bei den Fragezeichen sind Erläuterungen zu den Kennzahlen

- Aufgabenschwierigkeit,
- Trennschärfe,
- Diskriminationsindex und
- Cronbachs alpha (CA)

zu entnehmen, die nachfolgend am Beispiel erläutert werden. Im letzten Abschnitt des Kurzberichts befindet sich die tabellarische Übersicht der Kennwerte je Aufgabe.

Zudem finden sich erste Einblicke in Häufigkeitsverteilungen der erzielten Punkte bzw. der o. g. Kennzahlen (außer CA), wenn man mit dem Mauszeiger über den entsprechenden Histogrammen verweilt.

Zunächst steht jedoch die Notenverteilung und eine mögliche Anpassung der Bestehensgrenze als nächster Abschnitt im Vordergrund.

Notenverteilung und Bestehensgrenzen

	Notengrenze		Prüflinge		0 % 25 % 50 % 75 % 100 %
Note	Prozent	Punkte	Anzahl	Anteil	
1	90	44	3	14	
2	80	39	6	27	
3	70	34	3	14	
4	60	29	4	18	
5	0	0	6	27	

Die obige Tabelle zeigt die Notenverteilung der ursprünglichen Auswertung. Daraufhin sind die Bestehensgrenzen entsprechend der Studienordnung tabellarisch dargestellt, um die Gleitklausel zu überprüfen (hier die Werte der gegebenen Klausur):

Bestehensgrenze	Punkte	Prozent	Kommentar
absolut	29,4	60	in der Regel 60 % der Gesamtpunktzahl (Studienordnung Humanmedizin Würzburg)
relativ	26,77	55	Abweichung um nicht mehr als 22 % von der durchschnittlichen Prüfungsleistung aller Erstteilnehmenden, mindestens jedoch 50 %

Ist die *relative* Bestehensgrenze kleiner als die *absolute* Bestehensgrenze, ist anhand der Erstteilnehmenden zu überprüfen, ob die Gleitklausel angewendet werden sollte.

In diesem Fall **muss** also die Gleitklausel angewendet werden (55 < 60).

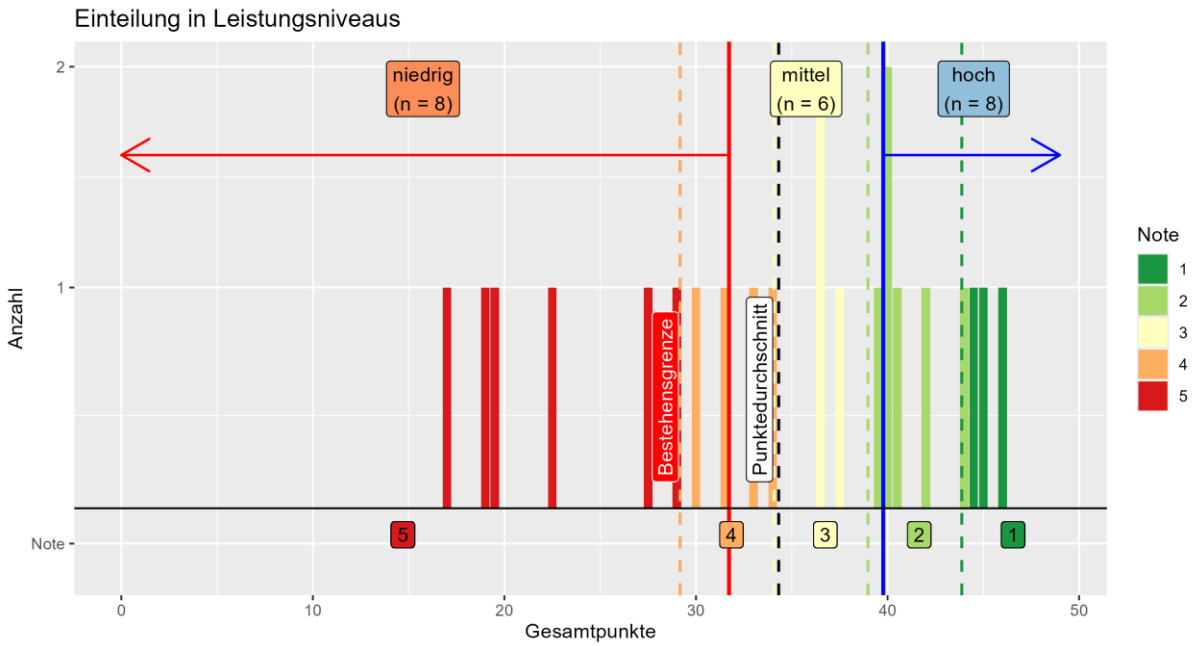
Es folgt daher der Vorschlag einer angepassten Bestehensgrenze und entsprechender Notengrenzen (Notenshift):

	Notengrenze		Prüflinge		0 % 25 % 50 % 75 % 100 %
Note	Prozent	Punkte	Anzahl	Anteil	
1	89	44	4	18	
2	78	38	5	23	
3	66	32	5	23	
4	55	27	4	18	
5	0	0	0	0	

Kommentare

Während der Klausur ist es den Prüflingen möglich, Kommentare zu Aufgaben zu hinterlassen. Einen ersten Überblick, wie viele Kommentare pro Aufgabe abgegeben wurden, liefert die entsprechende Tabelle im Kurzbericht, da **viele** Kommentare oft auf Probleme mit der Aufgabe hinweisen.

Punkteverteilung



Einen ersten Überblick über das Abschneiden der Prüflinge liefert das Histogramm der Punkteverteilung (im Kurzbericht nur als *Tooltip* in der ersten Tabelle; im ausführlichen Bericht auch separat einsehbar).

Hierin sind farblich auch die Noten markiert (siehe Legende rechts) und durch gestrichelte Linien abgegrenzt.

Zudem werden die Prüflinge in Gruppen mit hoher, mittlerer und niedriger Leistungsfähigkeit in dieser Klausur eingeteilt. Diese Einteilung wird für die Berechnung des Diskriminationsindex benötigt und orientiert sich in der Regel am oberen und unteren 27 %-Perzentil (nach Möltner et al.).

Aufgabenschwierigkeit

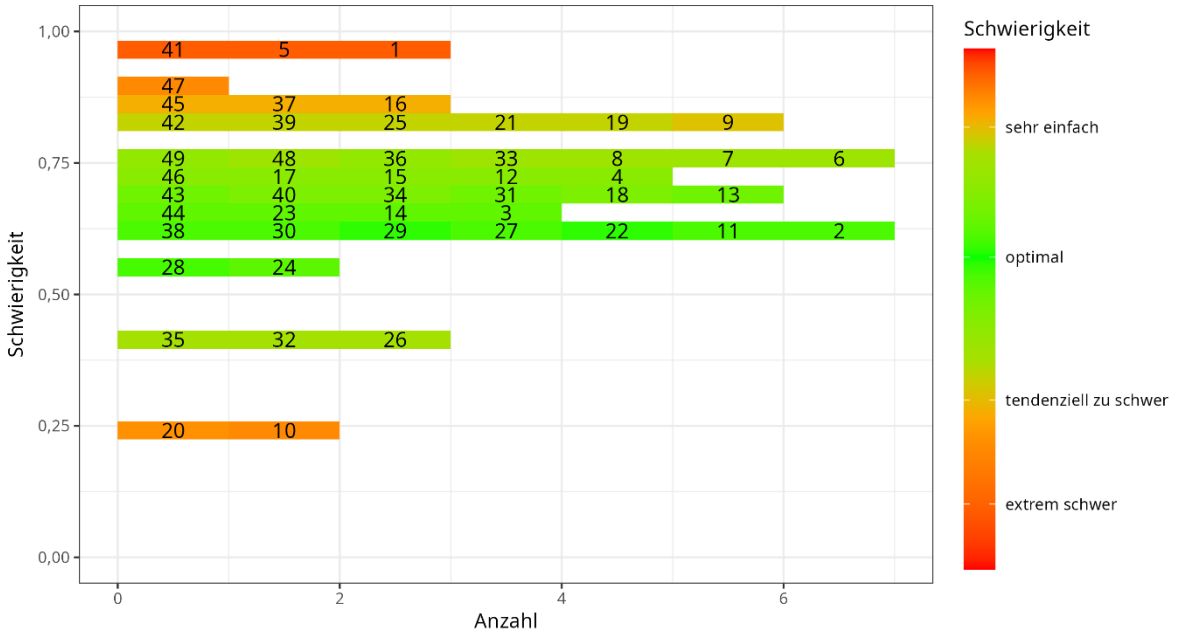
Definition:
mittlere, relative erreichte Punktzahl p, bezogen auf die maximal erreichbare Punktzahl

Synonyme: Erfolgsrate, Lösungswahrscheinlichkeit, Englisch: difficulty

Beachte:
p liegt zwischen 0 und 1 (oder 0 % und 100 %),
je größer p, desto leichter die Aufgabe

Richtwert:
p sollte im Bereich von 0,4 – 0,8 streuen

Histogramm für die Beispielklausur:



Dem obigen Histogramm ist die Verteilung der Aufgabenschwierigkeiten unter Angabe der Aufgabennummern zu entnehmen. Auf der Farbskala (rechts) ist die Einteilung in bestimmte Bereiche angegeben, die sich auch im Abschnitt „Richtwerte für Kenndaten und Einordnung der Aufgaben“ wiederfinden (sowohl im Kurzbericht als auch in der ausführlichen Variante):

Bereich	**0 < p ≤ 0,25**	**0,25 < p ≤ 0,4**	**0,4 < p ≤ 0,8**	**0,8 < p ≤ 0,9**	**p > 0,9**
Einordnung	extrem schwer	tendenziell zu schwer	optimale Schwierigkeit	sehr einfach	extrem leicht
Aufgaben	10, 20	(keine)	2, 3, 4, 6, 7, 8, 11, 12, 13, 14, 15, 17, 18, 22, 23, 24, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 38, 40, 43, 44, 46, 48, 49	9, 16, 19, 21, 25, 37, 39, 42, 45	1, 5, 41, 47

So lassen sich hier bereits Tendenzen erkennen, in welchem Schwierigkeitsbereich sich die einzelnen Aufgaben befinden und ob es zu einer Häufung in kritischen Bereichen kommt.

Zur Darstellung der Auswahlhäufigkeiten einzelner Antwortoptionen siehe ausführlicher Bericht.

Trennschärfe

Definition:
(Pearson-)**Korrelation** r‘ von erreichter Punktzahl mit der Gesamtsumme aller anderen Aufgaben in der Prüfung (“part-whole-korrigierte” Korrelation).
Die Trennschärfe gibt an, ob mit der Aufgabe bezogen auf die Gesamtpunktzahl “starke” und “schwache” Prüflinge unterschieden werden können.

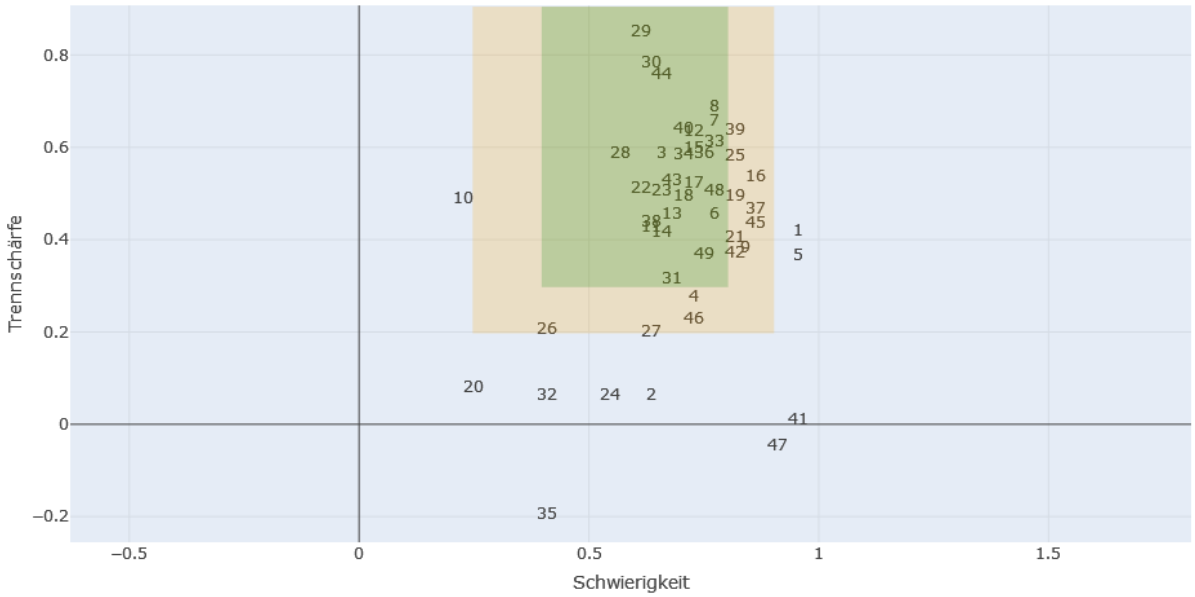
Synonym: (Item-)Selektivität (Englisch: selectivity)

Richtwerte:

r' ≥ 0,3	gut (Möltner, Schellerg, und Jünger 2006): hohe Trennschärfe, d. h. Prüflinge, die eine Aufgabe/Frage bzw. ein Item richtig beantworten, schneiden auch insgesamt gut ab
0,2 ≤ r' < 0,3	akzeptabel
0 ≤ r' < 0,2	gering
r'< 0	schlechte Trennschärfe: "starke" Prüflinge sind verwirrt, erreichen weniger Punkte als "schwache"

An dieser Stelle soll nun die gemeinsame Darstellung von Aufgabenschwierigkeit und Trennschärfe zur Darstellung kommen, die erstere auf der Abszisse und die Trennschärfe auf er Ordinate aufträgt und die Lage in diesem neuen Koordinatensystem mit deren Aufgabennummer kenntlich macht.

Der **grün** dargestellte Bereich umfasst dabei das Optimum der Werte für Trennschärfe (> 0,3) und Schwierigkeit (0,4-0,8), **orange** die akzeptablen Bereiche.



Unterhalb dieser Grafik erfolgt nun eine Auflistung von Aufgaben, die auffällig hinsichtlich übermäßiger Auswahl von Distraktoren sind, d. h. die Auswahlhäufigkeit überschreitet eine angenommene Ratewahrscheinlichkeit von 50 % für die Lösung durch die Anzahl der Distraktoren.

Für eine eingehendere Analyse der Auswahlhäufigkeiten und Trennschärfe eignen sich die grafischen Darstellungen der ausführlichen Auswertung.

Diskriminationsindex

Definition:
Der Diskriminationsindex ist die Differenz zwischen den Werten der Schwierigkeit für Prüflinge mit hoher und niedriger Leistungsfähigkeit (oberes vs. unteres Drittel bzw. 27/73-stes Quantil nach Möltner bzgl. Gesamtpunktzahl - s. Abschnitt Punkteverteilung).

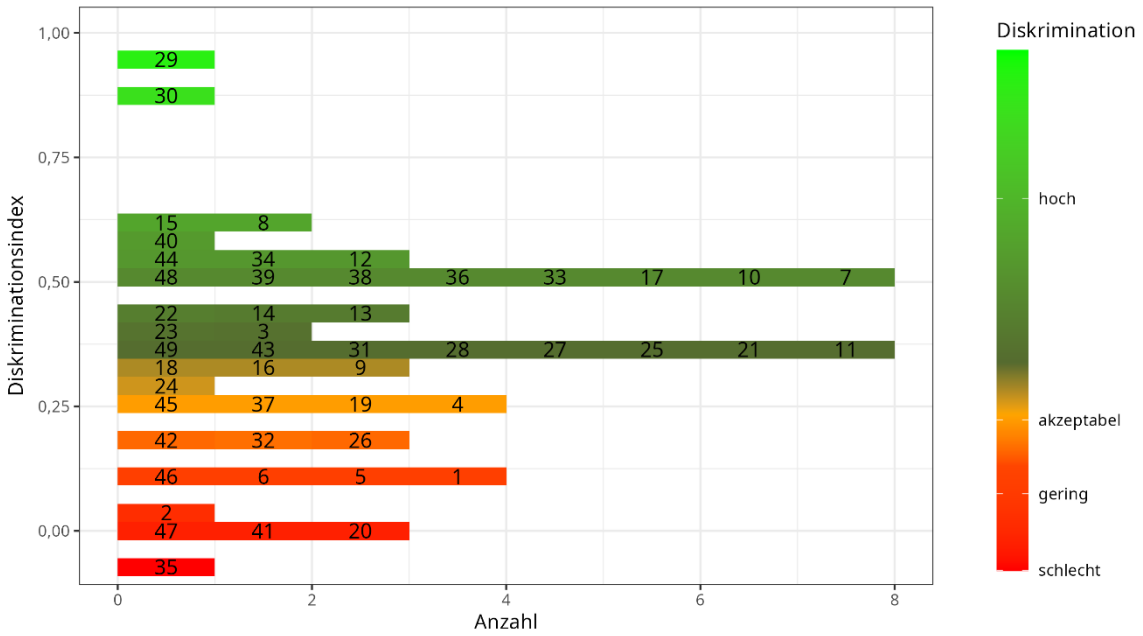
Synonym: Trennfähigkeitsindex, Englisch: discrimination</p

Richtwerte:
Zum **Diskriminationsindex D** schreibt A. Möltner (Möltner, Schellerg, und Jünger 2006):

„Trennt die Aufgabe gut, so ist die Differenz D offensichtlich groß, trennt sie schlecht, liegt sie nahe bei 0, negative D erhält man bei […] paradoxen Antwortmustern.“

Eine detailliertere Unterteilung mit Empfehlungen zur Überarbeitung der Aufgaben gibt E. Backhoff (Backhoff, Larrazolo, und Rosas 2000), ergänzt um die **hier** angewandten Grenzen in der ersten Spalte und einer Übersetzung der Kommentare in der letzten Spalte:
Table I. Discrimination power of the answers according to their D value

hier:	**D =**	**Quality**	**Recommendations**	**Kommentar**
D’ ≥ 0,4	> 0.39	Excellent	Retain	Aufgabe/Item unterscheidet effektiv zwischen leistungsstarken und -schwachen Prüflingen
0,3 ≤ D’ < 0,4	0.30 - 0.39	Good	Possibilities for improvement	Potential zur Verbesserung
0,2 ≤ D’ < 0,3	0.20 - 0.29	Mediocre	Need to check/review	Kontrolle erforderlich
0 ≤ D’ < 0,2	0.00 - 0.20	Poor	Discard or review in depth	Notwendigkeit zur Überarbeitung
D’ < 0	< - 0.01	Worst	Definitely discard	dringende Notwendigkeit zur Überarbeitung



Das obige Histogramm zeigt die Verteilung von D in der Beispielklausur und deutet auf Probleme mit den Aufgaben geringer und schlechter (sogar negativer) Diskrimination hin.

Die Darstellung der Auswahlhäufigkeiten einzelner Antwortoptionen im ausführlichen Bericht machen dies deutlich.

Cronbachs Alpha

Definition:
Cronbach´s Alpha (“interne Konsistenz”) ist der Grad der Genauigkeit, mit dem ein Merkmal gemessen wird – ganz egal, was gemessen wird.
Je homogener ein Konstrukt misst, desto höher fällt der Reliabilitätskoeffizient aus (Wertebereich zwischen 0 und 1).

Richtwert: > 0,7

Kenndaten der Aufgaben

Im letzten Abschnitt des Kurzberichts finden sich die Kenndaten für Aufgabenschwierigkeit, Trennschärfe und Diskriminationsindex in tabellarischer Form. Durch Klick auf die Spaltentitel lassen sich die Werte auf- bzw. absteigend sortieren. Mit der Suchfunktion (oben rechts) lässt sich die Anzeige filtern, bspw. nach bestimmten Aufgabentypen oder Suchworten auch innerhalb der nicht direkt angezeigten Aufgabenstellung bzw. den Antwortoptionen.

Aufgabenstellung und Angabe der Lösung(en)

Die gesamte Aufgabenstellung mit tabellarischer Angabe der korrekten Lösung(en) kann über einen *Tooltip* angezeigt werden, wenn man mit dem Mauszeiger über den entsprechenden Aufgabennummern verweilt wird. Diese erscheinen an allen Stellen, wo Aufgaben in den Berichtstexten genannt werden:

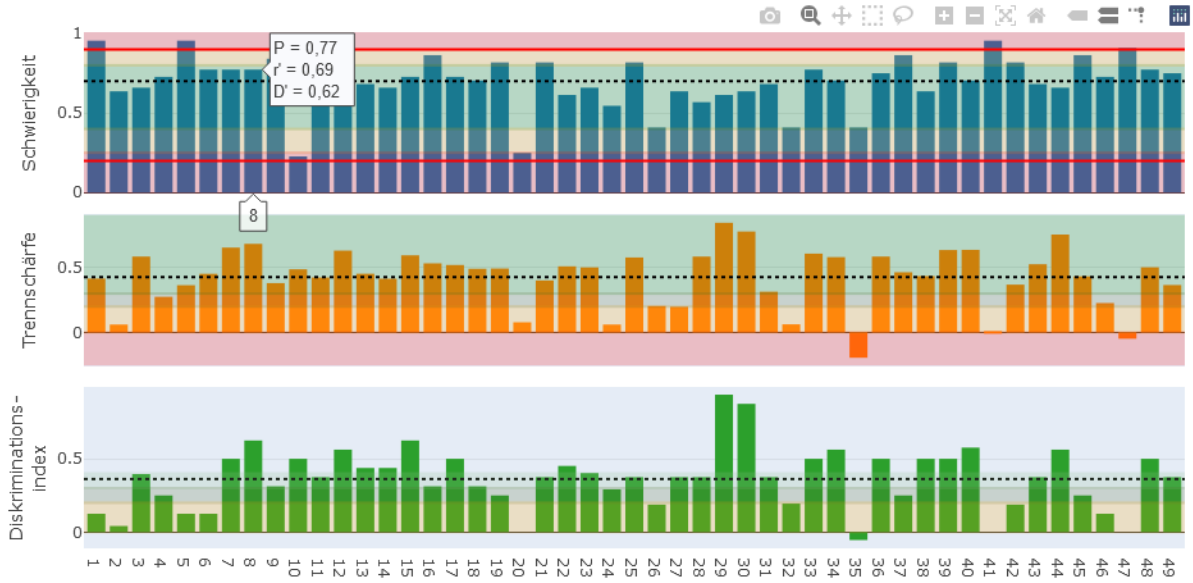
- in den o. g. Tabellen zu Richtwerten der Kennzahlen oder
- unterhalb der gemeinsamen Darstellung von Aufgabenschwierigkeit und Trennschärfe bzgl. Auffälligkeiten.

In der ausführlichen Darstellung kann dorthin direkt per Link gesprungen werden (auch aus den grafischen Darstellungen heraus - außer den Histogrammen).

Ausführliche Darstellung

Download des Berichts

Grafische Darstellungen der Kennwerte für alle Aufgaben



Die obige Grafik zeigt die Kennwerte für Aufgabenschwierigkeit, Trennschärfe und Diskriminationsindex aller Aufgaben der Beispielklausur als Balkendiagramme übereinander.

In der Symbolleiste oben rechts finden sich einige Werkzeuge zum Speichern, Zoomen und wiederherstellen der Standardansicht (Haussymbol).

Belässt man den Mauszeiger über den Balken, werden die entsprechenden Werte (obere beide Grafiken zu Aufgabenschwierigkeit und Trennschärfe) und Verbindungslinien zwischen den einzelnen Grafiken angezeigt, um die verschiedenen Werte besser miteinander vergleichen zu können.

Zusätzlich werden in der unteren Grafik zum Diskriminationsindex (s. u.) auch die Aufgabenstellung mit Lösung(en) angezeigt. Tipp hierzu: von unten an die Balken heranfahren.

Details zu den Aufgaben

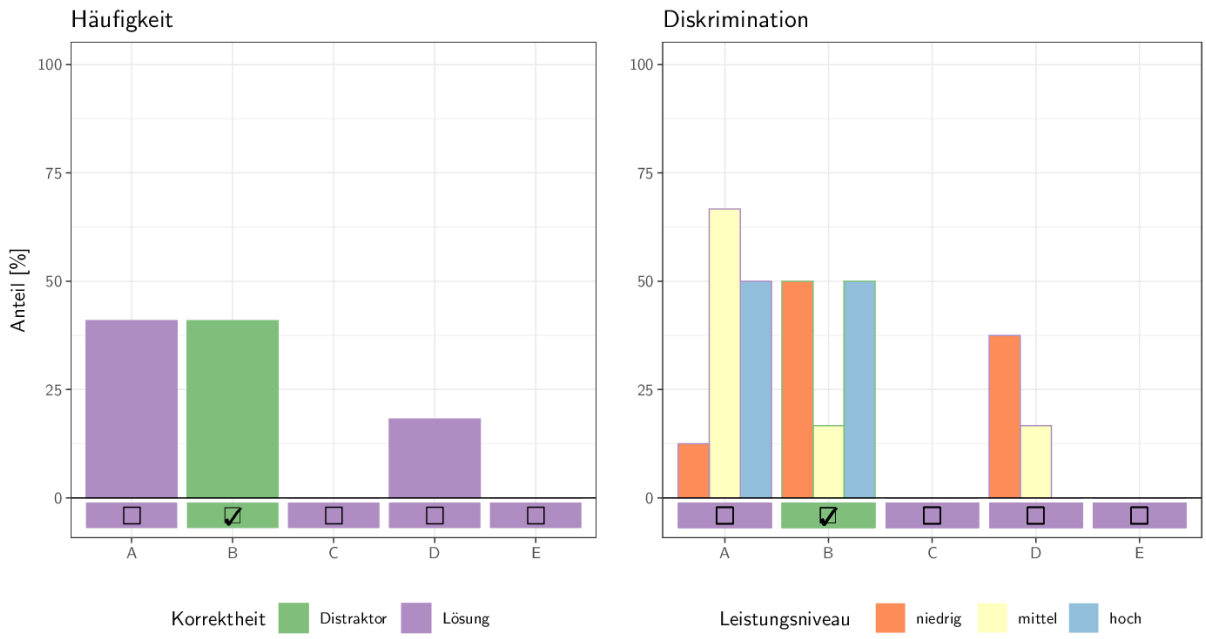
Die Aufgaben werden mit folgenden Details dargestellt:

- Aufgabentyp und Abschnitt, sowie Anzahl der Antwortmöglichkeiten und geforderten/richtigen Antworten (z. B.: „Aufgabentyp: TypA, Abschnitt: 35, Antwortmöglichkeiten: 5, Antworten gefordert: 1“)
- gesamte Aufgabenstellung mit Vignette und Grafiken
- Antwortoptionen mit Angabe der Richtigkeit sowie Schwierigkeit p, Trennschärfe r und Diskriminationsindex D (einzeln berechnet; der Übersichtlichkeit halber stehen die Werte bei Aufgaben des Typs EMQ und Kprim mit zusätzlicher Auflistung der Werte für alle Distraktoren nach der grafischen Darstellung und können durch Klick auf die Spaltenüberschriften beliebig umsortiert werden; für Text-, Wort- und Zahlenfragen erfolgt

zunächst eine aggregierte Ausgabe - absteigend sortiert nach Punkten mit Angabe der Anzahl von gleichen Antworten in Klammern dahinter, getrennt durch Komma - erst danach die volle Ausgabe mit Einzelwerten je gegebener Antwortoption)

- Kommentare der Prüflinge zu der Aufgabe
- Kennwerte p, r' und D' für die Aufgabe insgesamt mit Farbkodierung entsprechend der Richtwerte zu Aufgabenschwierigkeit, Trennschärfe und Diskriminationsindex (z. B.: „Aufgabenschwierigkeit p: 0,41, Trennschärfe r': -0,2, Diskriminationsindex D': -0,056"
- grafische Darstellung der Auswahlhäufigkeiten insgesamt bzw. unterteilt nach Leistungsgruppen (der Übersichtlichkeit halber bei EMQ nur Kombinationen, die final gewählt wurden), sowie (optional) die Visualisierung der Trennschärfe.

Grafische Darstellung der Auswahlhäufigkeiten je Antwortoption



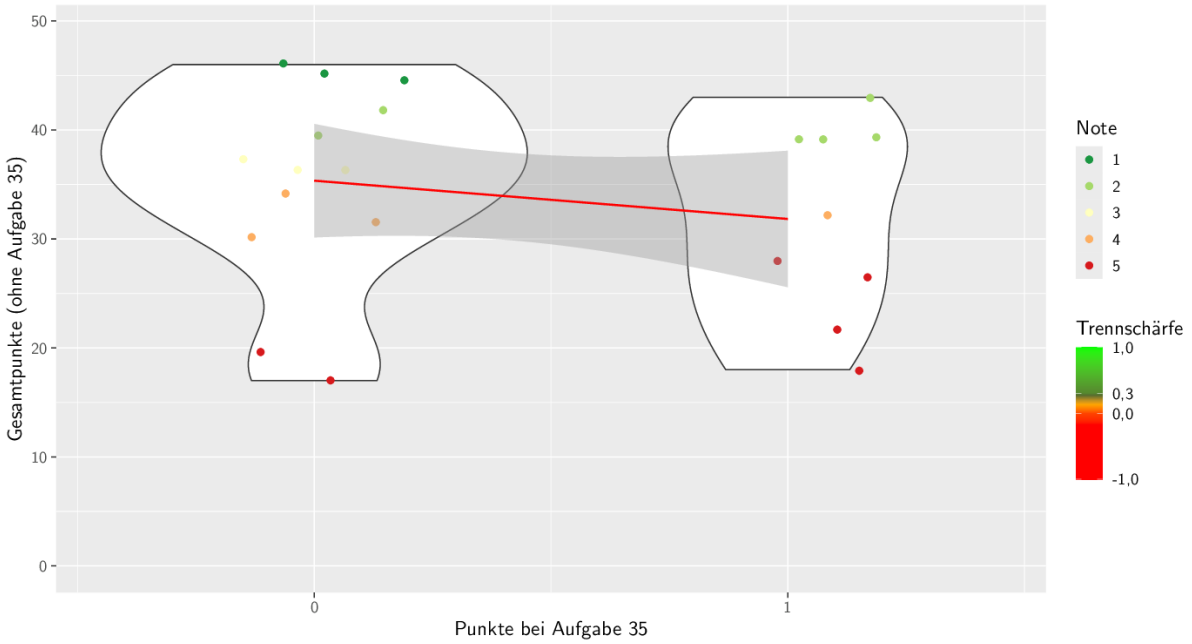
Die obige Grafik zeigt die Auswahlhäufigkeiten (links: insgesamt; rechts: nach Leistungsgruppen) der Aufgabe 35 der Beispielklausur.

Dabei wird links deutlich, dass Distraktor A die gleiche Auswahlhäufigkeit aufweist, wie die Lösung. Rechts wird dabei sichtbar, dass besonders Prüflinge der hohen Leistungsgruppe diese (falsche) Antwortoption ausgewählt haben. Daher rührt auch der negative Diskriminationsindex von **-0,056.**

Grafische Darstellung der Trennschärfe je Aufgabe (optional)

Die Visualisierung der Trennschärfe veranschaulicht die Berechnung der Trennschärfe: die Gesamtpunktzahl (abzüglich der erreichbaren Punktzahl der jeweiligen Aufgabe) aller Prüflinge wird aufgetragen über der Punktzahl bei der jeweiligen Aufgabe. Zur besseren Sichtbarkeit der einzelnen Punkte sind diese gestreut und entsprechend der Notenverteilung eingefärbt.

Zusätzlich sind Geigenplots (Violin-Plots) hinterlegt, die die Verteilung der erzielten Gesamtpunkte in Form einer gespiegelten Verteilungskurve zeigt, deren Breite der Anzahl von Prüflingen mit gleicher Punktzahl entspricht.



Hier eine negative Trennschärfe von **-0,2** (Regressionsgerade daher rot, weil negative Werte eine schlechte Trennschärfe darstellen), wo Prüflinge mit besseren Noten (siehe Farbkodierung der Punkte) häufiger die falsche(n) Option(en) gewählt haben (siehe Breite der Violin-Plots, die die Häufigkeit der erzielten Punkte darstellen).

Literaturverzeichnis

Backhoff, E., N. Larrazolo, und M. Rosas. 2000. „The Level of Difficulty and Discrimination Power of the Basic Knowledge and Skills Examination (EXHCOBA)“. Journal Article. *Revista Electrónica de Investigación Educativa* 2 (1): 16. http://redie.uabc.mx/vol2no1/contents-backhoff.html .

Möltner, Andreas, Dieter Schellerg, und Jana Jünger. 2006. „Grundlegende quantitative Analysen medizinischer Prüfungen“. Journal Article. *GMS Zeitschrift für Medizinische Ausbildung* 23 (3): 11.