

Qualitätssicherung (digitaler) schriftlicher Prüfungen: Von Aufgabenmerkmalen bis zur diagnostischen Analyse

Dr. Marlit Annalena Lindner

IPN Kiel – Leibniz Institut für die Pädagogik der Naturwissenschaften und Mathematik

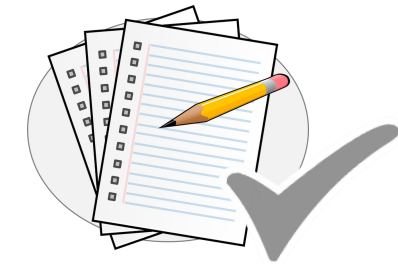
18.11.2021
Digitales Symposium der E-Assessment Alliance Berlin



Qualitätsmaßstab für (digitale) Hochschulprüfungen

Qualitätssicherung bei der Aufgabenkonstruktion

- ⇒ *Auswahl passender Aufgabenformate*
- ⇒ *Konstruktionsfehler kennen und vermeiden*
- ⇒ *Qualitätssicherung im Konstruktionsprozess*



Post Hoc Analyse der Aufgabenqualität



- ⇒ *Aufgabenkennwerte berechnen und interpretieren*
- ⇒ *Evidenzbasierte Reflektion der eigenen Prüfungspraxis*

Qualitätsmaßstab für Prüfungen



Gütekriterien für Prüfungen an Hochschulen



INHALTSVALIDITÄT

- Abdeckung der fachlichen Breite und Tiefe → fachspezifisch abzuleitende Kompetenzen
- Repräsentative Auswahl von Inhalten und Schwierigkeiten
- Sachliche Qualität der Aufgaben, Vermeidung von Konstruktionsfehlern

MESSGENAUIGKEIT

- Gleiche Resultate bei gleicher Kompetenz
- Gute Differenzierung von Leistungsniveaus

OBJEKTIVITÄT

- Transparente und objektive Bewertung

FAIRNESS

- Gleiche Bedingungen für alle Prüflinge

(Digitale) Prüfungspraxis in Pandemiezeiten



Gütekriterien für Prüfungen an Hochschulen

INHALTSVALIDITÄT

- Abdeckung der fachlichen Breite und Tiefe → fachspezifisch abzuleitende Kompetenzen
- Repräsentative Auswahl von Inhalten und Schwierigkeiten
- Sachliche Qualität der Aufgaben, Vermeidung von Konstruktionsfehlern

MESSGENAUIGKEIT

- Gleiche Resultate bei gleicher Kompetenz
- Gute Differenzierung von Leistungsniveaus

OBJEKTIVITÄT

- Transparente und objektive Bewertung


FAIRNESS

- **Gleiche Bedingungen für alle Prüflinge**
 - **Unberechtigte Vorteilnahme sowie unverschuldete Nachteile ausschließen**
 - **Authentizität der Prüfungsleistung sicherstellen**



Gütekriterien für Prüfungen an Hochschulen

INHALTSVALIDITÄT

- 
- Abdeckung der fachlichen Breite und Tiefe → fachspezifisch abzuleitende Kompetenzen
 - Repräsentative Auswahl von Inhalten und Schwierigkeiten
 - **Sachliche Qualität der Aufgaben, Vermeidung von Konstruktionsfehlern**


MESSGENAUIGKEIT

- Gleiche Resultate bei gleicher Kompetenz
- Gute Differenzierung von Leistungsniveaus

OBJEKTIVITÄT

- Transparente und objektive Bewertung

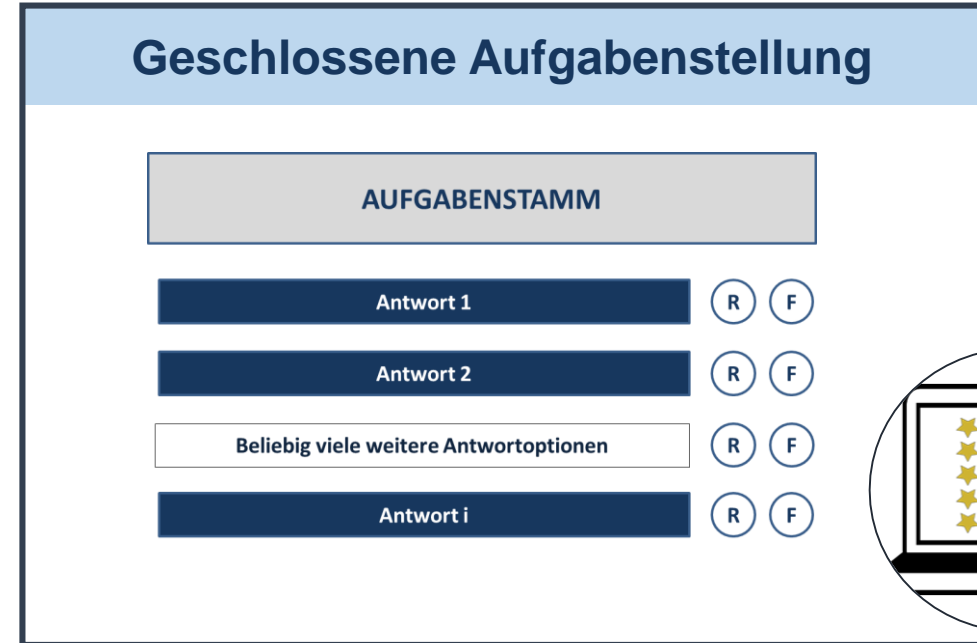
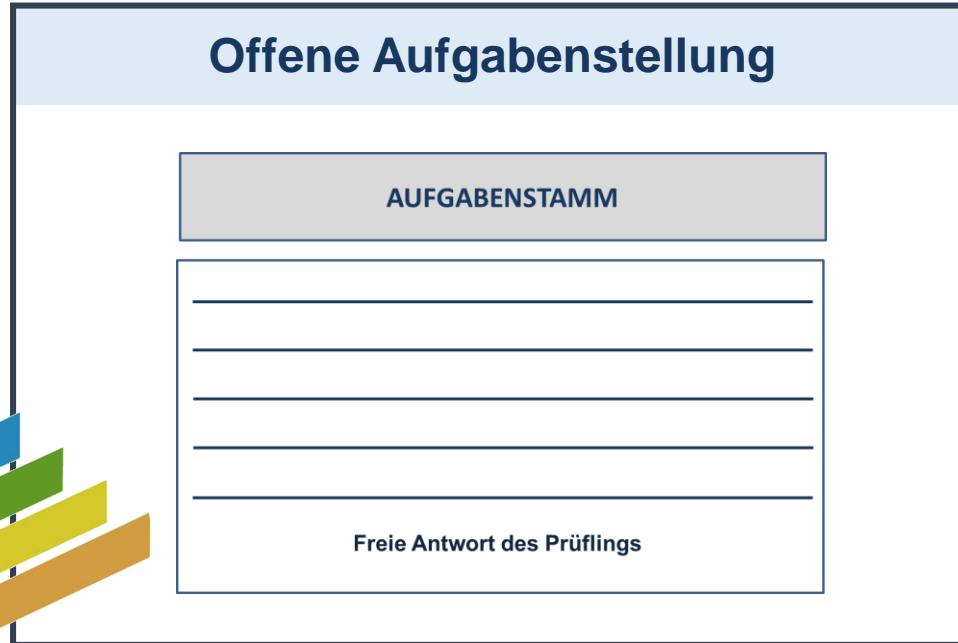
FAIRNESS

- 
- Gleiche Bedingungen für alle Prüflinge
 - Unberechtigte Vorteilnahme sowie unverschuldete Nachteile ausschließen
 - Authentizität der Prüfungsleistung sicherstellen

Aufgabenkonstruktion



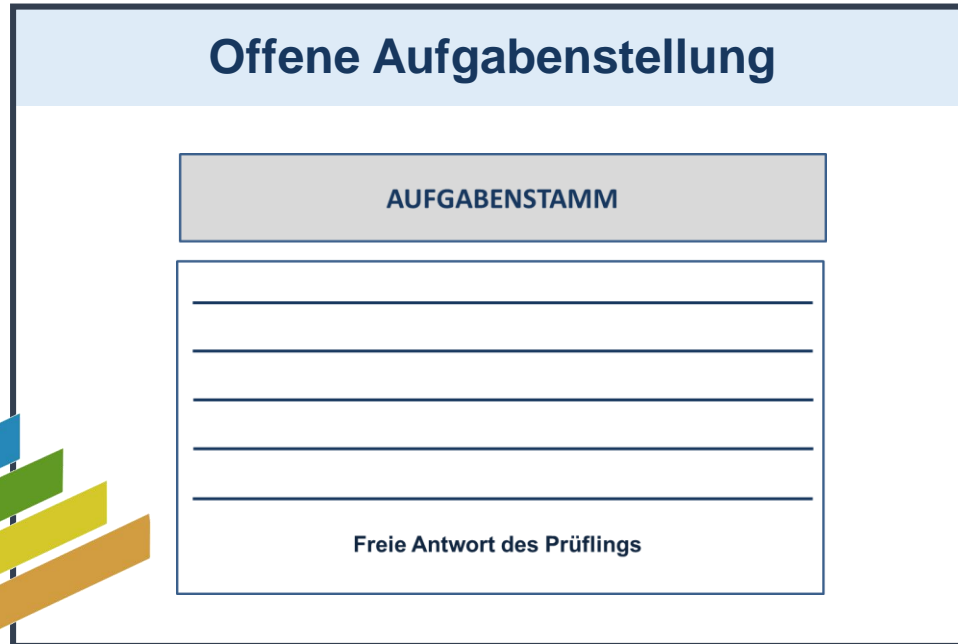
Offene und geschlossene Aufgabenformate



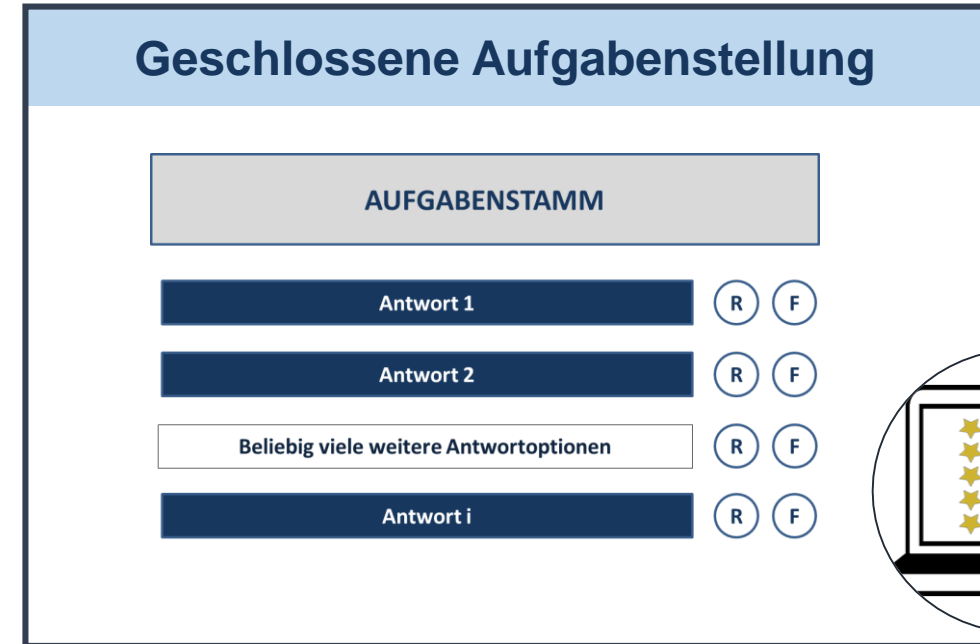
- Effiziente Konstruktion der Aufgaben
- Aufwendige Korrektur der Antworten
- Prüfung höherer Lernziele sowie von Transfer und Kreativität leicht möglich
- Viel Schreibarbeit, weniger Aufgaben
- Raten spielt kaum eine Rolle

- Anspruchsvolle Konstruktion
- Große Vielfalt verfügbarer Formate
- Automatisierte Auswertung möglich
- Prüfung höherer Lernziele schwieriger
- Mehr Aufgaben pro Zeiteinheit
- Raten muss berücksichtigt werden

Offene und geschlossene Aufgabenformate

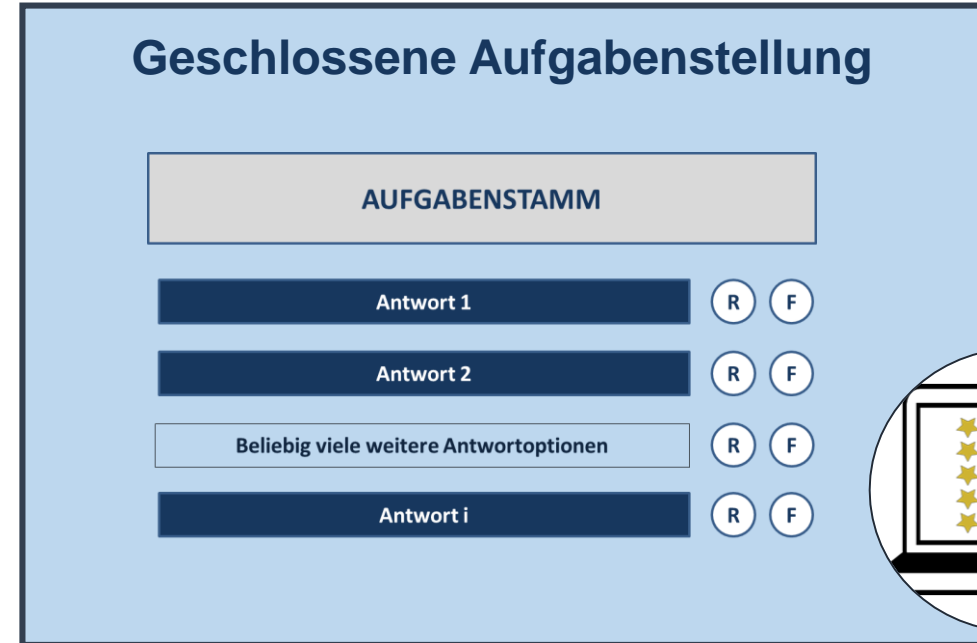
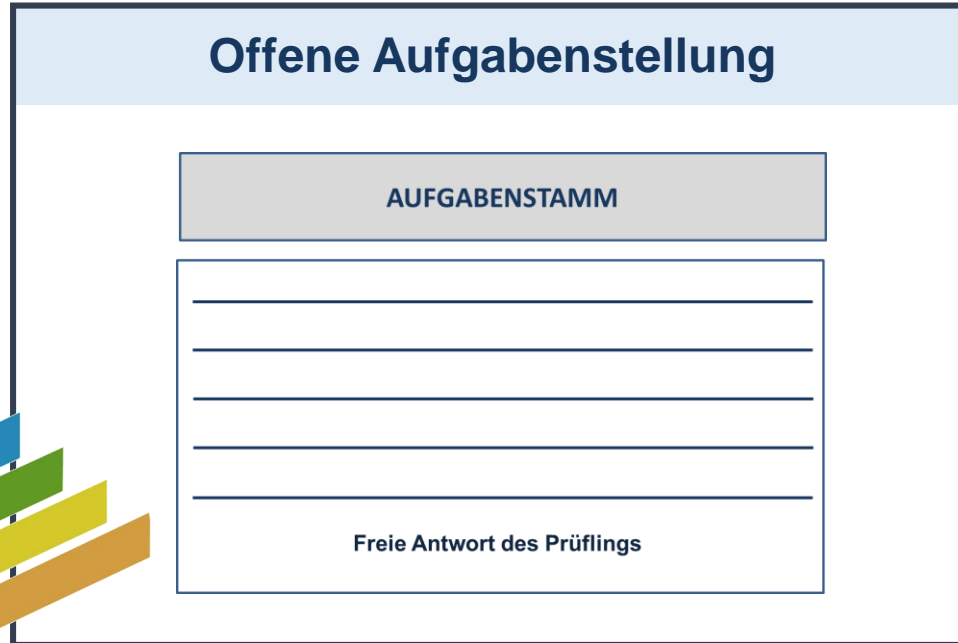


- Effiziente Konstruktion der Aufgaben
- **Aufwendige Korrektur der Antworten**
- Prüfung höherer Lernziele sowie von Transfer und Kreativität leicht möglich
- **Viel Schreibearbeit, weniger Aufgaben**
- Raten spielt kaum eine Rolle



- **Anspruchsvolle Konstruktion**
- Große Vielfalt verfügbarer Formate
- Automatisierte Auswertung möglich
- **Prüfung höherer Lernziele schwieriger**
- Mehr Aufgaben pro Zeiteinheit
- **Raten muss berücksichtigt werden**

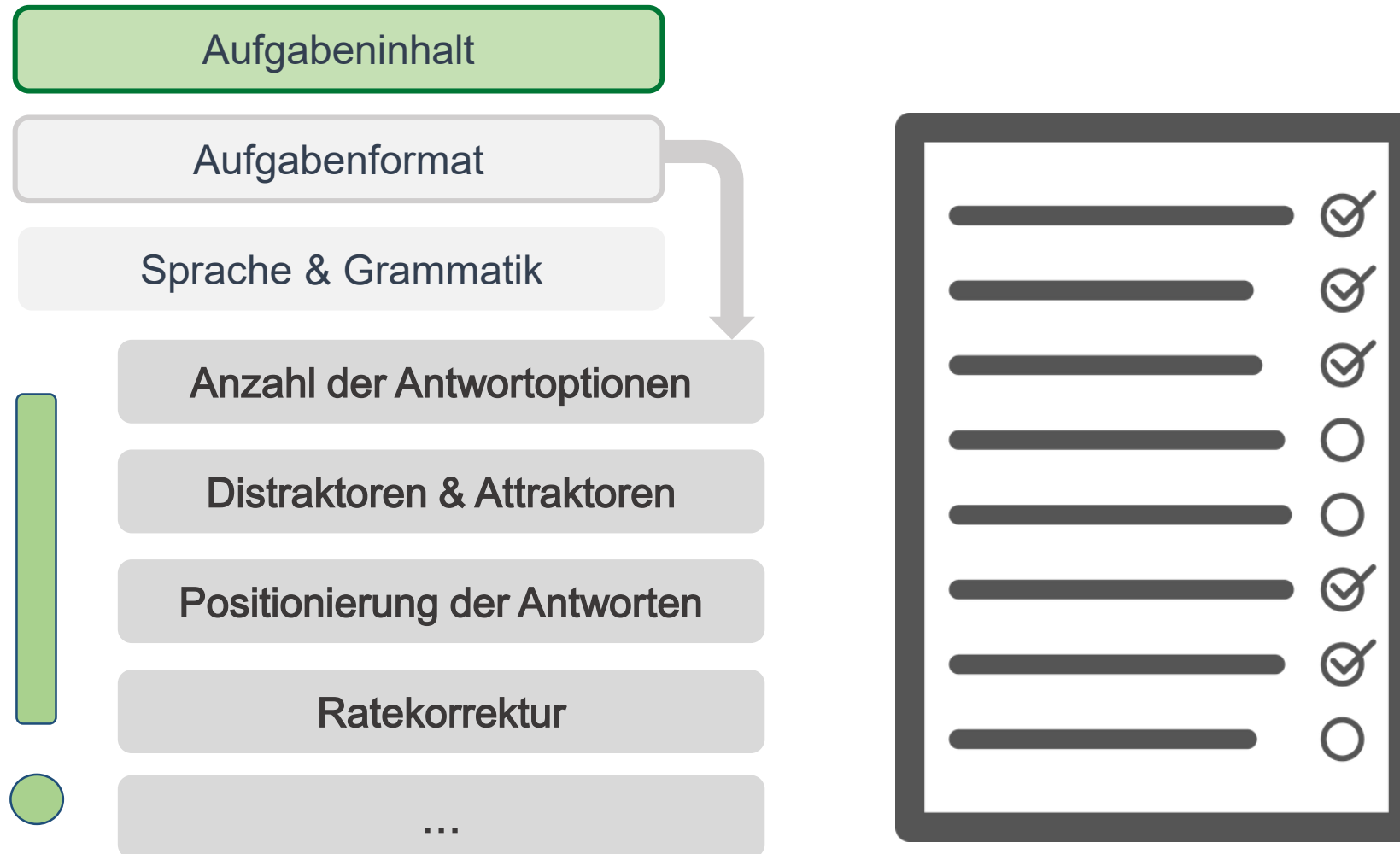
Offene und geschlossene Aufgabenformate



- Effiziente Konstruktion der Aufgaben
- Aufwendige Korrektur der Antworten
- Prüfung höherer Lernziele sowie von Transfer und Kreativität leicht möglich
- Viel Schreiarbeit, weniger Aufgaben
- Raten spielt kaum eine Rolle

- **Anspruchsvolle Konstruktion!!**
- Große Vielfalt verfügbarer Formate
- Automatisierte Auswertung möglich
- Prüfung höherer Lernziele schwieriger
- Mehr Aufgaben pro Zeiteinheit
- Raten muss berücksichtigt werden

Konstruktion von (geschlossenen) Prüfungsaufgaben



A 3D rendered chessboard with a black and white checkered pattern. The chessboard is shown from a low angle, with the pieces appearing to be on a raised platform. The pieces are white and include a king, queen, rook, knight, bishop, and pawns. The text "TESTWISENESS" is overlaid in the center of the board in a large, white, bold, sans-serif font with a black outline. The lighting is dramatic, with strong shadows and highlights on the pieces and the board.

TESTWISENESS

Testwiseness-Strategien bedenken

Testwiseness

„Fähigkeit, Eigenschaften geschlossener Testformate strategisch auszunutzen, um ein besseres Testergebnis zu erzielen.“

(z.B. Dodeen, 2008; Millman, Bishop & Ebel, 1965)

- ⇒ Testwiseness-Strategien können die Prüfungsleistung unabhängig vom fachlichen Wissen systematisch verzerren *(Cohen, 2006; Haladyna & Downing, 2004)*
- ⇒ **Voraussetzung:** Testkonstruktionsfehler in geschlossenen Aufgaben z.B. schlechte Distraktoren, logische Abhängigkeiten der Antworten, grammatikalische Fehler *(vgl. z.B. Haladyna & Downing, 2002)*

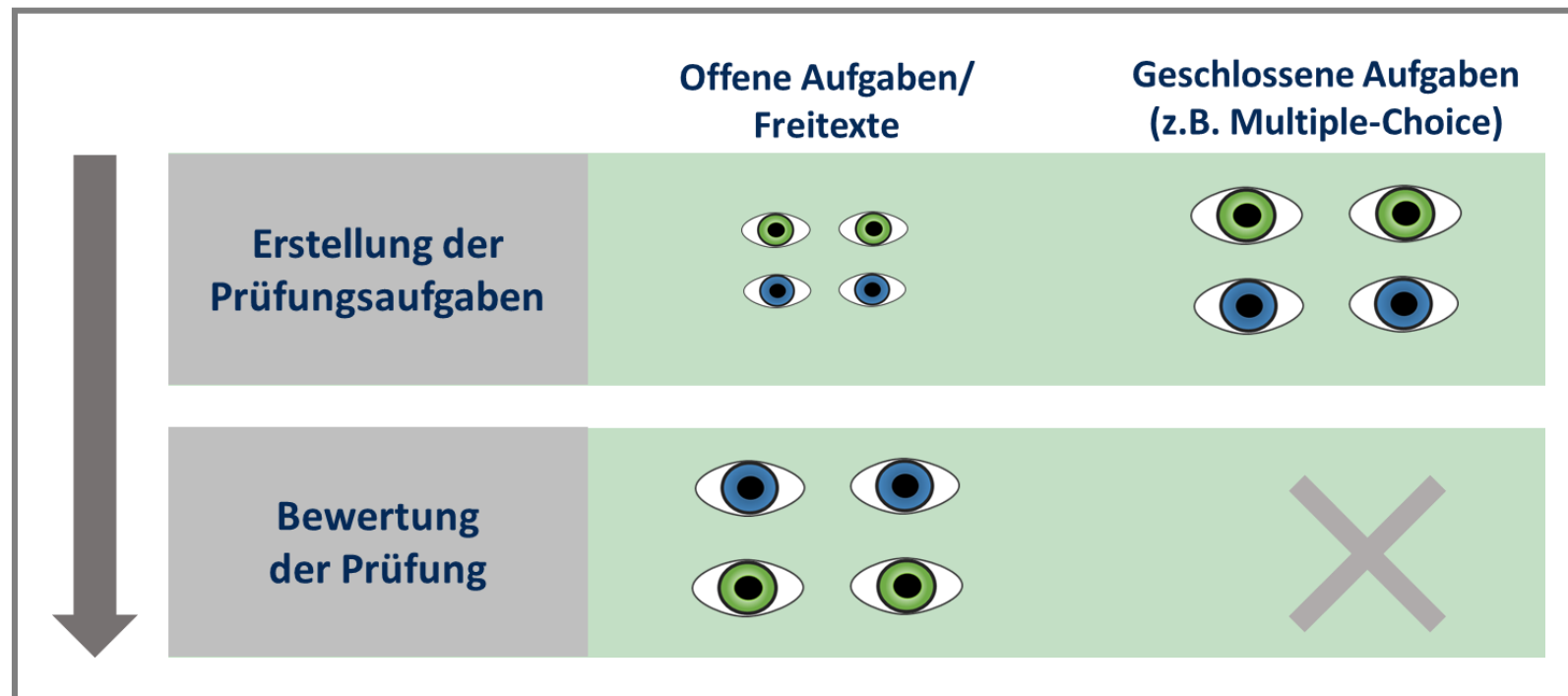
Beispiele für typische Strategien der Antwortwahl...

- Antworten in mittlerer Position oder längste Antwort
- Antworten, die Aspekte des Aufgabenstamms wiederholen
- Selektion von Antworten nach Indikatorwörtern (z.B. nie/ immer)
- Ausschluss von Antworten die grammatikalisch unpassend sind
- Ausschluss anhand von logischen Abhängigkeiten

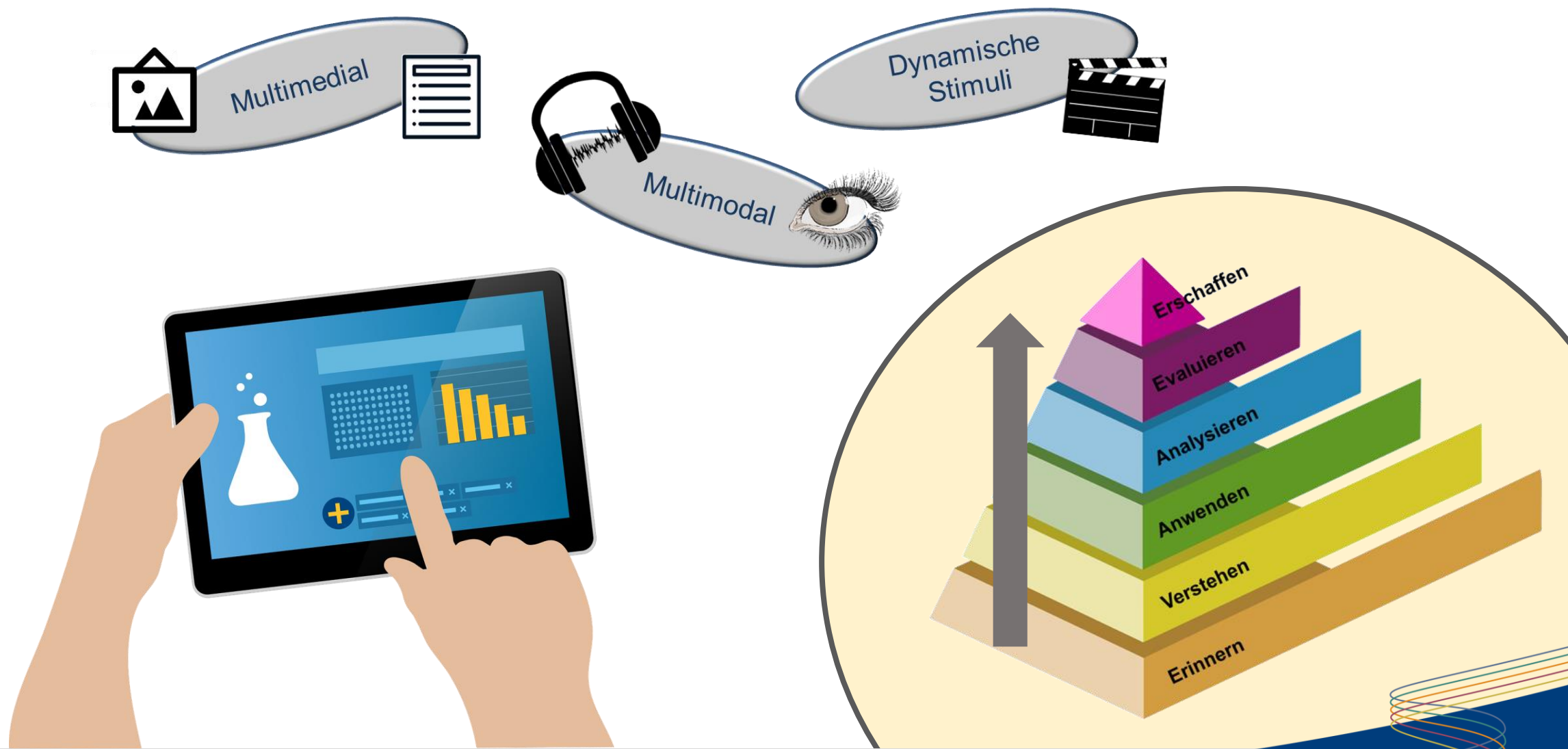
Qualitätssicherung durch Vier-Augen-Prinzip

Validierung durch Zweitprüfende

- Sicherstellung der Aufgabenqualität im Konstruktionsprozess (vor der Prüfung)
- Sicherstellung einer fairen und objektiven Bewertung (nach der Prüfung)



Potenziale digitaler Aufgabenstellungen



Analyse der Aufgabenqualität



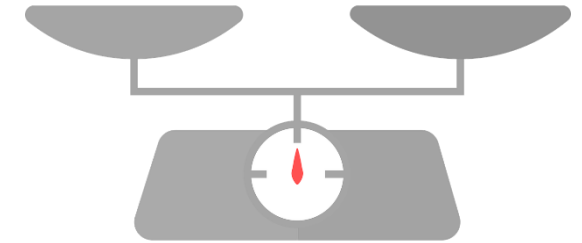


- Antworten von Studierenden liefern empirische Daten
- in computerbasierten Prüfungen liegen die Daten (im besten Fall) unmittelbar digital vor
 - Daten müssen dennoch für die Analyse vorbereitet werden → erfordert Aufwand und Know-How von Prüfenden (Automatisierung wünschenswert)
- Basierend auf den Daten können verschiedene Parameter zur besseren Beurteilung der Aufgabenqualität berechnet werden (Aufgabendiagnostik)
- Berechnete Parameter stellen z.B. auffällige Aufgaben heraus und können helfen Probleme in der Aufgabenkonstruktion und Optimierungsmöglichkeiten abzuleiten
- Analyse gleicher Aufgaben über verschiedene Kohorten ist besonders aufschlussreich

Aufgabenschwierigkeit

Frage

Wie schwierig war die Aufgabe für die aktuelle Studienkohorte?



Berechnung

- Anteil der Personen, die eine Aufgabe richtig beantwortet haben
 - ✓ Untersuchung getrennt nach Leistungsgruppen liefert zusätzliche Informationen
 - Die Schwierigkeit einer Aufgabe sollte in höheren Leistungsgruppen [bemessen am Gesamtergebnis] abnehmen, d.h. Personen mit guter Leistung sollten die Aufgabe häufiger korrekt lösen

Interpretation

- Klassifikation von Aufgaben, z.B. als schwer, mittel und leicht kann bei der (zukünftigen) Zusammenstellung von Prüfungen helfen
- Identifikation von zu schweren Aufgaben (Konstruktionsfehler?) und zu leichten Aufgaben (z.B. Aufgabe bekannt geworden)

Aufgabentrennschärfe (Diskrimination)



Frage

Wie gut differenziert die Aufgabe zwischen Prüflingen mit höherem und niedrigerem Leistungsniveau? Wie nützlich ist die Aufgabe zur Leistungsdifferenzierung?

Berechnung

- Korrelation (linearer Zusammenhang) zwischen der Antwortrichtigkeit in der untersuchten Aufgabe und dem Gesamtergebnis der Prüfung
- Korrelationswerte liegen zwischen -1 und 1
 - ✓ positive Korrelation [$>.30$] spricht für eine gute Differenzierung der Aufgabe zwischen Personen mit mehr und weniger Wissen

Interpretation

- Repräsentativität der Aufgabe für die Prüfungsleistung (inwiefern spiegelt sich die Gesamtleistung der Studierenden in der untersuchten Aufgabe wider)
- Potenzial der Aufgabe zwischen Leistungsniveaus zu differenzieren

Distraktoranalysen (nur geschlossene Aufgaben)

Frage

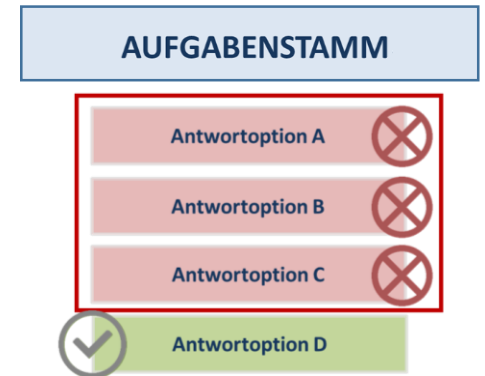
Wie gut sind meine Distraktoren (falsche Antworten) in geschlossenen Aufgaben konstruiert? Tragen sie zur Leistungsdifferenzierung bei?

Berechnung(en)

- Distraktortrennschärfe: Korrelation der Distraktorwahl mit dem Gesamtscore im Test (sollte im negativen Bereich liegen)
- Wahlhäufigkeit: Wie oft wurde ein Distraktor gewählt (z.B. in Prozent)
 - ✓ Untersuchung in verschiedenen Leistungsgruppen liefert zusätzliche Information (siehe Beispiel)

Interpretation

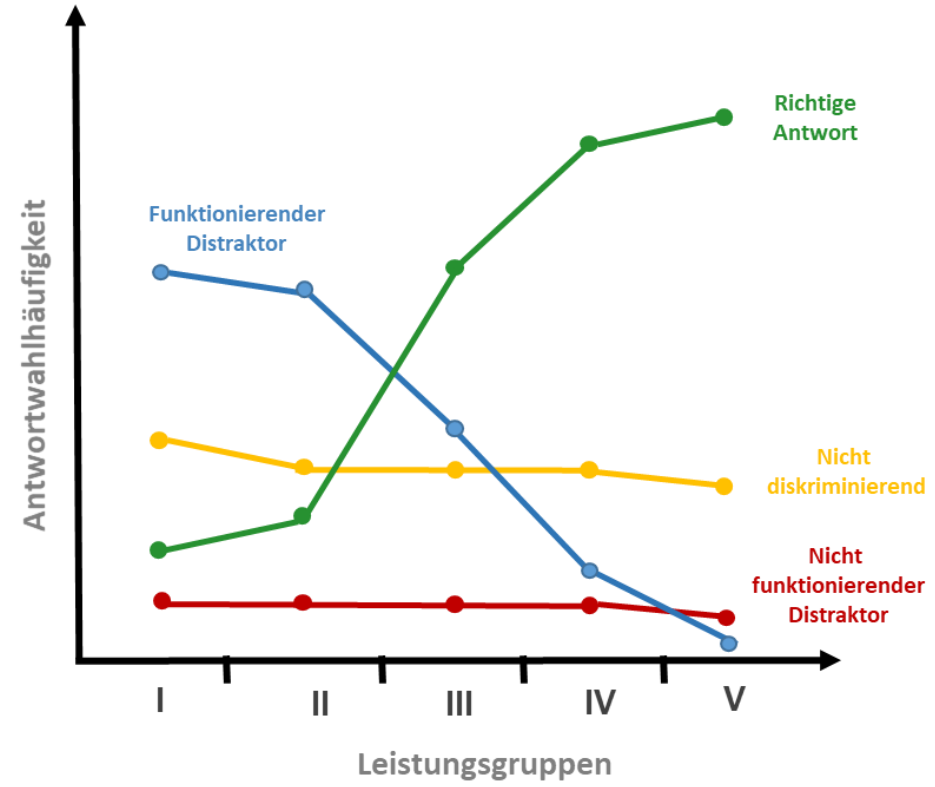
- Attraktivität und Ablenkungspotenzial einzelner Distraktoren in geschlossenen Aufgaben
- Welche Distraktoren sind besonders geeignet fachlich kompetente Studierende von weniger kompetenten Studierenden zu trennen (Distraktortrennschärfe)



Beispiel für eine Distraktoranalyse

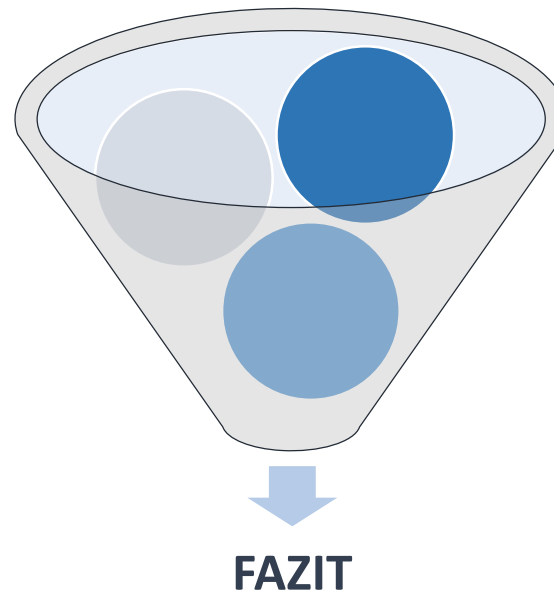
(Fiktive) Beispielanalyse für eine Single-Choice-Aufgabe

AUFGABENSTAMM	
Antwort 1	F
Antwort 2	R
Antwort 3	F
Antwort 4	F



vgl. z.B. Forthmann, B., Förster, N., Schütze, B., Hebbecker, K., Flessner, J., Peters, M. T., & Souvignier, E. (2020). How much g is in the distractor? Re-thinking item-analysis of multiple-choice items. *Journal of Intelligence*. <https://www.mdpi.com/2079-3200/8/1/11>

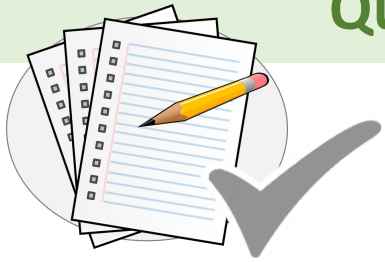
Zusammenfassung und Fazit



Fazit: Gute Prüfungspraxis



Qualitätssicherung bei der Aufgabenkonstruktion



Offene und geschlossene Formate sinnvoll kombinieren

Testwiseness – Konstruktionsfehler vermeiden

Vier-Augen-Prinzip (Prüfungserstellung und Bewertung)

Potenzielle digitaler Aufgabengestaltung nutzen

Post Hoc Analyse der Aufgabenqualität



Aufgabenkennwerte berechnen und interpretieren

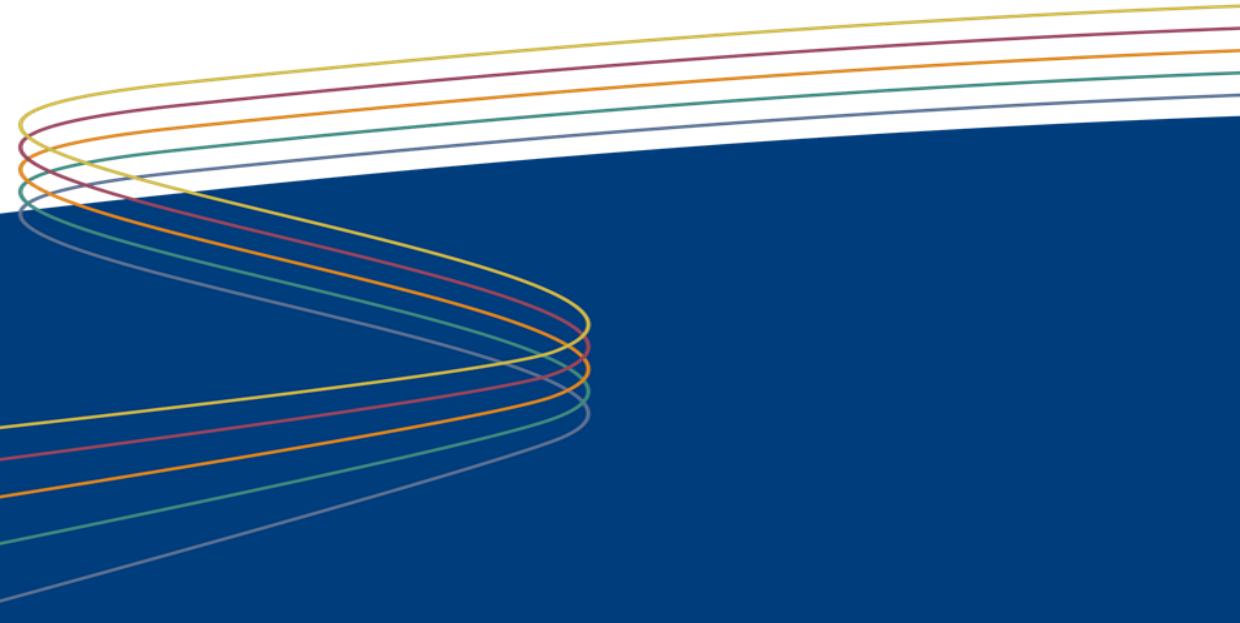
Kritische Reflektion der eigenen Aufgaben

Gezielte Adaptation der Prüfung mittels empirischer Hinweise

Vielen Dank!



mlindner@leibniz-ipn.de



Übersichtsarbeiten zum Weiterlesen



Ein Plädoyer zur Qualitätssicherung schriftlicher Prüfungen im Psychologiestudium

Marlit A. Lindner¹, Jörn R. Sparfeldt², Olaf Köller¹, Josef Lukas³ und Detlev Leutner⁴

Psychologische Rundschau, 72, 93–105
<https://doi.org/10.1026/0033-3042/a000524>

2021

Multiple-Choice-Prüfungen an Hochschulen?

Ein Literaturüberblick und Plädoyer für mehr praxisorientierte Forschung

Marlit A. Lindner, Benjamin Strobel und Olaf Köller

Zeitschrift für Pädagogische Psychologie, 29, 133–149

<https://doi.org/10.1024/1010-0652/a000156>

2015