Sebastian Rampp

ISSN 2568-0331

Theoretische Struktur vs. empirische Realität
–
Eine empirische Untersuchung des BRT 2 zur dimensionalen Architektur und seiner Eignung als Instrument zur Diagnostik von Rechenschwäche in der 2. Jahrgangsstufe

Mathematikdidaktik im Kontext

Heft 13

Inhaltsverzeichnis

1 Einleitung und Zusammenfassung 6
1.1 Begriffsbestimmung von Rechenschwäche 6
1.2 Zielsetzung und forschungsleitende Fragen 7
2 Theoretischer Hintergrund 9
2.1 Mathematische Kompetenzentwicklung in der 2. Jahrgangsstufe und das Drei-Säulen-Modell 9
2.2 Diagnostik im Grundschulalter und Vorstellung des Testverfahrens 10
2.3 Methodologische und Testtheoretische Grundlagen 11
2.3.1 Verteilungscharakteristika und nicht-parametrische Hypothesenprüfung 11
2.3.2 Prüfung auf Varianzhomogenität 12
2.3.3 Mathematische Modellierung manifester und latenter Zusammenhänge 13
2.3.4 Axiomatik der Klassischen Testtheorie (KTT) und Itemmetrik 13
2.3.5 Konzepte der klassischen und modellbasierten Reliabilitätsbeurteilung 14
2.3.6 Graphentheoretische Modellierung von Fehlermustern (Item-Netzwerke) 14
2.3.7 Personenzentrierte Typenbildung: Clusteranalyse und Validierung 15
2.3.8 Fundamente der multidimensionalen Eigenschafts- und Faktorenanalyse (EFA/CFA) 16
2.3.9 Einführung in die Item-Response-Theorie (IRT) 17
2.3.10 Testfairness und Invarianzprüfung: DIF-Analyse 18
2.3.11 Hierarchische Ladungsarchitekturen: Das Bifaktor-Modell 20
3 Methodik 22
3.1 Stichprobenbeschreibung 22
3.2 Das Erhebungsinstrument: Struktur der 18 Items (A1-A18) und Bepunktungssystem 23
3.3 Statistische Datenanalyse und Voraussetzungsprüfung 24
3.3.1 Überprüfung der Normalverteilungsannahme 26
3.3.2 Nicht-parametrische Verfahren zur Ergebnisprüfung und Ergebnisgrafik 26
3.3.3 Zusammenhangsanalyse 27
3.3.4 Geschlechtsspezifische Analysen 27
3.3.5 Prüfung der Varianzhomogenität 28
3.3.6 Extremgruppenvergleich 29
3.3.7 Methoden zur Bestimmung der Reliabilität und internen Konsistenz 29

3.3.8 Analyse der Aufgabenschwierigkeit 30
3.3.9 Methoden zur Analyse bedingter Fehlerwahrscheinlichkeiten und Item-Netzwerke 30
3.3.10 Methodisches Vorgehen bei der Durchführung der Clusteranalyse 31
3.3.11 Prozedurales Vorgehen zur empirischen Strukturaufklärung 32
3.3.12 Prozedurales Vorgehen zur probabilistischen Testanalyse (IRT) 33
3.3.13 Prozedurales Vorgehen zur Analyse ungleicher Aufgabenfunktionen 35
3.3.14 Prozedurales Vorgehen zur Analyse der latenten Personenparameter 35
3.3.15 Prozedurales Vorgehen zur hierarchischen Bifaktor-Modellierung 36
4 Ergebnisse 39
4.1 Deskriptive und inferenzstatistische Leistungsanalysen der Stichprobe 40
4.1.1 Leistungsunterschiede zwischen den mathematischen Teilbereichen 41
4.1.2 Interkorrelation der mathematischen Teilbereiche 44
4.1.3 Geschlechtsspezifische Leistungsunterschiede 45
4.1.4 Analyse der Leistungsstreuung und Heterogenität 48
4.1.5 Extremgruppenvergleich – Oberes vs. unteres Leistungsviertel 51
4.2 Psychometrische Skalen- und Itemanalyse des BRT-2 53
4.2.1 Interne Konsistenz und Reliabilität der Kompetenzbereiche 53
4.2.2 Itemmetrische Kennwerte: Aufgabenschwierigkeiten 54
4.2.3 Bedingte Fehlerwahrscheinlichkeiten und Fehlernetze 56
4.3 Personenzentrierte und strukturprüfende Modellanalysen 58
4.3.1 Identifikation typischer Leistungsprofile 59
4.3.2 Strukturaufklärung und faktorielle Validität des BRT-2 64
4.3.3 Probabilistische Analysen und diagnostische Eignung 67
4.3.4 Analyse ungleicher Aufgabenfunktionen (Differential Item Functioning) 72
4.3.5 Verteilungsanalysen und Gruppenvergleich der latente Personenparameter 74
4.3.6 Untersuchung der dimensionalen Architektur mittels Bifaktor-Modellierung 77
5 Diskussion 84
5.1 Integration und Interpretation der zentralen Befunde 84
5.2 Methodische Limitationen 88
5.3 Implikationen für die Weiterentwicklung des Instruments 89
5.3.1 Itemmetrische Optimierung und Schwellenbereinigung 89

5.3.2 Konstruktionstechnische Kompensation von Geschlechtereffekten (DIF) 89
5.3.3 Transformation in ein Computergestütztes Adaptives Testverfahren (CAT) ... 90
5.3.4 Schnittstelle zur qualitativen Vertiefung: Kopplung mit der Bayreuther Förderdiagnostik 91
6 Fazit..... 92
Literaturverzeichnis 94
Anhang 101
Eidesstattliche Erklärung 114

1 Einleitung und Zusammenfassung

Die Entwicklung solider mathematischer Basiskonzepte in der Primarstufe gilt als ein entscheidender Meilenstein für den schulischen Erfolg von Kindern (vgl. Duncan et al., 2007). Bildungsstandards und Lehrpläne unseres Schulsystems sehen daher explizit vor, dass Schülerinnen und Schüler in der Grundschule tragfähige Vorstellungen zu den natürlichen Zahlen und den vier Grundrechenarten entwickeln. In der schulischen Praxis lässt sich dieser Anspruch jedoch nicht immer vollumfänglich umsetzen: „Ein nennenswerter Anteil an Schülerinnen und Schülern verlässt die Grundschule, ohne ein tragfähiges Verständnis für natürliche Zahlen, für das dezimale Stellenwertsystem und für Rechenoperationen aufgebaut zu haben." (Steinecke & Richter, 2025, S. 2).

Das Risiko dieser Defizite liegt in ihrer potenziellen Unsichtbarkeit im frühen Arithmetikunterricht: „Rechenschwache Schülerinnen und Schüler können die Grundschule durch das Auswendiglernen von Ergebnissen, ein gutes Gedächtnis und fleißiges Üben durchaus mit einer Note 3 im Fach Mathematik abschließen" (Steinecke & Richter, 2025, S. 2). Die Quittung folgt spätestens im Mathematikunterricht der weiterführenden Schulen. Hier stehen diese Kinder vor massiven Barrieren. Wenn der neue kumulative Lernstoff, von Brüchen und negativen Zahlen bis hin zu Variablen, hinzukommt, fehlt das Fundament. Mangels Basiswissen lässt sich das Ganze allenfalls als mechanisches, sinnbefreites Regelwerk auswendig lernen (Steinecke & Ulm, 2025). Eine unentdeckte Rechenschwäche führt in den weiterführenden Schulen somit unweigerlich zu chronischer Überforderung und Lernblockaden (Steinecke & Ulm, 2025).

Damit Lehrkräfte individuelle Lernbarrieren sensibel wahrnehmen und betroffene Lernende rechtzeitig unterstützen können, bedarf es verlässlicher pädagogischer Diagnoseverfahren. „Genau zu diesem Zweck wurde das *Bayreuther Testpaket zur Erfassung von Rechenschwäche im Mathematikunterricht der Jahrgangsstufe 2* entwickelt" (Steinecke & Richter, 2025, S. 2), das den Bayreuther Rechentest für die Jahrgangsstufe 2 enthält. Während die Handreichung selbst primär für die schulpraktische, quantitative Identifikation und anschließende Förderdiagnostik konzipiert wurde, widmet sich die vorliegende Masterarbeit einer umfassenden quantitativen und theoretischen Evaluation des Instruments. Anhand eines großen empirischen Datensatzes von insgesamt $N = 574$ Lernenden wird das Verfahren psychometrisch auf Herz und Nieren geprüft, um die theoretischen Ansprüche der Testkonstruktion empirisch abzusichern.

1.1 Begriffsbestimmung von Rechenschwäche

Die wissenschaftliche Einordnung und Benennung von besonderen Schwierigkeiten beim Lernen von Mathematik variiert je nach Fachdisziplin. In medizinisch-psychologischen Zugängen wird traditionell mit Begriffen wie Dyskalkulie oder Rechenstörung gearbeitet – insbesondere auf Basis der Definition des internationalen Klassifikationssystems ICD-11 der Weltgesundheitsorganisation:

> „Eine Lernentwicklungsstörung mit Beeinträchtigung in Mathematik ist gekennzeichnet durch bedeutsame und anhaltende Schwierigkeiten beim Erlernen akademischer Fähigkeiten im Zusammenhang mit Mathematik oder Arithmetik, wie z. B. Zahlensinn, Auswendiglernen von Zahlenfakten, genaues Rechnen, flüssiges Rechnen und genaues mathematisches Denken. Die Leistungen der Person in Mathematik oder Arithmetik liegen deutlich unter dem, was für das chronologische oder Entwicklungsalter und intellektuelle Funktionsniveau zu erwarten wäre, und führen zu einer bedeutsamen Beeinträchtigung der akademischen oder beruflichen Leistungsfähigkeit der Person. Die Lernentwicklungsstörung mit Beeinträchtigung in Mathematik ist nicht auf eine Störung der intellektuellen Entwicklung, eine sensorische Beeinträchtigung (Seh- oder Hörvermögen), eine neurologische Störung, mangelnde Verfügbarkeit von Bildung, mangelnde Beherrschung der Sprache des akademischen Unterrichts oder psychosoziale Widrigkeiten zurückzuführen.“ (Weltgesundheitsorganisation [WHO], 2019)

Hierbei handelt es sich um ein sogenanntes Diskrepanzkriterium, da nur die Lernenden mit einem durchschnittlichen oder überdurchschnittlichen Intelligenzquotienten eine Rechenstörung haben können (Urhahne et al., 2019).

Dahingegen wird in der Mathematikdidaktik und in pädagogischen Kontexten der Begriff der Rechenschwäche verwendet. Hierbei wird ausdrücklich betont, dass hier kein unheilbares Krankheitsbild vorliegt, sondern es sich vielmehr um ein Verständnisdefizit handelt, das durch gezieltes, inhaltsbasiertes Training minimiert oder vollständig beseitigt werden kann (Steinecke & Ulm, 2025).

Diese pädagogische Sichtweise prägt auch die Arbeiten von Gaidoschik et al. (2021). Er definiert Rechenschwäche im Kern als das folgenschwere Ausbleiben des Übergangs vom zählenden Rechnen zu denkenden, strukturierten Lösungsstrategien. Der Lernende verharrt beim mühsamen, fehleranfälligen Abzählen einzelner Einheiten, oft unter heimlicher Zuhilfenahme der Finger oder anderer Hilfsmittel. Die Ursache liegt laut Gaidoschik et al. (2021) nicht im Gehirn des Kindes verankert, sondern ist das Resultat eines unvollständigen Lernprozesses. Den betroffenen Lernenden fehlt schlicht das tragfähige Fundament der Vorstellung von natürlichen Zahlen des dezimalen Stellenwertsystems und ein tiefes, operatives Verständnis für die Bedeutung der Rechenoperationen. Sie haben sich in einer sackgassenartigen Überlebensstrategie eingerichtet. Genau an dieser Stelle setzt das kriterienorientierte Screening des BRT-2 an, um diese Verhaltensmuster im Klassenverbund frühzeitig aufzudecken.

1.2 Zielsetzung und forschungsleitende Fragen

Da die für diese Arbeit verwendete Definition von Rechenschwäche von einer Aufteilung in drei Kompetenzbereiche ausgeht und auch das Bayreuther Testpaket anhand dieser Definition entwickelt wurde, sollen in dieser Arbeit die folgenden zwei Forschungsfragen untersucht werden:

1. Spiegelt sich die theoretische Struktur in den empirischen Daten wider?
2. Eignet sich der Bayreuther Rechentest für die Jahrgangsstufe zwei als Diagnoseinstrument für Rechenschwäche?

Um diese Frage zu beantworten und darüber hinaus noch die Eignung des Testinstruments zu prüfen, wurden 13 forschungsleitende Fragen formuliert, die in dieser Arbeit geklärt werden sollen, um zu einem finalen Ergebnis über die empirische Struktur zu gelangen. Diese 13 forschungsleitenden Fragen lauten:

1. Schneiden die Lernenden in den verschiedenen Kompetenzbereichen unterschiedlich ab?
2. Wenn ein Lernender in einem der Kompetenzbereiche gute Leistungen erzielt, tut er dies dann in den anderen auch?
3. Gibt es bei den Gesamtleistungen bzw. bei den Leistungen in den Kompetenzbereichen Geschlechterunterschiede?
4. Ist die Leistungsstreuung über alle Kompetenzbereiche hinweg konstant?
5. Wie unterscheiden sich die besten von den schlechtesten 25 % der Lernenden?
6. Sind die Aufgaben der Kompetenzbereiche konsistent?
7. Existieren besonders leichte oder schwere Aufgaben?
8. Existieren typische Leistungsprofile der Lernenden?
9. Inwiefern spiegeln die Zusammenhänge zwischen den einzelnen Aufgaben die theoretisch angenommenen mathematischen Kompetenzbereiche wider?
10. Wenn eine Aufgabe falsch beantwortet wurde, mit welcher Wahrscheinlichkeit wurde dann eine andere Aufgabe falsch beantwortet?
11. Wie gut eignen sich die Aufgaben zur Diagnose von Rechenschwäche und gibt es besonders gute oder schlechte Aufgaben?
12. Existieren Aufgaben, bei denen die Geschlechter trotz gleicher Fähigkeit unterschiedliche Leistungen erbringen?
13. Existiert ein Generalfaktor oder beschreibt ein dreidimensionales Modell das Verhalten am besten?

2 Theoretischer Hintergrund

Das vorliegende Kapitel stellt das theoretische und methodische Fundament für die anschließende empirische Überprüfung des Testverfahrens dar. Zur Verknüpfung der fachdidaktischen Intention des Messinstruments mit dessen statistischer Validierung gliedert sich der theoretische Rahmen in drei Kernbereiche. Zunächst werden die mathematikdidaktischen Grundlagen der Kompetenzentwicklung im zweiten Schuljahr sowie das dem Test zugrundeliegende Drei-Säulen-Modell skizziert (Kapitel 2.1). Darauf aufbauend folgt eine diagnostische Verortung des untersuchten Screening-Verfahrens (Kapitel 2.2). Den Schwerpunkt dieses Kapitels bildet schließlich die Darlegung der testtheoretischen und methodologischen Grundlagen (Kapitel 2.3). Diese definieren die theoretische Basis für die fortgeschrittenen statistischen Analysen dieser Arbeit, welche von der klassischen Itemanalyse über strukturprüfende Faktorenanalysen bis hin zur probabilistischen Modellierung reichen.

2.1 Mathematische Kompetenzentwicklung in der 2. Jahrgangsstufe und das Drei-Säulen-Modell

Die mathematische Kompetenzentwicklung im zweiten Schuljahr markiert eine kritische Phase beim Übergang von zählenden Rechenprozeduren zu abstrakten, flexiblen Rechenstrategien sowie beim Aufbau eines tieferen Verständnisses für Zahlbeziehungen und arithmetische Strukturen (für einen umfassenden Überblick zur arithmetischen Entwicklung in der Primarstufe vgl. Padberg, 2015). Um den Anforderungen des Mathematikunterrichts dauerhaft gewachsen zu sein und das Entstehen einer Rechenschwäche (im Sinne der Ausführung in Kapitel 1.1) zu vermeiden, müssen Kinder ein tiefgehendes Verständnis in drei zentralen, miteinander vernetzten Inhaltsbereichen entwickeln. Das Bayreuther Testpaket operationalisiert das mathematische Konstrukt daher über eine theoretische Dreiteilung (Steinecke & Richter, 2025):

- Verständnis für natürliche Zahlen (VNZ): Hierbei werden die Zahlauffassungen, der Kardinal- und Ordinalzahlaspekt sowie eine flexible Orientierung im erweiterten Zahlenraum fokussiert.
- Verständnis für das Stellenwertsystem (VSWS): In diesem Inhaltsbereich steht die Einsicht in Bündelungsprinzipien sowie die Struktur des dekadischen Systems im Mittelpunkt
- Verständnis für Rechenoperationen (VRO): Hier wird das relationale Verständnis von Grundoperationen (Addition und Subtraktion) sowie das Erkennen von Aufgabenbeziehungen geprüft.

Diese drei Säulen bilden das fachdidaktische Fundament, auf dessen Basis die strukturellen und probabilistischen Analysen der vorliegenden Arbeit durchgeführt werden.

2.2 Diagnostik im Grundschulalter und Vorstellung des Testverfahrens

Um betroffenen Kindern frühzeitig evidenzbasierte Unterstützung zukommen zu lassen, ist eine präzise Erfassung des individuellen Förderbedarfs unabdingbar. In der Schulpraxis besteht hierbei ein hoher Bedarf an diagnostischen Verfahren, die von Lehrkräften ökonomisch, reliabel und ohne massiven organisatorischen Aufwand direkt im regulären Schulalltag als Screening-Instrument im Klassenverband durchgeführt werden können (vgl. Ingenkamp & Lissmann, 2008; Steinecke & Richter, 2025).

Zu diesem Zweck wurde am Lehrstuhl für Didaktik der Mathematik der Universität Bayreuth das *Bayreuther Testpaket zur Erfassung von Rechenschwäche im Mathematikunterricht* entwickelt. Der in der vorliegenden Arbeit untersuchte *Bayreuther Rechentest für die Jahrgangsstufe 2 (BRT 2)* ist als standardisiertes Screening-Verfahren für den Einsatz im Klassenverband konzipiert. Ziel des Instruments ist es nicht, eine zeitaufwendige klinische Einzelfalldiagnose zu ersetzen. Vielmehr soll es Lehrkräften als ökonomisches Werkzeug dienen, um mathematische Risikoprofile bei Lernenden frühzeitig zu identifizieren und spezifische qualitative Defizite aufzudecken (Steinecke & Richter, 2025).

Damit grenzt sich der BRT-2 von klassischen Schultests ab. Die auf dem Markt dominierenden Verfahren sind fast ausschließlich normorientiert (Moosbrugger & Kelava, 2020). Sie vergleichen die individuelle Leistung eines Kindes mit einer repräsentativen Eichstichprobe, um die relative Position des Einzelnen innerhalb der Bezugsgruppe zu bestimmen (Moosbrugger & Kelava, 2020). Am Ende steht ein nackter Prozentrang. Der BRT-2 bricht mit dieser Tradition durch seinen kriterienorientierten Ansatz (Steinecke & Richter, 2025). Er fragt nicht, wie das Kind im Durchschnitt der Masse abschneidet, sondern ganz direkt, ob fundamentale mathematische Basiskonzepte inhaltlich verstanden wurden.

Ein gravierender Unterschied zeigt sich in der strukturellen Kopplung. Herkömmliche Diagnostik endet abrupt mit der Auswertung des Testhefts. Das Bayreuther Testpaket verknüpft die Produkt- direkt mit der Prozessdiagnostik (Steinecke & Richter, 2025). Das 25-minütige Screening filtert die Risikogruppe im Klassenverband heraus. Erst dadurch wird der ökonomische Einsatz der anschließenden, qualitativen Bayreuther Förderdiagnostik (BFD) im Einzelinterview überhaupt leistbar (Steinecke & Richter, 2025).

Die konzeptionelle und inhaltliche Struktur des BRT 2 spiegelt exakt die in Kapitel 2.1 beschriebene dreiteilige fachdidaktische Modellierung (VNZ, VSWS, VRO) wider (Steinecke & Richter, 2025). Während die Handreichung dem Praktiker primär qualitative Auswertungshinweise bietet, widmet sich die vorliegende Arbeit einer umfassenden quantitativen und psychometrischen Überprüfung dieses Strukturmodells. Die exakte metrische Zusammensetzung der 18 Testaufgaben sowie das zugrundeliegende Bepunktungssystem der empirischen Stichprobe werden in Kapitel 3.2 dargelegt.

2.3 Methodologische und Testtheoretische Grundlagen

Die empirische Erfassung, Strukturierung und Modellierung komplexer akademischer Fähigkeiten erfordert ein präzises, definiertes mathematisches Regelwerk. Um den spezifischen Verteilungs- und Strukturcharakteristika des Bayreuther Rechentests (BRT-2) in einer Stichprobe von Grundschulkindern gerecht zu werden, bedarf es einer methodischen Verzahnung klassischer, nicht-parametrischer, netzwerktheoretischer sowie moderner multidimensionaler probabilistischer Analyseverfahren.

2.3.1 Verteilungscharakteristika und nicht-parametrische Hypothesenprüfung

Die mathematische Normalverteilung bildet das theoretische Ideal quantitativer Analysen, stellt jedoch in der pädagogisch-psychologischen Leistungsdiagnostik selten die empirische Realität dar. Leistungstests im Grundschulalter weisen aufgrund von Lehrplaneffekten häufig ausgeprägte Asymmetrien auf, die über die formalen Parameter der Schiefe und des Exzesses quantifiziert werden (Döring et al., 2016; Moosbrugger & Kelava, 2020).

Typisch sind hierbei Deckeneffekte (ceiling effects), bei denen ein großer Teil der Probanden Aufgaben fehlerfrei löst, was zu linksschiefen Verteilungen und Varianzeinschränkungen führt (Döring et al., 2016). Die formale Absicherung der Verteilungsform erfolgt über den Shapiro-Wilk-Test, welcher die Nullhypothese (H_0) einer exakten Normalverteilung in der Grundgesamtheit über den Quotienten aus der quadrierten Linearkombination der geordneten Stichprobenwerte und der Quadratsummen der Abweichung prüft (Shapiro & Wilk, 1965):

$$W_{SW} = \frac{\left(\sum_{i=1}^{N} a_i x_{(i)}\right)^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Bei größeren Stichproben reagiert dieser jedoch hochgradig sensitiv auf marginale, praktisch irrelevante Abweichungen, weshalb eine komparative Triangulation mit visuellen Inspektionen von Histogrammen methodisch zwingend erforderlich ist (Field et al., 2012). Liegt eine manifeste Verletzung der Normalverteilung vor, ist der Übergang zu rangbasierten, nicht-parametrischen Hypothesentests mathematisch indiziert:

1. Vergleich abhängiger Stichproben: Um Leistungsunterschiede zwischen verbundenen Messungen (den mathematischen Kompetenzbereichen derselben Lernenden) zu prüfen, wird der globale Friedman-Test als nicht-parametrisches Äquivalent zur Varianzanalyse mit Messwiederholung herangezogen (Friedman, 1937). Die Teststatistik χ_F^2 basiert auf den Rangsummen R_j der Anzahl der unabhängigen Variablen k und der Gesamtstichprobengröße N (Field et al., 2012).

$$\chi_F^2 = \left[\frac{12}{Nk(k+1)}\sum_{i=1}^{k} R_i^2\right] - 3N(k+1)$$

Zur Bestimmung der standardisierten Effektstärke respektive des globalen Übereinstimmungsmaßes wird hierbei Kendalls Konkordanzkoeffizient (W) berechnet, welcher die relative Abweichung der tatsächlichen Rangsummenquadrate vom maximal möglichen Wert quantifiziert (Bortz & Lienert, 2008; Kendall, 1938):

$$W_{Kendall} = \frac{12 \sum_{j=1}^{k} \left(R_j - \bar{R}\right)^2}{N^2 (k^3 - k)}$$

Im Falle eines signifikanten Globalbefundes werden paarweise Post-hoc-Vergleiche mittels des Wilcoxon-Vorzeichenrangtests (Wilcoxon signed-rank test) durchgeführt (Field et al., 2012; Wilcoxon, 1945), wobei die Teststatistik T der kleinere Wert aus den Summen der positiven (T^+) und negativen (T^-) Rangplätze ist. Die Alpha-Fehler-Kumulierung über multiple Vergleiche wird über die Bonferroni-Korrektur (Field et al., 2012) kontrolliert.

2. Vergleich unabhängiger Stichproben: Die Prüfung auf signifikante Differenzen zwischen den zentralen Tendenzen zweier unabhängiger Gruppen (Jungen und Mädchen) erfolgt über den Wilcoxon-Rangsummentest bzw. Mann-Whitney-U-Test (Bortz & Schuster, 2010; Mann & Whitney, 1947), dessen Teststatistik U die paarweisen Rangplatzüberschreitungen bestimmt:

$$U_1 = N_1 N_2 + \frac{N_1 (N_1 + 1)}{2} - R_1$$
$$U = \min(U_1, N_1\, N_2 - U_1)$$

Zur Quantifizierung der praktischen Relevanz (Effektstärke r) werden die extrahierten Teststatistiken in standardisierte z-Werte überführt und über die Transformation asymptotisch normiert (Rosenthal, 1991):

$$r = \frac{|z|}{\sqrt{N}}$$

2.3.2 Prüfung auf Varianzhomogenität

Die Prüfung der Varianzgleichheit stellt eine minimale Prämisse für die Validität von Gruppenvergleichen dar. Da klassische, parametrische Verfahren hochgradig sensitiv auf die Verletzung der Normalverteilung reagieren und Fehler erster Art artifiziell inflationieren, wird für unabhängige Stichproben der rangbasierte Fligner-Killeen-Test implementiert (Fligner & Killeen, 1976). Dieses Verfahren zeichnet sich durch eine extreme Robustheit gegenüber non-normalen Verteilungsformen und Ausreißern aus. Hierbei repräsentiert R_i die Rangposition der berechneten Abweichungsbeträge, während a_N die Gewichtungsfunktion darstellt, die diesen Rängen spezifische Werte zuweist, um die Teststatistik T_N zu bilden:

$$T_N = \sum_{i=1}^{m} a_N(R_i)$$

Ein nicht-signifikantes Ergebnis ($p \geq 0{,}05$) bestätigt die Annahme homogener Varianzen. Für abhängige, auf unterschiedlichen Skalenmetriken basierende Teststrukturen wird die

Dispersionsgleichheit komparativ über relative Interquartilsabstände (IQR) evaluiert; weiterhin wird der Wilcoxon-Vorzeichenrangtest bezüglich der Absolutabweichung herangezogen. Um die Dispersionsgleichheit zwischen Teststrukturen mit unterschiedlichen Rohwertmetriken statistisch prüfen zu können, wurden die Primärdaten einer linearen Transformation auf eine einheitliche Prozentmetrik unterzogen. Im Gegensatz zu nicht-linearen Normwerten bleibt bei dieser Form der Skalierung das Intervallskalenniveau sowie die empirische Verteilungsform erhalten (Bortz & Schuster, 2010).

2.3.3 Mathematische Modellierung manifester und latenter Zusammenhänge

Zur Bestimmung ungerichteter Assoziationen zwischen diagnostischen Teilbereichen bei verletzter Normalverteilungsannahme oder dem Vorliegen kategorialer Datenstrukturen wird die Spearman-Rang-Korrelation (r_s) berechnet (Spearman, 1904). Im Gegensatz zur Pearson-Produkt-Korrelation basiert dieses Verfahren auf den Differenzen d_i der monotonen Rangplätze der Datenpunkte (Kendall, 1938):

$$r_s = 1 - \frac{6 \sum_{i=1}^{N} d_i^2}{N(N^2 - 1)}$$

Dadurch agiert das Maß vollständig robust gegenüber extremen Ausreißern. Die Interpretation der praktischen Relevanz der resultierenden Assoziationen folgt den Richtlinien von Cohen (1988), welche Koeffizienten ab $|r_s| \geq 0{,}10$ als schwach, ab $|r_s| \geq 0{,}30$ als moderat und ab $|r_s| \geq 0{,}50$ als stark klassifizieren.

2.3.4 Axiomatik der Klassischen Testtheorie (KTT) und Itemmetrik

Das traditionelle Fundament der psychometrischen Analyse ruht auf dem Messfehlermodell der klassischen Testtheorie. Die fundamentale Axiomatik postuliert, dass sich ein beobachteter Messwert (X) additiv aus einem konstanten wahren Wert (T, *true score*) und einem stochastischen, unkorrelierten Messfehler (E, *error*) zusammensetzt (Moosbrugger & Kelava, 2020).

Daraus leitet sich die klassische Itemanalyse ab, welche primär über den Schwierigkeitsindex (P_i) operationalisiert wird. Für gemischt-kategoriale Teststrukturen mit maximal erreichbaren Punktwerten x_{max} berechnet sich dieser über den empirischen Mittelwert $\bar{x}_i$:

$$P_i = \frac{\bar{x}_i}{x_{max}}$$

Der resultierende Kennwert gibt den relativen Punkteertrag eines Items im Verhältnis zum erreichbaren Maximum an und dient als zentraler Indikator zur deskriptiven Beurteilung der Aufgabenanforderungen innerhalb der Stichprobe. Die KTT leidet jedoch unter einer strikten Stichprobenabhängigkeit ihrer Parameter, da das mathematische Fähigkeitsniveau der Untersuchungspopulation direkt auf die Höhe des Index einwirkt.

2.3.5 Konzepte der klassischen und modellbasierten Reliabilitätsbeurteilung

Die Bestimmung der internen Konsistenz als Indikator der Messgenauigkeit erfolgt traditionell über Cronbachs Alpha (α) unter Berücksichtigung der Itemzahl k, der Itemvarianz σ_i^2 und der Gesamttestvarianz σ_t^2 (Cronbach, 1951):

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_t^2}\right)$$

Da Alpha eine strikte τ-Äquivalenz (alle Items besitzen dieselbe Steigung im Bezug auf das latente Merkmal, also identische Ladungen; Moosbrugger & Kelava, 2020) fordert, führt jede Ungleichheit der Item-Faktorladungen zu einer systematischen Unterschätzung der wahren Reliabilität (Mair, 2018). Als robustere, ladungsgewichtete Alternative wird daher McDonalds Omega Total (ω_t) berechnet (Moosbrugger & Kelava, 2020; Rodriguez et al., 2016), welches die wahren Ladungen λ_i und die Residualvarianzen θ_i aus einem latenten Ein-Faktor-Modell mathematisch verknüpft:

$$\omega_t = \frac{\left(\sum_{i=1}^{k}\lambda_i\right)^2}{\left(\sum_{i=1}^{k}\lambda_i\right)^2 + \sum_{i=1}^{k}\theta_i}$$

Die qualitative Einordnung folgt George & Mallery (2016), welche Werte ab 0,70 als akzeptabel, ab 0,80 als gut und ab 0,90 als exzellent definieren. Für hierarchisch aufgebaute, multidimensionale Skalenstrukturen wird zudem McDonalds Omega Hierarchical (ω_h) berechnet, um den Varianzanteil zu isolieren, der rein auf einen übergeordneten Generalfaktor (G) unter Berücksichtigung der Ladungen der m spezifischen Gruppenfaktoren (F) zurückzuführen ist (Zinbarg et al., 2005):

$$\omega_h = \frac{\left(\sum_{i=1}^{k}\lambda_{G_i}\right)^2}{\left(\sum_{i=1}^{k}\lambda_{G_i}\right)^2 + \sum_{j=1}^{m}\left(\sum_{i\in F_j}\lambda_{F_{ji}}\right)^2 + \sum_{i=1}^{k}\theta_i}$$

Ein Wert von $\omega_h \geq 0{,}70$ gilt hierbei als statistische Legitimation für die Bildung eines globalen Gesamtwerts (Moosbrugger & Kelava, 2020).

2.3.6 Graphentheoretische Modellierung von Fehlermustern (Item-Netzwerke)

Die netzwerktheoretische Psychometrie interpretiert statistische Abhängigkeiten zwischen Aufgabenkomplexen als komplexe adaptive Systeme interagierender Komponenten (Epskamp et al., 2018). Ein topologisches Netzwerk wird formal als mathematischer Graph $G = (V, E)$ definiert (Newman, 2010), wobei die Knoten (V) die Testaufgaben und die Kanten (E) deren funktionale Beziehung repräsentieren.

Die Modellierung basiert auf der Berechnung paarweise bedingter Wahrscheinlichkeiten innerhalb einer quadratischen asymmetrischen Fehlermatrix (Wickham, 2007). Die bedingte Wahrscheinlichkeit $P(A|B)$, eine Aufgabe A falsch zu lösen, falls Aufgabe B bereits

fehlerhaft bearbeitet wurde, berechnet sich als Quotient aus der gemeinsamen Wahrscheinlichkeit beider Fehler und der Wahrscheinlichkeit des bedingten Fehlers B:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Die räumliche Anordnung und visuelle Verdichtung enger Fehlerpfade erfolgt über ein kraftgesteuertes Fruchterman-Reingold-Layout (Fruchterman & Reingold, 1991), welches die Kantenstärken proportional zu den bedingten Übergangswahrscheinlichkeiten als elastische Federn simuliert (Pedersen, 2026a, 2026b).

2.3.7 Personenzentrierte Typenbildung: Clusteranalyse und Validierung

Im Gegensatz zu variablenzentrierten Ansätzen zielen personenzentrierte Methoden auf die Identifikation homogener Subgruppen von Individuen mit identischen Leistungsprofilen ab. Das mathematische Standardverfahren bildet das partitionierende k-Means-Clustering (Charrad et al., 2014; MacQueen et al., 1967), welches Probanden iterativ so verschiebt, dass die Summe der quadrierten euklidischen Distanzen innerhalb der k Cluster zu den jeweiligen Zentroidtransformationen μ_j, die *Within-Cluster Sum of Squares* (WSS), minimiert wird (Charrad et al., 2014):

$$WSS = \sum_{j=1}^{k} \sum_{i \in C_j} \left\| x_i - \mu_j \right\|^2$$

Die mathematische Bestimmung der optimalen Clusterzahl (k) erfolgt sequenziell über die Triangulation dreier Validierungsindizes (Charrad et al., 2014; Kaufman & Rousseeuw, 2005):

1. Ellbogen-Kriterium (WSS): Grafische Inspektion des Verlaufs der intra-clusterzentrischen Varianz, bei der der Punkt des steilsten Abflachungswinkels den optimalen Partitionierungsraum anzeigt.
2. Durchschnittliche Silhouettenbreite: Ein mathematischer Validitätsindex, welcher für jedes Objekt i die Differenz zwischen der mittleren Distanz zum nächsthöheren Cluster $b(i)$ und der mittleren Intracluster-Distanz $a(i)$ normiert (Rousseeuw, 1987):
$$s(i) = \frac{b(i) - a(i)}{\max\big(a(i), b(i)\big)}$$
Maximale gemittelte Werte über alle Objekte indizieren eine optimale Trennung.
3. Gap-Statistik nach Tibshirani: Ein simulationsbasiertes Verfahren, welches die logarithmische In-Cluster-Varianz W_k der empirischen Daten mit dem Erwartungswert einer über Monte-Carlo-Simulationen generierten, gleichverteilten Nullreferenzverteilung E_n^* vergleicht (Tibshirani et al., 2001):
$$Gap(k) = E_n^*\{\log(W_k)\} - \log(W_k)$$

Zur geometrischen Qualitätskontrolle hochdimensionaler Clustergrenzen wird eine bivariate 2D-Partitionierung durchgeführt, bei welcher die multidimensionalen Merkmalsräume über eine integrierte Hauptkomponentenanalyse (PCA) auf eine zweidimensionale Ebene projiziert werden (Kassambara & Mundt, 2026; Mair, 2018).

2.3.8 Fundamente der multidimensionalen Eigenschafts- und Faktorenanalyse (EFA/CFA)

Die mathematische Aufklärung latenter Variablenstrukturen setzt eine fundierte Überprüfung der Datenmatrix voraus. Die statistische Eignung der Item-Korrelationsmatrix für faktorenanalytische Verfahren wird über das Kaiser-Meyer-Olkin-Kriterium (KMO/ *Measure of Sampling Adequacy*) abgesichert, welches das Verhältnis der quadrierten Korrelationen r_{ij}^2 zur Summe aus quadrierten Korrelationen und quadrierten Partialkorrelationen a_{ij}^2 berechnet (Kaiser, 1974):

$$KMO = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} a_{ij}^2}$$

Ergänzend prüft der Bartlett-Test auf Sphärizität über eine χ^2-Transformation der Determinante der Korrelationsmatrix $|R|$ die Nullhypothese, dass die vorliegende Matrix eine Identitätsmatrix darstellt und somit ungeeignet für Dimensionsreduktion ist (Bartlett, 1950).

Während die spätere konfirmatorische Überprüfung an ein festes theoretisches Modell gebunden ist, verfolgt die Explorative Faktorenanalyse (EFA) das Ziel, die dimensionale Struktur der 18 Testaufgaben des BRT-2 datengeleitet offenzulegen. Psychometrisch geschieht dies unter der Prämisse, alle 18 Items simultan und ohne apriorische Annahmen über eine feste Zuordnung zu bestimmten mathematischen Kompetenzbereichen zu analysieren (Mair, 2018).

Das mathematische Fundament der EFA basiert auf dem Grundtheorem der klassischen Faktorenanalyse. Dieses postuliert, dass sich der standardisierte Messwert einer manifesten Variablen (Aufgabe) x_i als Linearkombination aus m gemeinsamen, latenten Faktoren (f_i) und einem itemspezifischen Residuumsfaktor (e_i, *unique factor*) darstellen lässt (vgl. Moosbrugger & Kelava, 2020):

$$x_i = \lambda_{i1} f_1 + \lambda_{i2} f_2 + \cdots + \lambda_{im} f_m + e_i = \sum_{j=1}^{m} \lambda_{ij} f_j + e_i$$

Hierbei repräsentiert λ_{ij} die Faktorenladung der Variablen i auf dem latenten Faktor j. In kompakter Matrixnotation lässt sich das Modell für den gesamten Vektor der manifesten Aufgabenvariablen x wie folgt formalisieren (Mair, 2018):

$$x = \Lambda f + e$$

wobei Λ die $p \times m$-Matrix der Faktorenladungen darstellt (mit p als Anzahl der Items). Unter der Bedingung standardisierter Variablen und unkorrelierter Fehlerglieder ergibt sich daraus die fundamentale Varianzkomposition eines jeden Items. Die Gesamtvarianz ($Var(x_i) = 1$) spaltet sich additiv in die Kommunalität (h_i^2) und die Einzelrestvarianz (u_i^2) (Moosbrugger & Kelava, 2020):

$$1 = h_i^2 + u_i^2 = \sum_{j=1}^{m} \lambda_{ij}^2 + Var(e_i)$$

Die Kommunalität h_i^2 definiert dabei denjenigen Varianzanteil einer Aufgabe, der durch die extrahierten gemeinsamen Faktoren simultan aufgeklärt werden kann, während sich die Einzelrestvarianz u_i^2 aus der spezifischen Aufgabenvarianz und dem stochastischen Messfehler zusammensetzt (Field et al., 2012). Die Bestimmung der exakten Anzahl der zu extrahierenden gemeinsamen Faktoren m basiert auf der mathematischen Triangulation aus dem Kaiser-Guttman-Kriterium (Guttman, 1954), der Parallelanalyse nach Horn (Horn, 1965) und der visuellen Inspektion der Scree-Plots.

Da die primär extrahierte Faktorenlösung mathematisch unendlich viele Koordinatensysteme zulässt, ist eine Faktorenrotation zur Interpretation der Ladungsmuster erforderlich. Für Konstrukte, bei denen theoretisch von gekoppelten kognitiven Fähigkeiten ausgegangen werden muss, wird eine schiefwinklige (oblique) Promax-Rotation implementiert (Hendrickson & White, 1964). Im Gegensatz zu orthogonalen Rotationsverfahren lässt Promax eine Korrelation zwischen den latenten Faktoren rechnerisch zu, da kognitive Teildimensionen in der psychologischen Realität meist hochgradig miteinander interagieren (Field et al., 2012).

Zur empirischen Absicherung der globalen Modellpassung der extrahierten Ladungsarchitektur (Λ) sowie der Kovarianzmatrix der Faktoren (Φ) werden die Daten im Anschluss einer Konfirmatorischen Faktorenanalyse (CFA) unterzogen, welche die implizierte Kovarianzmatrix Σ schätzt (Rosseel, 2012):

$$\Sigma = \Lambda \Phi \Lambda^T + \Theta$$

2.3.9 Einführung in die Item-Response-Theorie (IRT)

Die moderne Item-Response-Theorie definiert die Wahrscheinlichkeit für das Lösen einer Aufgabe als mathematische Funktion der kontinuierlichen, latenten Personenfähigkeit (θ) (Moosbrugger & Kelava, 2020). Für gemischt-kategoriale Teststrukturen bildet das Generalized Partial Credit Model (GPCM) den mathematischen Schätzstandard (Muraki, 1992). Die Wahrscheinlichkeit $P_{ix}(\theta)$, dass eine Person auf Item i die Bewertungsstufe x innerhalb der Antwortkategorien m_i erzielt, berechnet sich unter Einbezug des Diskriminationsparameters a_i und des Schwellenparameters b_{iv} wie folgt (Mair, 2018):

$$P_{ix}(\theta) = \frac{\exp[\sum_{v=0}^{x} a_i(\theta - b_{iv})]}{\sum_{c=0}^{m_i} \exp[\sum_{v=0}^{c} a_i(\theta - b_{iv})]}$$

Die funktionale Darstellung dieser Lösungswahrscheinlichkeiten erfolgt über die Item-Charakteristischen Kurven (ICC/ *Category Response Curves*, CRC). Die lokale Messpräzision wird über die Item-Informationskurven (IIC) bestimmt, welche für das GPCM über das erste und zweite Moment der Kategoriestufen formalisiert wird (Reckase, 2009):

$$I_i(\theta) = a_i^2 \sum_{x=0}^{m_i} x^2 P_{ix}(\theta) - \left[a_i \sum_{x=0}^{m_i} x P_{ix}(\theta) \right]^2$$

Die Summierung aller Einzelinformationen liefert die globale Test-Informationsfunktion (TIF) $I(\theta) = \sum I_i(\theta)$ (Chalmers, 2012), aus welcher direkt der kontinuierliche Standardmessfehler abgeleitet wird (Moosbrugger & Kelava, 2020):

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}}$$

Die Modellparameter werden rechenintensiv über numerische EM-Algorithmen geschätzt, wobei zur Stabilisierung der hochdimensionalen Integrale adaptive Quasi-Monte-Carlo-Approximationen (QMCEM) implementiert werden (vgl. Mair, 2018).

Der globale Fit wird über die Limited-Information-M_2-Statistik (Maydeu-Olivares & Joe, 2006) und der lokale Item-Fit über die informationsbasierte $S - X^2$-Statistik (Orlando & Thissen, 2000) abgesichert, während die lokale stochastische Unabhängigkeit über Yens Q_3-Statistik (Residuenkorrelation; Yen, 1984) evaluiert wird.

2.3.10 Testfairness und Invarianzprüfung: DIF-Analyse

Das psychometrische Gütekriterium der Testfairness bildet eine fundamentale Säule für die Validität diagnostischer Schlussprozesse. Es fordert den mathematischen Nachweis, dass das eingesetzte Messinstrument über verschiedene Subpopulationen (z. B. definiert durch Geschlecht, soziokulturellen Hintergrund oder Migrationsstatus) hinweg strukturell identische Messeigenschaften aufweist. Formal wird dieses Konzept über die Messinvarianz (measurement invariance) operationalisiert (Mellenbergh, 1989; Moosbrugger & Kelava, 2020).

Mathematisch liegt vollständige Messinvarianz für ein Item vor, wenn die bedingte Wahrscheinlichkeit einer bestimmten Antwort bei gegebener latenter Fähigkeitsausprägung θ vollständig unabhängig von der Gruppenzugehörigkeit G ist. Für ein polytomes oder dichotomes Item x_i mit den Antwortstufen x lässt sich diese Unabhängigkeit wie folgt darstellen (Mellenbergh, 1989):

$$P(x_i = x|\theta, G = Reference) = P(x_i = x|\theta, G = Focal)$$

Hierbei bezeichnet die Reference-Gruppe die mathematische Vergleichsbasis (z. B. die Majoritätspopulation) und die Focal-Gruppe die psychometrisch zu überprüfende Zielgruppe. Wird diese Gleichung signifikant verletzt, resultiert daraus das Vorliegen von Dif-

ferentiellen Aufgabenfunktionen bzw. Differential Item Functioning (DIF). Ein Item mit DIF verzerrt den diagnostischen Messprozess, da Personen zweier Teilpopulationen trotz exakt identischer latenter Merkmalsausprägung θ systematisch unterschiedliche Wahrscheinlichkeiten besitzen, die Aufgabe erfolgreich zu lösen (Mair, 2018; Moosbrugger & Kelava, 2020).

In der probabilistischen Testtheorie wird kriteriengeleitet zwischen zwei mathematischen Ausprägungsformen des DIF unterschieden:

1. Uniformes DIF: Liegt vor, wenn der Effekt der Gruppenzugehörigkeit über das gesamte kontinuierliche Fähigkeitsspektrum θ hinweg konstant bleibt (Mair, 2018). Im mathematischen Framework des Generalized Partial Credit Models (GPCM) manifestiert sich uniformes DIF als eine systematische Verschiebung der Schwellenparameter (b_{iv}) zwischen den Gruppen, während der Diskriminationsparameter (a_i) invariant bleibt ($a_{i,Ref} = a_{i,Foc}$ mit $b_{iv,Ref} \neq b_{iv,Foc}$). Die Item-Charakteristischen Kurven (ICCs) der beiden Gruppen verlaufen in diesem Szenario parallel verschoben zueinander (Mair, 2018; Moosbrugger & Kelava, 2020).
2. Nicht-uniformes DIF: Ist durch eine statistische Interaktion zwischen der latenten Fähigkeit θ und der Gruppenzugehörigkeit G charakterisiert. Dies bedeutet, dass das Ausmaß und die Richtung der Benachteiligung vom konkreten Punkt auf dem Fähigkeitskontinuum θ abhängen. Mathematisch ist dies der Fall, wenn sich die Diskriminationsparameter des Items zwischen den Subpopulationen signifikant unterscheiden ($a_{i,Ref} \neq a_{i,Foc}$). Die ICCs der Gruppen weisen in diesem Fall unterschiedliche Steigungen auf und schneiden sich regelhaft im Koordinatenraum (Mair, 2018).

Die psychometrische Detektion von DIF-Effekten innerhalb einer latenten Modellarchitektur erfolgt standardmäßig über den Likelihood-Quotienten-Test (*Likelihood Ratio Test*, LRT). Hierbei wird für jedes Item sequenziell ein kompaktes Nullmodell (M_0), welches die Parameter des untersuchten Items über beide Gruppen hinweg restriktiv gleichsetzt (Invarianzannahme), gegen ein unrestringiertes Alternativmodell (M_1) getestet, in welchem die Itemparameter für die Focal- und Reference-Gruppe frei geschätzt werden dürfen. Die Teststatistik G^2 basiert auf dem Verhältnis der maximalen Likelihood-Werte der beiden Modelle und ist asymptotisch χ^2-verteilt (Moosbrugger & Kelava, 2020):

$$G^2 = -2 \ln\left(\frac{L_{M_0}}{L_{M_1}}\right) = 2(\ln L_{M_1} - \ln L_{M_0})$$

Die Anzahl der Freiheitsgrade (Δdf) entspricht dabei exakt der Anzahl der zusätzlich freigesetzten Itemparameter im Alternativmodell M_1 (ein Freiheitsgrad für uniformes DIF durch Freisetzung der Lokation; zwei Freiheitsgrade für die simultane Prüfung auf nicht-uniformes DIF durch zusätzliche Freisetzung der Diskrimination; Moosbrugger & Kelava, 2020).

Da im Rahmen einer umfassenden Testvalidierung die DIF-Analyse simultan für alle p Aufgaben einer Skala durchgeführt werden muss, tritt das Problem der Alpha-Fehler-Inflationierung durch multiples Testen auf. Um das Risiko einer artifiziellen Identifikation von DIF-Effekten (Fehler erster Art) zu kontrollieren, ohne die statistische Power drastisch zu vermindern (wie es bei der konservativen Bonferroni-Korrektur der Fall wäre), wird das Benjamini-Hochberg-Verfahren implementiert (Benjamini & Hochberg, 1995). Dieses Framework kontrolliert direkt die *False Discovery Rate* (FDR), also den erwarteten Anteil fälschlicherweise abgelehnter Nullhypothesen an der Gesamtzahl aller signifikanten Befunde (Benjamini & Hochberg, 1995).

Das mathematische Prozedere ist wie folgt definiert. Die aus den einzelnen Likelihood-Quotienten-Tests der Items extrahierten empirischen p-Werte werden zunächst in aufsteigender Reihenfolge sortiert (Benjamini & Hochberg, 1995):

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(p)}$$

Für ein vorab definiertes nominales Signifikanzniveau α (z. B. $\alpha = 0{,}05$) wird anschließend der größte Rangindex k bestimmt, für den die folgende Ungleichheit erfüllt ist (Field et al., 2012):

$$p_{(k)} \leq \frac{k}{p} \cdot \alpha$$

Ist dieser maximale Index k identisch, werden die Nullhypothesen für alle Items, deren p-Werte einen Rangplatz kleiner oder gleich k einnehmen ($i \leq k$), verworfen. Für diese Aufgaben gilt das Vorliegen eines signifikanten und populationsabhängigen Differential Item Functioning als mathematisch nachgewiesen.

2.3.11 Hierarchische Ladungsarchitekturen: Das Bifaktor-Modell

Das orthogonale Bifaktor-Modell postuliert, dass die Varianz einer jeden manifesten Testaufgabe x_i (mit $i = 1, \ldots, p$) simultan durch zwei Arten von latenten Variablen determiniert wird, einem globalen Generalfaktor (G), welcher die gemeinsame Varianz über alle p Items der Skala hinweg aufklärt, und eine Menge von m zueinander vollständig orthogonalen spezifischen Gruppenfaktoren (F_j), welche die verbleibende Cluster-Varianz innerhalb bestimmter Aufgabenkomplexe binden (Reise, 2012). Formal lässt sich das Modell über die folgende Linearkombination definieren (Moosbrugger & Kelava, 2020):

$$x_i = \lambda_{Gi} G + \lambda_{F_{ji}} F_j + e_i$$

Unter den strikten Orthogonalitätsbedingungen ($Cov(G, F_j) = 0;\ Cov(F_j, F_k) = 0$) spaltet sich die Gesamtvarianz einer jeden Aufgabe additiv auf:

$$Var(x_i) = 1 = \lambda_{Gi}^2 + \lambda_{F_{ji}}^2 + Var(e_i)$$

Ein Verfahren zur rechnerischen Schätzung dieser orthogonalen Ladungsarchitekturen aus einer primär extrahierten, korrelierten Faktorlösung höherer Ordnung stellt die

Schmid-Leiman-Transformation dar (Schmid & Leiman, 1957). Lädt ein Sekundärfaktor höherer Ordnung einen Primärfaktor f_i mit dem Koeffizienten γ_j und lädt dieser Primärfaktor das manifeste Item i mit der Ladung a_{ij}, so transformieren sich die Bifaktor-Ladungen mathematisch wie folgt (vgl. Revelle, 2026):

$$\lambda_{Gi} = a_{ij} \cdot \gamma_j$$

$$\lambda_{F_{ji}} = a_{ij} \cdot \sqrt{1 - \gamma_j^2}$$

Die zentrale psychometrische Fragestellung bei der Anwendung lautet, ob ein komplexer multidimensionaler Testraum im Sinne der Messpraxis als essenziell eindimensional eingestuft und über einen einzigen, aggregierten Score abgerechnet werden darf (Rodriguez et al., 2016). Zur mathematischen Absicherung dieses Zustands werden zwei strukturanalytische Indizes trianguliert:

1. Explained Common Variance (ECV): Quantifiziert das proportionale Verhältnis der durch den Generalfaktor erklärten gemeinsamen Varianz zur gesamten aufgeklärten gemeinsamen Varianz aller Modellfaktoren (Rodriguez et al., 2016):
$$ECV = \frac{\sum_{i=1}^{p} \lambda_{Gi}^2}{\sum_{i=1}^{p} \lambda_{Gi}^2 + \sum_{j=1}^{m} \sum_{i \in F_j} \lambda_{F_{ji}}^2}$$
2. McDonalds hierarchisches Omega (ω_h): Bestimmt den Anteil der Gesamtvarianz der Testscores, der exakt und ausschließlich auf den Generalfaktor zurückgeht. Die formale Definition und mathematische Herleitung entsprechen der in Abschnitt 2.3.5 dargestellten Skalenformel (Revelle, 2026).

Erreicht ein Datensatz Werte von $ECV > 0{,}70$ in Kombination mit einem $\omega_h > 0{,}70$, gilt der Nachweis einer essenziellen Eindimensionalität als mathematisch erbracht (Reise, 2012; Rodriguez et al., 2016). In diesem Fall ist die Operationalisierung des gesamten Testraums über einen einzigen, aggregierten Gesamtwert mathematisch vollständig legitimiert.

3 Methodik

Das vorliegende Kapitel beschreibt das methodische Vorgehen zur empirischen Überprüfung und psychometrischen Validierung des Bayreuther Rechentests (BRT-2). Zur systematischen und replizierbaren Beantwortung der 13 forschungsleitenden Fragestellungen dieser Arbeit ist eine transparente Darlegung des Untersuchungsdesigns erforderlich. Zu diesem Zweck wird zunächst die empirische Stichprobe (N = 574) hinsichtlich ihrer demografischen Zusammensetzung sowie des konkreten Ablaufs der Datenerhebung charakterisiert (Kapitel 3.1). Darauf aufbauend widmet sich Kapitel 3.2 dem eingesetzten Erhebungsinstrument, indem es die metrische Struktur der 18 Testaufgaben sowie das zugrundeliegende Bepunktungssystem detailliert aufschlüsselt. Abschließend werden in Kapitel 3.3 die verwendeten statistischen Analyseverfahren und die softwarebasierte Umsetzung in der Programmierumgebung R (R Core Team, 2026) dargelegt, welche das methodologische Fundament für die nachfolgende Ergebnisdarstellung bilden.

3.1 Stichprobenbeschreibung

Die empirische Basis der vorliegenden Arbeit bildet kein willkürliches Feldexperiment, sondern die originale Normierungsstichprobe des Bayreuther Rechentests (BRT-2). Dies sichert den nachfolgenden Analysen eine fundamentale Validität und Praxisrelevanz. Die Normierungsdaten wurden von den Entwicklerinnen des Testverfahrens im bayrischen Raum erhoben. Die Daten wurden zu Beginn des zweiten Schulhalbjahres in den Monaten Februar und März 2025 erhoben (Steinecke & Richter, 2025).

Die ursprüngliche Normierungsstichprobe der Handreichung umfasst insgesamt $N = 575$ Lernende der zweiten Jahrgangsstufe (Steinecke & Richter, 2025). Die Erhebung wurde als Vollerhebung in den jeweils kompletten zweiten Klassen der teilnehmenden Schulen durchgeführt. Um eine ausgewogene und für den bayrischen Raum repräsentative Stichprobe zu gewährleisten, wurden Schulen aus unterschiedlichen regionalen geografischen Räumen in Mittel- und Oberfranken ausgewählt (Steinecke & Richter, 2025):

- Großstädtischer Raum: Schulen in den Großstädten Nürnberg und Erlangen.
- Mittelstädtischer Raum: Schulen in Schwabach.
- Ländlicher Raum: Schulen in verschiedenen ländlichen Gemeinden Mittel- und Oberfrankens.

„Aufgrund dieser Verteilung der Schulen und der Durchführung von Vollerhebungen an den Schulen kann die Probandengruppe als weitgehend repräsentativ für die Gruppe aller Schülerinnen und Schüler in der Jahrgangsstufe 2 in Bayern angesehen werden“ (Steinecke & Richter, 2025).

Für die im Rahmen dieser Arbeit durchgeführten komplexen, modellbasierten Analysen, insbesondere die konfirmatorische Faktorenanalyse (CFA), die Item-Response-Theorie-Modellierung (IRT) sowie die sensitiven Analysen zum Differential Item Functioning

(DIF), war jedoch eine strikte Datenbereinigung unumgänglich. Da die genannten multivariaten Verfahren äußerst sensibel auf fehlende Daten reagieren, mussten alle Testprotokolle mit unvollständigen Aufgabenbearbeitungen entfernt werden. Hiernach umfasste der Datensatz noch $N = 574$ Lernende. Für manche Verfahren mussten auch Daten mit fehlendem Geschlechtseintrag aus dem Analysedatensatz ausgeschlossen werden. Hierbei wird eine Teilstichprobe von $N = 467$ Lernenden erhalten, die sich wie folgt zusammensetzt:

- Mädchen: $n = 243$ (52,03 % der bereinigten Stichprobe)
- Jungen: $n = 224$ (47,96 % der bereinigten Stichprobe)

Dieses ausgewogene Verhältnis ist für die statistische Power und die Fehlerstabilität der geschlechtsspezifischen DIF-Analysen ideal, da es mathematische Verzerrungen durch stark asymmetrische Gruppengrößen effektiv minimiert.

Da alle Daten aus dem bayrischen Raum stammen, profitiert die Stichprobe von einer extrem hohen curricularen Homogenität.

3.2 Das Erhebungsinstrument: Struktur der 18 Items (A1-A18) und Bepunktungssystem

Zur Erfassung der mathematischen Basiskompetenzen der Lernenden im zweiten Schuljahr wurde der in Kapitel 2.2 theoretisch verortete Bayreuther Rechentest (BRT-2) eingesetzt. Das Instrument besteht in der vorliegenden Form aus insgesamt 18 standardisierten Testaufgaben (Items A1 bis A18). Entsprechend der fachdidaktischen Konzeption des Drei-Säulen-Modells wurde das Konstrukt der arithmetischen Kompetenz über drei Subskalen operationalisiert, die jeweils exakt sechs Aufgaben umfassen (Steinecke & Richter, 2025):

1. Verständnis für natürliche Zahlen (VNZ): Items A1 bis A6
2. Verständnis für das Stellenwertsystem (VSWS): Items A7 bis A12
3. Verständnis für Rechenoperationen (VRO): Items A13 bis A18

Die Aufgabenformate variieren je nach mathematischem Inhaltsbereich, um die spezifischen kognitiven Operationen der Lernenden adäquat abzubilden. Während im Bereich des Stellenwertsystems (VSWS) und der Rechenoperationen (VRO) vorwiegend halboffene Notationsverfahren oder das Eintragen in symbolische Strukturen gefordert sind, erfordern die Aufgaben des Zahlverständnisses (VNZ) teils mehrteilige Anforderungsarchitekturen, wie beispielsweise das qualitative Einordnen von Zahlen oder das Fortsetzen mathematischer Muster (Steinecke & Richter, 2025).

Für die testtheoretische Analyse und die anschließende Modellierung der Daten ist die metrische Beschaffenheit des Bepunktungssystems (Scoring-Modell) von zentraler Bedeutung. Der BRT-2 nutzt in der Auswertung kein rein dichotomes Format, sondern ist als gemischt-polytomes Testverfahren strukturiert (Steinecke & Richter, 2025):

- Dichotome Items: Ein Teil der Aufgaben (Items A1, A2, A7, A8) wird streng dichotom kodiert. Für eine vollständig korrekte Lösung wird ein Punkt vergeben, bei einer fehlerhaften oder nicht bearbeiteten Aufgabe null Punkte (Steinecke & Richter, 2025).
- Polytome Items: Die restlichen existierenden Aufgaben erlauben eine abgestufte Bepunktung. Aufgrund mehrteiliger Lösungsanforderungen können Lernende hier neben einem Minimalwert von 0 Punkten auch Teilpunkte (z. B. 1 Punkt) oder die volle Punktzahl (2 Punkte bei den Items A3, A4, A9, A10, A13, A14, A15, A16, A17 und A18 sowie 4 Punkte bei den Items A5, A6, A11 und A12) für partiell korrekte Leistungen erzielen (Steinecke & Richter, 2025).

Diese gemischt-polytome Datenstruktur hat fundamentale Auswirkungen auf die mathematischen Voraussetzungen der psychometrischen Analysen. Sie bedingt direkt, dass klassische Koeffizienten der internen Konsistenz (wie Cronbachs α), welche eine strikte Einhaltung der essenziellen τ-Äquivalenz sowie rein dichotome oder kontinuierliche Daten voraussetzen, verzerrte Schätzungen liefern können (vgl. McNeish, 2018; Moosbrugger & Kelava, 2020). Dies bildet die methodische Rechtfertigung für den in Kapitel 3.3 und Kapitel 4.2 dargelegten vergleichenden Einsatz moderner, latenter Reliabilitätsmaße wie McDonalds ω (Dunn et al., 2014).

Die Durchführung der Datenerhebung erfolgte nach einem standardisierten Protokoll im Klassenverband, um externe Störvariablen zu minimieren. Der BRT-2 wurde als ökonomisches, papierbasiertes Testverfahren appliziert, wobei die Instruktionen durch die jeweiligen Lehrkräfte anhand einer fest vorgegebenen Vorleseanweisung vermittelt wurden (vgl. Steinecke & Richter, 2025). Die Bearbeitungszeit war mit „[...] 25 Minuten, gegebenenfalls zuzüglich Nachteilsausgleich“ (Steinecke & Richter, 2025, S. 15) so dimensioniert, dass das Verfahren den Charakter eines Speed-Tests aufweist. Zudem wurde die benötigte Bearbeitungszeit, wie für zeiteingeschränkte Verfahren üblich, protokolliert (Moosbrugger & Kelava, 2020). Zusätzliche Hilfsmittel wie Taschenrechner oder didaktische Rechenmaterialien waren nicht zugelassen.

3.3 Statistische Datenanalyse und Voraussetzungsprüfung

Um die aufgestellten Forschungshypothesen zu überprüfen, wurden die Daten mit der Statistiksoftware R (Version 4.6.0; R Core Team, 2026) quantitativ ausgewertet. Das Signifikanzniveau für alle inferenzstatistischen Tests wurde vorab auf $\alpha = 0{,}05$ festgesetzt.

Um über alle ergebnisrelevanten Kapitel hinweg eine strikte visuelle Konsistenz, typografische Homogenität und mathematische Präzision der Befunde zu garantieren, folgt die grafische Aufbereitung einem standardisierten gestalterischen Protokoll. Die typografische Standardisierung und präzise Vektorisierung der gewählten Schrift (Aptos) über alle Rendering-Engines hinweg wird systemweit durch das Paket *systemfonts* (Version 1.3.2; Pedersen et al., 2026) abgesichert. Um die Grafiken als hochauflösende Abbildungen zu exportieren, wird das R-Paket *svglite* (Version 2.2.2; Wickham et al., 2026) genutzt. Die

prozedurale Erstellung der Abbildung teilt sich in drei funktionale Darstellungsklassen auf.

Die erste Klasse umfasst die Verteilungs- und Partitionierungsdiagramme. Die univariaten und bivariaten Verteilungsstrukturen der manifesten Rohwerte sowie der extrahierten latenten Parameter werden über das universelle Grafikpaket *ggplot2* (Version 4.0.3; Wickham, Chang et al., 2026) abgebildet, wobei die notwendige Datenstrukturierung und die Matrix-Transformationen über das Kern-Ökosystem des *tidyverse* (Version 2.0.0; Wickham et al., 2019) sowie über das Paket *reshape2* (Version 1.4.5; Wickham, 2010) vorbereitet werden. Hierzu zählen Histogramme zur Darstellung von Häufigkeitsverteilungen, 2D-Partitionierungen zur Abgrenzung von Merkmalsräumen sowie Boxplots zur Veranschaulichung von Lage- und Dispersionsmaßen. Das mathematische Kriterium für die automatische Kennzeichnung von Ausreißerpunkten innerhalb der Boxplots ist matrixweit über das 1,5-Fache des Interquartilsabstand (IQR) oberhalb des dritten bzw. unterhalb des ersten Quartils definiert (Bortz & Schuster, 2010). Um komplexe, mehrteilige Abbildungsstrukturen und Panel-Vergleiche exakt aufeinander abzustimmen, wird das Paket *patchwork* (Version 1.3.2; Pedersen, 2019) zur strukturierten Zusammenführung individueller Plots implementiert.

Als zweite Klasse werden probabilistische Funktionskurven extrahiert. Die grafische Repräsentation der item- und personentheoretischen Kennwerte aus den IRT-Modellierungen wird entweder direkt über die mathematischen Plot-Routinen des Pakets *mirt* (Version 1.46.1; Chalmers, 2012) gesteuert oder über extrahierte mathematische Stützpunkte in eine hochauflösende *ggplot2*-Umgebung überführt. Dies umfasst die Item-Informationskurven (IIC) zur Lokalisierung der messgenauen Fähigkeitsbereiche, die aggregierten Test-Informationskurven (TIC) zur Bestimmung der globalen Testpräzision sowie die Kategorie-Antwort-Kurven (Category Response Curves, CRC) zur visuellen Analyse der stufenweisen Lösungswahrscheinlichkeiten auf Itemebene.

Die dritte Klasse betrifft multivariate Matrix- und Relationsgrafiken, die zur Aufdeckung komplexer Kennwertstrukturen, mathematischer Abhängigkeiten sowie dimensionaler Diagnostiken herangezogen werden. Zur empirischen Fundierung der optimalen Faktoren- und Clusterzahl im Item- und Probandenraum werden hierbei mithilfe der Pakete *factoextra* (Version 2.0.0; Kassambara & Mundt, 2026) und NbClust (Version 3.0.1; Charrad et al., 2014) Scree-Plots zur Evaluation des Eigenwertverlaufs nach dem klassischen Ellbogen-Kriterium, die Gap-Statistik nach Tibshirani et al. zum Abgleich der clusterinternen Varianz mit einer simulierten Nullreferenzverteilung sowie die durchschnittliche Silhouettenbreite (Average Silhouette Width) als absoluter Validitätsindex der Zuordnungsgüte grafisch aufbereitet. Die Interkorrelationsmatrizen der mathematischen Teilbereiche werden als farbcodierte Heatmaps ausgegeben, während hierarchische Gruppierungen über Dendrogramme dargestellt werden. Die Abbildung von bedingten Fehlerwahrscheinlichkeiten und typischen Aufgabenkopplungen erfolgt über topologische Netzwerkgrafen, welche über die mathematischen Relationspakete igraph (Version 2.3.2;

Csárdi et al., 2006) und ggraph (Version 2.2.2; Pedersen, 2026a) parametrisiert werden, um die assoziativen Verbindungsstärken und Fehlernetzwerke räumlich zu visualisieren.

Eine methodische Ausnahme im Rahmen dieser Systematik bildet eine spezifische strukturelle Grafik, die aufgrund ihrer programmtechnischen Eigenheit direkt aus den latenten Modellstrukturen abgeleitet und nicht mit dem Paket *ggplot2* erzeugt wurde. Abweichend von den allgemeinen Prinzipien der grafischen Repräsentation wird die Darstellung exklusiv in ihrem entsprechenden Fachkapitel behandelt, um sie im unmittelbaren Kontext des empirischen Befundberichts textlich tiefergehend zu analysieren.

Zunächst wurden die Verteilungseigenschaften der relevanten Variablen überprüft, um die Wahl zwischen parametrischen und nicht-parametrischen Verfahren methodisch zu begründen.

3.3.1 Überprüfung der Normalverteilungsannahme

Die Prüfung auf Normalverteilung der metrischen Variablen (darunter die Testwerte der drei mathematischen Kompetenzbereiche sowie die Gesamtpunktzahlen) erfolgte zweistufig durch eine Kombination aus einer explorativen grafischen Analyse und einem formalen Signifikanztest:

1. Grafische Überprüfung: Zur visuellen Beurteilung der Verteilungsformen wurden für die Werte der Kompetenzbereiche sowie für die Gesamtergebnisse Histogramme erstellt. Anhand dieser konnte ein erster Eindruck bezüglich der Symmetrie, Modalität und eventueller Schiefe der Datenverteilung gewonnen werden.
2. Formale Überprüfung: Ergänzend zur grafischen Untersuchung wurde eine formale Überprüfung mittels des Shapiro-Wilk-Tests durchgeführt. Dieser Test prüft die Nullhypothese (H_0), dass die Daten in der Grundgesamtheit normalverteilt sind. Ein signifikantes Ergebnis ($p < 0{,}05$) führt zur Ablehnung der H_0 und deutet auf eine signifikante Abweichung von der Normalverteilung hin. Da der Shapiro-Wilk-Test jedoch bei größeren Stichproben hochgradig sensitiv auf selbst minimale, praktisch irrelevante Abweichungen von der Normalverteilung reagiert, ist die gemeinsame Betrachtung mit der grafischen Inspektion der Histogramme methodisch essenziell (vgl. Döring et al., 2016).

Da die Shapiro-Wilk-Tests und die Histogramme eine signifikante Abweichung von der Normalverteilung bestätigten, wurde für die nachfolgenden Analysen auf nicht-parametrische Verfahren zurückgegriffen.

3.3.2 Nicht-parametrische Verfahren zur Ergebnisprüfung und Ergebnisgrafik

Für den Vergleich der zentralen Tendenzen zwischen den abhängigen Stichproben (den drei verschiedenen Kompetenzbereichen innerhalb der gleichen Stichprobenstruktur) wurden folgende nicht-parametrische Verfahren herangezogen:

1. Friedman-Test: Um zu prüfen, ob systematische Leistungsunterschiede zwischen den drei Teilbereichen bestehen, wurde der globale Friedman-Test als nicht-parametrisches Äquivalent zur Varianzanalyse mit Messwiederholung herangezogen. Er prüft die Nullhypothese, dass die Verteilungen der Rangplätze der drei Kompetenzbereiche über eine Stichprobe hinweg identisch sind (Bortz & Schuster, 2010).
2. Wilcoxon-Vorzeichenrangtest (Wilcoxon signed-rank test): Im Falle eines signifikanten globalen Friedman-Tests wurden für die Beantwortung der spezifischen Fragestellungen paarweise post-hoc-Vergleiche mittels Wilcoxon-Vorzeichenrangtests durchgeführt. Zur Kontrolle der Alpha-Fehler-Kumulierung bei multiplen Vergleichen wurde die Bonferroni-Korrektur angewendet (Bortz & Schuster, 2010).
3. Grafische Veranschaulichung: Um die Ergebnisse der beiden vorausgehenden Tests, des Friedman- und des Wilcoxon-Test, visuell zusammenzufassen, wurden Boxplots erstellt. Diese Grafiken veranschaulichen die Leistungsunterschiede zwischen den Kompetenzen mittels ihrer Mediane, Quartile und Streubereiche, wodurch die inferenzstatistisch abgesicherten Ergebnisse deskriptiv sichtbar gemacht werden.

3.3.3 Zusammenhangsanalyse

Zur Überprüfung von ungerichteten Zusammenhängen zwischen den Bereichen wurde die Spearman-Rang-Korrelation (Spearman-Rho r_s) berechnet. Im Gegensatz zur Pearson-Produkt-Moment-Korrelation setzt dieses Verfahren keine Normalverteilung oder lineare Beziehung voraus, sondern basiert auf den Rängen der Datenpunkte.

Damit ist der Koeffizient robust gegenüber Ausreißern und eignet sich zur Analyse der vorliegenden nicht-normalverteilten Variablen. Die Interpretation der Effektstärke ρ erfolgt nach den gängigen Konventionen nach Cohen (1988):

- $|r_s| = 0{,}10$: schwacher Zusammenhang
- $|r_s| = 0{,}30$: mittlerer Zusammenhang
- $|r_s| = 0{,}50$: starker Zusammenhang

3.3.4 Geschlechtsspezifische Analysen

Zur Analyse möglicher geschlechtsspezifischen Unterschiede wurde ein mehrstufiges Verfahren angewendet:

1. Voraussetzungsprüfung: Für die inferenzstatistische Untersuchung der geschlechtsspezifischen Unterschiede wurde die Annahme der Normalverteilung für die relevanten Variablen getrennt nach Geschlecht überprüft. Hierbei wurde analog zu den Ausführungen in Kapitel 3.3.1 eine zweistufige Überprüfung mittels Histogrammen und Shapiro-Wilk-Tests durchgeführt.
2. Deskriptive Statistik: Die Kompetenzbereiche wurden zunächst getrennt nach Geschlecht deskriptiv beschrieben. Hierzu wurden das zentrale Lagemaß (der Median) sowie die Streuungsmaße (Spannweite und Interquartilsabstand [IQR])

der relevanten Variablen bestimmt und zur übersichtlichen Gegenüberstellung in Tabellenform aufgearbeitet.

3. Inferenzstatistischer Gruppenvergleich und Ergebnisveranschaulichung: Da die formalen Voraussetzungsprüfungen eine signifikante Verletzung der Normalverteilungsannahme in den Kompetenzbereichen und für das Gesamtergebnis zeigten (siehe Kapitel 4.1.3), wurde für den Gruppenvergleich ein nicht-parametrisches Verfahren gewählt. Die Prüfung auf signifikante Unterschiede zwischen den zentralen Tendenzen der unabhängigen Gruppen erfolgte mittels des Wilcoxon-Rangsummentests (auch bekannt als Mann-Whitney-U-Test). Mithilfe des R-Pakets *coin* (Version 1.4-3; Hothorn et al., 2005) wurden außerdem für die Wilcoxon-Rangsummentests die zugehörigen Effektstärken (*r*) berechnet. Zur visuellen Untermauerung und grafischen Präsentation der Testergebnisse wurden Boxplots erstellt, welche die Verteilungen der Testwerte von Jungen und Mädchen direkt gegenüberstellt (Bortz & Schuster, 2010).

3.3.5 Prüfung der Varianzhomogenität

Die Untersuchung der Varianzhomogenität erfolgte in zwei differenzierten Schritten, um sowohl den unabhängigen als auch den abhängigen Datenstrukturen der Untersuchung methodisch gerecht zu werden:

1. Geschlechtsspezifische Varianzhomogenität (unabhängige Stichprobe): Zur Überprüfung der Varianzgleichheit zwischen den beiden unabhängigen Gruppen (Jungen und Mädchen) wurde der Fligner-Killeen-Test (berechnet über das R-Standardpaket *stats*, Version 4.6.0) herangezogen. Da für die geschlechtsspezifischen Teilstichproben bereits Abweichungen von der Normalverteilungsannahme festgestellt wurden (siehe Kapitel 3.3.4), bietet dieses rangbasierte Verfahren eine robuste Alternative zu parametrischen Tests (wie dem Bartlett-Test), da es hochgradig unempfindlich gegenüber Abweichungen von der Normalverteilung und gegenüber Ausreißern ist. Ein nicht-signifikantes Testergebnis ($p \geq 0{,}05$) bestätigt die Annahme homogener Varianzen (Fligner & Killeen, 1976).
2. Varianzhomogenität zwischen den Teilbereichen (abhängige Stichproben): Da die verschiedenen Teilbereiche des Tests von denselben Lernenden bearbeitet wurden, liegen hier verbundene (bzw. abhängige) Messungen vor. Klassische Varianzhomogenitätstests für unabhängige Stichproben sind somit für diese spezifische Datenstruktur mathematisch unzulässig. Aus diesem Grund wurde die Gleichheit der Streuung zwischen den einzelnen Teilbereichen über einen Vergleich der Interquartilsabstände (IQR) evaluiert. Um eine valide Vergleichbarkeit der Teilbereiche zu gewährleisten, da diese auf unterschiedlichen Maximalpunktzahlen basieren, wurde nicht der absolute, sondern der relative Interquartilsabstand herangezogen. Dieser wurde berechnet, indem der absolute IQR des jeweiligen Kompetenzbereichs durch dessen maximal erreichbare Punktzahl dividiert wurde. Weiterhin kam der Wilcoxon-Vorzeichenrangtest auf Basis der Absolutabweichung der pro-

zentual erreichten Punkte vom jeweiligen Skalenmedium zum Einsatz. Dies ermöglicht eine fundierte, normierte und gegenüber Ausreißern resistente Einschätzung der Varianzhomogenität bei abhängigen, nicht-normalverteilten Variablen (Bortz & Schuster, 2010).

3.3.6 Extremgruppenvergleich

Im Anschluss erfolgt die Durchführung eines Extremgruppenvergleichs, bei dem die leistungsschwächsten 25 % (untere Leistungsgruppe) mit den leistungsstärksten 25 % (obere Leistungsgruppe) der Lernenden gegenübergestellt werden. Die Gruppenzuordnung erfolgt auf Basis der Quantile des erreichten Gesamtwerts im BRT-2.

Für den Vergleich der Leistungsprofile innerhalb der drei mathematischen Inhaltsbereiche wird bewusst auf eine inferenzstatistische Absicherung (z. B. mittels Wilcoxon-Rangsummentests) verzichtet. Da der Gesamtwert die mathematische Summe der einzelnen Teilbereiche darstellt, sind signifikante Unterschiede zwischen den Extremgruppen auf Ebene der Kompetenzbereiche durch das zirkuläre Selektionsdesign bereits mathematisch determiniert. Ein Hypothesentest besäße hier folglich keine empirische Aussagekraft. Der Extremgruppenvergleich wird daher rein deskriptiv über prozentuale Interquartilsabstände (relative IQRs) und Mediane sowie visuell über Boxplots ausgewertet, um die spezifischen Stärken- und Schwächenprofile im oberen und unteren Leistungsspektrum zu kontrastieren.

3.3.7 Methoden zur Bestimmung der Reliabilität und internen Konsistenz

Zur Überprüfung der Messgenauigkeit des verwendeten Bayreuther Rechentests (BRT-2) werden sowohl für die drei mathematischen Inhaltsbereiche als auch für die Gesamtskala die Reliabilitätskoeffizienten Cronbachs Alpha (α) und McDonalds Omega Total (ω_t) berechnet. Da es sich beim BRT-2 um ein multidimensionales Testverfahren handelt, welches hierarchisch aufgebaut ist, wird für die Gesamtskala zusätzlich McDonalds Omega Hierarchical (ω_h) bestimmt (Zinbarg et al., 2005).

Da der klassische Alpha-Koeffizient eine strikte, essenzielle Tau-Äquivalenz voraussetzt, reagiert er bei einer vorliegenden Linksschiefe und einer geringen Itemzahl mit einer systematischen Unterschätzung der wahren Reliabilität. Aus diesem Grund erfolgt die psychometrische Bewertung der Skalen, den Empfehlungen der modernen Testtheorie folgend (Field et al., 2012), primär auf Basis der robusteren, ladungsgewichteten Omega-Koeffizienten. Die Einordnung der Koeffizienten hinsichtlich der internen Konsistenz orientiert sich an den gängigen Richtlinien von George & Mallery, 2016):

- $\alpha\ bzw.\ \omega_t \geq 0{,}70$: akzeptabel (solide) bewertet
- $\alpha\ bzw.\ \omega_t \geq 0{,}80$: gut bewertet
- $\alpha\ bzw.\ \omega_t \geq 0{,}90$: exzellent bewertet

Für das hierarchische Omega gilt ein Schwellenwert von $\omega_h > 0{,}70$ als statistische Legitimation für die Bildung eines globalen Gesamtwerts, da ein Erreichen dieser Schwelle anzeigt, dass der übergeordnete Generalfaktor den Großteil der Testvarianz erklärt.

3.3.8 Analyse der Aufgabenschwierigkeit

Zur systematischen Beurteilung der empirischen Anforderungen der einzelnen Testitems des BRT-2 wird im ersten Schritt der klassische Schwierigkeitsindex (P_i) berechnet. Da das Testverfahren sowohl dichotome als auch polytome Aufgaben umfasst, wird ein erweitertes Berechnungsverfahren angewandt. Hierzu wird die empirisch ermittelte, mittlere Punktzahl der Stichprobe für jedes Skalenelement bestimmt und im Folgenden durch die maximal erreichbare Punktzahl der jeweiligen Aufgabe dividiert. Der resultierende, linear transformierte Wert liegt definitionsgemäß im Intervall zwischen 0 und 1.

Für die testtheoretische Interpretation dieses Kennwerts gilt das Prinzip der inversen Proportionalität. Je höher der berechnete Index ausfällt, desto geringer ist die mathematische Hürde und umso leichter wurde die Aufgabe von Lernenden gelöst. Werte nahe 1 indizieren demnach extrem leicht zu bewältigende Aufgaben, während Werte gegen 0 auf eine sehr hohe Aufgabenschwierigkeit im Leistungsquerschnitt hindeuten. Das Ziel dieser Analyse ist die Überprüfung, ob die Itemarchitektur eine differenzierte Erfassung des Fähigkeitsspektrums erlaubt.

3.3.9 Methoden zur Analyse bedingter Fehlerwahrscheinlichkeiten und Item-Netzwerke

Um die bedingten Wahrscheinlichkeiten bestimmen und mathematische Abhängigkeiten modellieren zu können, müssen die vorliegenden Datensätze zunächst dichotomisiert werden. Dazu wurde eine Aufgabe als „falsch“ klassifiziert, wenn weniger als 50 % der maximalen Punkte erzielt wurden. Andernfalls erfolgte die Einstufung als korrekt. Die Hürde von 50 % wurde hierbei bewusst gewählt, um einerseits das Risiko zu minimieren, dass eine Aufgabe aufgrund eines isolierten Leichtsinnsfehlers im Teilpunktebereich vorschnell als manifestiertes Unvermögen fehlinterpretiert wird und andererseits sicherzustellen, dass die Schwelle für ein unzureichendes Ergebnis testdiagnostisch hinreichend sensitiv ist.

Im Anschluss wurden die paarweise bedingten Wahrscheinlichkeiten berechnet und in Form einer quadratischen Fehlermatrix dargestellt. Dazu wurde das R-Paket *reshape2* (Version 1.4.5) genutzt. Für weiterführende topologische Analysen und zur Optimierung der visuellen Veranschaulichung wurden die bedingten Wahrscheinlichkeiten zusätzlich als gerichtete Netzwerkgraphen dargestellt. Hierbei wurde eine Relevanz-Schwelle auf 60 % für die Kantenplatzierung festgesetzt, um ausschließlich substanzielle informationelle Pfade im Graphen abzubilden und die Übersichtlichkeit beizubehalten. Für die Erstellung und algorithmische Anordnung der Netzwerkgraphen (unter Anwendung des

Fruchterman-Reingold-Layouts; Fruchterman & Reingold, 1991) wurden die R-Pakete *ggraph* (Version 2.2.2;) und *tidygraph* (Version 1.3.1; Pedersen, 2026b) genutzt.

3.3.10 Methodisches Vorgehen bei der Durchführung der Clusteranalyse

Nachdem die testtheoretischen Grundlagen der personenzentrierten Typenbildung in Kapitel 2.3.7 dargelegt wurden, beschreibt der vorliegende Abschnitt die konkrete softwaretechnische und algorithmische Umsetzung im Rahmen der Datenauswertung. Die Analysen wurden innerhalb der statistischen Programmierumgebung R durchgeführt.

3.3.10.1 Datenvorbereitung und Variablenkonditionierung

Für die Identifikation der mathematischen Leistungsprofile wurden die drei primären Kompetenzbereiche des Bayreuther Rechentests (BRT-2) als clusterrelevante Indikatorvariablen herangezogen:

1. Verständnis für natürliche Zahlen (VNZ)
2. Verständnis für das Stellenwertsystem (VSWS)
3. Verständnis für Rechenoperationen (VRO)

Da die Rohpunktwerte der drei Skalen unterschiedliche Maximalpunktzahlen aufweisen, wurden die Variablen für die Clusteranalyse in prozentuale Lösungswerte (Wertebereich von 0 % bis 100 %) transformiert. Durch diese identische Skalierung der mathematischen Kompetenzbereiche war eine vorab durchgeführte *z*-Standardisierung der Daten nicht erforderlich. Die inhärente Metrik und die direkte Vergleichbarkeit der Mittelwerte blieben somit vollständig erhalten.

Da der *k*-Means-Algorithmus hochgradig sensitiv gegenüber fehlenden Werten ist, wurde vorab eine listenweise Falllöschung (listwise deletion) über die R-Funktion `na.omit()` durchgeführt. Datensätzen mit unvollständigen Werten in den drei Zielskalen wurden somit restriktiv aus dem Clustering-Prozess ausgeschlossen, sodass eine bereinigte Stichprobe von $N = 574$ Lernenden in die Analyse einging.

3.3.10.2 Algorithmische Parameter und Ablaufstruktur

Die Partitionierung der Daten erfolgte über die Funktion `kmeans()` des R-Standardpakets *stats* (Version 4.6.0). Um dem mathematischen Problem der Konvergenz in einem lokalen Minimum (statt dem globalen Optimum) zu begegnen, wurde die Anzahl der zufälligen Startkonfigurationen der Zentroiden über das Argument `nstart = 25` definiert. Der Algorithmus berechnet hierbei 25 voneinander unabhängige Startlösungen und wählt autonom die Partition mit der geringsten Varianz innerhalb der Cluster aus. Zur Gewährleistung einer lückenlosen Replikation der exakten Clustergrenzen und der resultierenden Gruppenzugehörigkeiten wurde der Zufallsgenerator im Skript vorab via `set.seed(123)` fixiert (MacQueen et al., 1967).

Das empirische Prüfverfahren zur Bestimmung der finalen Clusteranzahl erfolgte sequenziell in drei Schritten unter Verwendung der Zusatzpakete *factoextra* (Version 2.0.0) und *NbClust* (Version 3.0.1):

1. Berechnung der mathematischen Validitätsindizes: Simultane Schätzung des WSS-Verlaufs, der durchschnittlichen Silhouettenbreite sowie der Gap-Statistik für einen Suchraum von $k = 1$ bis $k = 10$ Clustern.
2. Geometrische Inspektion: Extraktion der ersten beiden Hauptkomponenten über eine integrierte Hauptkomponentenanalyse (PCA), um die dreidimensionalen Clusterstrukturen für die visuelle Qualitätskontrolle auf eine zweidimensionale Ebene zu projizieren.
3. Vektorrückspielung: Nach der finalen Festlegung auf die 6-Cluster-Lösung wurde der resultierende Cluster-Vektor als kategorialer Faktor über die Zeilenindizes wieder in den Hauptdatensatz integriert, um eine fehlerfreie Zuordnung für die nachfolgenden deskriptiven Analysen im Ergebnisteil (Kapitel 4.3.1) zu gewährleisten.

3.3.11 Prozedurales Vorgehen zur empirischen Strukturaufklärung

Die Überprüfung der faktoriellen Validität und die dreidimensionale Strukturaufklärung des Bayreuther Rechentests (BRT-2) erfolgen über eine explorative Faktorenanalyse (EFA), gefolgt von einer konfirmatorischen Absicherung mittels konfirmatorischer Faktorenanalyse (CFA). Die mathematisch-testtheoretischen Grundlagen der Verfahren sind in Kapitel 2.3.8 verortet. Das konkrete statistische Vorgehen bei der Analyse des vorliegenden Datensatz ($N = 574$) gliedert sich in fünf sequenzielle Schritte:

1. Berechnung der robusten Korrelationsbasis: Aufgrund der Verteilungscharakteristika und potenzieller Decken-Effekte der 18 kriterienorientierten Testitems (vgl. Kapitel 2.3.1) wird als Analysebasis explizit eine Spearman-Rangkorrelationsmatrix berechnet. Dieses Vorgehen sichert die nachfolgenden Extraktionsschritte gegen Verzerrungen durch nicht-normale Itemverteilungen ab.
2. Formale Eignungsprüfung der Matrix: Vor der Faktorisierung wird die Datenmatrix simultan über zwei Kriterien auf ihre Faktorabilität geprüft. Beim Bartlett-Test auf Sphärizität wird ein hochsignifikantes Ergebnis ($p < 0{,}05$) erwartet, um eine Identitätsmatrix formal auszuschließen und das Vorliegen bedeutsamer Linearzusammenhänge nachzuweisen. Für die Annahme hinreichender Stichprobenangemessenheit beim Kaiser-Meyer-Olkin-Kriterium (KMO) wird ein globaler Mindestwert von Overall $MSA \geq 0{,}60$ festgelegt. Auf Einzelitem-Ebene gilt ein Schwellenwert von $MSA \geq 0{,}60$ als obligatorisches Ausschlusskriterium für defizitäre Variablen (Kaiser, 1974).
3. Protokoll zur Bestimmung der Faktorenzahl: Zur datengeleiteten Ermittlung der optimalen Faktorenzahl werden drei Abbruchkriterien parallel evaluiert und trianguliert. Zunächst wird das Kaiser-Guttman-Kriterium herangezogen, welches eine Vorauswahl aller Faktoren mit einem empirischen Eigenwert $> 1{,}0$ vorsieht. Wei-

terhin dient der Cattell-Scree-Plot der visuellen Identifikation des statistischen Elbogens (Knick-Kriterium), ergänzt durch die Parallelanalyse nach Horn. Im Falle widersprüchlicher Indikationen besitzt die Parallelanalyse systematische Priorität vor dem Kaiser-Guttman-Kriterium. Ergänzend gilt das strukturelle Postulat der Faktorisierung, nach dem Faktoren, die nach der Rotation durch weniger als drei Items substanziell geladen werden, aus Gründen der Skalenstabilität restriktiv reduziert werden (Horn, 1965).

4. Faktorenextraktion, Rotation und Ladungsevaluation: Die Extraktion der Faktoren erfolgt über die Minimum-Residual-Methode (MinRes), um eine methodische Deckungsgleichheit mit den empirischen Analysen zu gewährleisten (Harman & Jones, 1966). Entsprechend der fachdidaktisch begründeten Annahme korrelierter mathematischer Teilkompetenzen (vgl. Kapitel 2.3.4) wird standardmäßig ein oblikques (schräges) Rotationsverfahren (Promax) angewendet (Hendrickson & White, 1964).
   Für die finale Itemzuordnung und Skalenbereinigung gelten die folgenden Grenzwerte. Eine Variable wird einem Faktor zugeordnet, wenn ihre Primärladung in der Mustermatrix einen Betrag von $\geq 0{,}30$ erreicht.
   Eine Querladung liegt vor, wenn ein Item auf einem Nebenfaktor ebenfalls eine Ladung von $\geq 0{,}30$ aufweist, sofern die Ladungsdifferenz zum Hauptfaktor weniger als 0,15 beträgt (Howard, 2016).
5. Konfirmatorische Absicherung: Zur Prüfung der im Ergebnisteil (Kapitel 4.3.2) generierten EFA-Struktur wird eine konfirmatorische Faktorenanalyse (CFA) mittels des R-Pakets *lavaan* (Version 0.6-21; Rosseel, 2012) durchgeführt. Aufgrund des nicht-normalen Charakters der Daten wird die Schätzung über den robusten Maximum-Likelihood-Schätzer (MLR) vollzogen. Die Güte des Modells wird anhand der klassischen Fit-Indizes nach Hu & Bentler (1999) bewertet, wobei ein akzeptabler bis guter Modell-Fit über die Schwellenwerte $CFI \geq 0{,}90$ (bzw. restriktiv $\geq 0{,}95$), $TLI \geq 0{,}90$ (bzw. restriktiv $\geq 0{,}95$), $RMSEA \leq 0{,}06$ und $SRMR \leq 0{,}08$ definiert ist.

3.3.12 Prozedurales Vorgehen zur probabilistischen Testanalyse (IRT)

Zur itemgenauen Erfassung der Messpräzision und zur Evaluation der differenziellen Eignung des BRT-2 für den diagnostischen Zielbereich wird eine probabilistische Testanalyse im Rahmen der Item-Response-Theorie (IRT) durchgeführt. Die mathematisch-theoretischen Grundlagen der IRT-Modelle sind in Kapitel 2.3.6 verortet. Das konkrete prozedurale Vorgehen bei der Schätzung und Absicherung der Modelle umfasst die nachfolgend beschriebenen Spezifikationen.

3.3.12.1 Modellwahl für gestufte Antwortformate

Aufgrund des polytomen (gestuften) Antwortformats des Testinstruments wird die Parameterschätzung über ein Generalized Partial Credit Modell (GPCM) vollzogen (Muraki, 1992). Die Schätzung erfolgt separat für die drei in Kapitel 4.3.2 extrahierten Faktoren, um

der dimensionalen Binnenstruktur Rechnung zu tragen. Die Berechnungen werden computergestützt unter Verwendung des R-Pakets *mirt* (Version 1.46.1) durchgeführt.

3.3.12.2 Protokoll zur Überprüfung der IRT-Modellvoraussetzungen

Bevor die Item- und Testinformationskurven interpretiert werden können, wird die mathematische Modellkonformität anhand von vier sequenziellen Prüfschritten abgesichert:

1. Essenzielle Eindimensionalität: Der globale Modell-Fit pro Faktor wird über die limitierte Informations-Fit-Statistik M_2 (Typ: „C2“) evaluiert. Als Kriterien für eine akzeptable Passung der unifaktoriellen IRT-Struktur werden approximative Fit-Indizes herangezogen ($CFI \geq 0{,}90,\ TLI \geq 0{,}90,\ RMSEA \leq 0{,}06$).
2. Lokale Unabhängigkeit: Die Überprüfung auf unerwünschte lokale Itemabhängigkeiten erfolgt über Yen's Q_3-Statistik auf Basis der Residuenmatrix. Werte von $Q_3 > 0{,}20$ werden als kritischer Indikator für lokale Abhängigkeiten definiert, sofern sie ein positives Vorzeichen aufweisen (Yen, 1984).
3. Lokaler Item-Fit: Die Passung jedes einzelnen Items an die mathematischen Annahmen des GPCM wird über die $S\text{-}X^2$-Statistik von Orlando & Thissen (2000) geprüft. Ein nicht-signifikantes Testergebnis ($p > 0{,}05$) indiziert eine hinreichende Modellkonformität des Items.
4. Schwellenordnung und Monotonie: Die Funktionalität der Antwortkategorien wird über die visuelle Inspektion der Kategorie-Antwortkurven (Trace Lines) sowie über die numerische Monotonieprüfung der Schwellenparameter (b_1 bis b_4) in der IRT-Metrik kontrolliert.

3.3.12.3 Prozedur zur Evaluation von Testinformation und Reliabilität

Nach erfolgreicher Modellabsicherung werden die lokalen Informationserträge extrahiert, um die diagnostische Validität des Instruments zu bestimmen:

1. Item-Informationskurven (IIC): Die lokale Messpräzision jedes Items wird über den Verlauf der Funktion $I(\theta)$ evaluiert. Dabei wird analysiert, bei welchem latenten Fähigkeitsniveau θ das Maximum (Peak) des Informationsertrags liegt, um festzustellen, in welchem Leistungsbereich eine Aufgabe am schärfsten trennt.
2. Test-Informationskurven (TIC) und Standardmessfehler: Durch die additive Bündelung der Item-Informationen wird die Gesamttestinformation pro Faktor berechnet. Diese steht in einem direkten mathematischen Kehrwert-Verhältnis zum lokalen Standardfehler $SE(\theta)$. Es wird geprüft, ob das Maximum der TIC im intendierten negativen Fähigkeitsbereich ($\theta < 0$) liegt.
3. Empirische Reliabilität: Zur globalen Beschreibung der Messgenauigkeit der Skalen wird die marginale (empirische) Reliabilität berechnet, welche die Präzision über die gesamte Fähigkeitsverteilung der Stichprobe aggregiert (Chalmers, 2012).

3.3.13 Prozedurales Vorgehen zur Analyse ungleicher Aufgabenfunktionen

Die Überprüfung der geschlechtsspezifischen Messinvarianz des BRT-2 auf Itemebene folgt einem festgelegten statistischen Auswertungsprotokoll, welches softwaregestützt innerhalb der R-Umgebung umgesetzt wird.

3.3.13.1 Modellierung und Gruppenkonfiguration

Die Analyse des Differential Item Functioning (DIF) wird auf Basis der geschlechtsbereinigten Teilstichprobe von *N* = 467 Lernenden durchgeführt. Die statistische Modellierung erfolgt über die Funktion `DIF()` des R-Pakets *mirt* (Version 1.46.1). Als mathematische Schätzbasis dient das polytome Generalized Partial Credit Model (GPCM), welches separat für die drei etablierten Kompetenzdimensionen spezifiziert wird.

3.3.13.2 Inferenzstatistische Absicherung und Selektionskriterien

Zur Kontrolle der Alpha-Fehler-Inflation im Zuge der simultanen Itemprüfung werden die aus den Likelihood-Ratio-Tests resultierenden empirischen p-Werte über die R-Funktion `p.adjust()` nach dem Verfahren von Benjamini und Hochberg (1995) adjustiert.

Ein Testitem wird dann als in seiner Messinvarianz verletzt klassifiziert, wenn der korrespondierende, adjustierte Wahrscheinlichkeitswert das Signifikanzniveau von $\alpha = 0{,}05$ unterschreitet. Als ergänzende mathematische Kriterien zur Bewertung der globalen Informationshaltigkeit der Modellverschiebung werden für jedes Item das Akaike-Informationskriterium (AIC) sowie das Bayesianische Informationskriterium (BIC) in Form von Differenzwerten (ΔAIC und ΔBIC) dokumentiert (Burnham & Anderson, 2002).

3.3.13.3 Prozedur zur graphischen Richtungs- und Tiefenanalyse

Die qualitative Evaluation der statistisch auffälligen DIF-Effekte erfolgt standardisiert in zwei aufeinanderfolgenden Schritten:

1. Extraktion der erwarteten Aufgaben-Scores: Über das geschätzte Modell werden die fähigkeitsspezifischen erwarteten Aufgaben-Scores (Expected Item Scores) über das latente Kontinuum (θ) extrahiert und grafisch als geschlechtsspezifische Funktionsgraphen dargestellt. Dies dient der Identifikation von uniformen und nicht-uniformen Verlaufsmustern.
2. Extraktion der Kategorie-Wahrscheinlichkeiten: Für die detaillierte Lokalisierung der Effekte werden die Item-Kategorie-Antwortwahrscheinlichkeiten (Category Response Curves) ausgegeben. Diese erlauben eine visuelle Analyse der geschlechtsspezifischen Verschiebung direkt auf Ebene der einzelnen Antwortstufen.

3.3.14 Prozedurales Vorgehen zur Analyse der latenten Personenparameter

Die quantitative Auswertung auf Personenebene umfasst die softwaregestützte Extraktion sowie die anschließende inferenzstatistische Analyse der individuellen Parameterschätzungen.

3.3.14.1 Extraktion der EAP-Schätzwerte

Die Berechnung der latenten Personenfähigkeiten (θ-Werte) erfolgt als Expected A Posteriori (EAP)-Schätzwerte auf Basis des zuvor spezifizierten Generalized Partial Credit Models (GPCM). Die Extraktion wird dimensionsspezifisch für die drei mathematischen Faktoren des BRT-2 über den Aufruf der R-Funktion `fscores()` durchgeführt. Die generierten individuellen Metriken werden anschließend in einer zentralen Datenmatrix für die finale Verteilungsprüfungen zusammengeführt (Bock & Mislevy, 1982).

3.3.14.2 Inferenzstatistische Gruppen- und Dimensionsvergleiche

Die inferenzstatistische Auswertung der extrahierten Parameter gliedert sich in zwei prozedurale Schritte:

1. Interindividuelle Differenzprüfung: Der geschlechtsspezifische Gruppenvergleich wird für jeden der drei Faktoren separat über die R-Funktion `wilcox.test()` berechnet. Diese Prozedur wird auf Basis der geschlechtsbereinigten Teilstichprobe ($N = 467$) operationalisiert. Für jeden Faktor werden die Prüfgröße (z), der asymptotische Wahrscheinlichkeitswert (p) sowie die Effektstärke (r) dokumentiert, wobei letztere über die mathematische Transformation $r = \frac{|z|}{\sqrt{N}}$ bestimmt wird (Rosenthal, 1991).
2. Intraindividuelle Differenzprüfung: Der dimensionsübergreifende Vergleich der individuellen Fähigkeitsausprägungen über die drei Faktoren hinweg wird über die R-Funktion `friedman.test()` durchgeführt. Diese Prozedur wird parallel für die unbereinigte Gesamtstichprobe ($N = 574$) sowie für die geschlechtsbereinigte Teilstichprobe angewendet. Dokumentiert werden die empirische Chi-Quadrat-Prüfgröße (χ^2), die Freiheitsgrade (df) und der korrespondierende Signifikanzwert (p).

3.3.15 Prozedurales Vorgehen zur hierarchischen Bifaktor-Modellierung

Zur empirischen Überprüfung der Annahme einer essenziellen Eindimensionalität wird eine hierarchische Bifaktor-Modellierung auf Basis der Gesamtstichprobe ($N = 574$) durchgeführt. Die mathematische-testtheoretischen Grundlagen dieser Modellfamilie sind in Kapitel 2.3.8 verortet. Die prozedurale Umsetzung gliedert sich in fünf rein operative Analysephasen.

3.3.15.1 Modellspezifikation und Schätzalgorithmen

Die mathematische Modellierung wird computergestützt innerhalb der R-Umgebung unter Verwendung des Pakets *mirt* (Version 1.46.1) realisiert. Zur kriteriengeleiteten Evaluation werden fünf konkurrierende Strukturmodelle im Rahmen des Generalized Partial Credit Models (GPCM) definiert und geschätzt:

1. Unifaktorielles Basismodell (GPCM-1F): Fixierung aller 18 Testaufgaben auf einem einzigen, globalen latenten Fähigkeitsfaktor (θ).

2. Theoriegeleitetes Dreifaktormodell (GPCM-Theorie): Zuordnung der Items zu den drei mathematischen Kernbereichen gemäß der fachdidaktischen Primärstruktur.
3. Explorativ hergeleitetes Dreifaktormodell (GPCM-EFA): Datengetriebene Zuordnung der Items auf Basis der in Kapitel 4.3.2 extrahierten Faktorladungen.
4. Theoriegeleitetes Bifaktor-Modell (Bifaktor-Theorie): Orthogonale Struktur, bei der jedes Item simultan auf einem allgemeinen Generalfaktor (G) sowie seinem theoretischen spezifischen Gruppenfaktor (F1 bis F3) lädt.
5. Empirisch-exploratives Bifaktor-Modell (Bifaktor-EFA): Orthogonale Struktur unter Beibehaltung des datengetriebenen EFA-Zuordnungsschemas für die spezifischen Faktoren.

Die Parameterschätzung der Bifaktor-Modelle erfolgt über den Quasi-Monte Carlo Expectation-Maximization-Schätzer (QMCEM; Cai, 2010; Chalmers, 2012). Zur Überprüfung der numerischen Stabilität werden die extrahierten Ladungsarchitekturen im Anschluss einer konfirmatorischen Absicherung im Paket lavaan (Version 0.6-21) unterzogen.

3.3.15.2 Inferenzstatistischer Modellvergleich und Gütekriterien

Die Bestimmung des empirisch überlegenen Strukturmodells erfolgt über eine komparative Synopse globaler und inkrementeller Fit-Indizes. Da die Testdaten kategoriale Verteilungscharakteristika aufweisen, werden sämtliche Gütemaße einheitlich über die Limited-Information-M_2-Statistik (Typ „C2“) berechnet. Als Evaluationskriterien für eine gute Modellpassung gelten die Schwellenwerte $RMSEA \leq 0{,}06$, $TLI \geq 0{,}95$, $CFI \geq 0{,}95$. Für den Modellvergleich bezüglich der Balance aus Erklärungsgehalt und Modellsparsamkeit werden das Akaike-Informationskriterium (AIC) sowie das Bayesianische Informationskriterium (BIC) herangezogen, wobei das Modell mit dem absoluten Minimum im Informationsraum als überlegen klassifiziert wird.

3.3.15.3 Psychometrische Indizes zur Bestimmung der Eindimensionalität

Zur operationalen Bewertung, ob der BRT-2 trotz vorliegender Subskalen als essenziell eindimensional interpretiert werden darf, werden zwei spezifische Strukturkoeffizienten berechnet und evaluiert.

1. Explained Common Variance (ECV): Dieser Koeffizient bestimmt das Verhältnis der Quadratsumme der Faktorladungen des Generalfaktors zur Gesamtsumme der Ladungsquadrate aller extrahierten Faktoren. Als kritisches Entscheidungskriterium für das Vorliegen einer hinreichenden Generalfaktordominanz wird ein Schwellenwert von $ECV > 0{,}70$ definiert (vgl. Reise, 2012).
2. McDonalds hierarchisches Omega (ω_h): Im Rahmen der modellbasierten Reliabilitätsanalyse extrahiert dieser Koeffizient denjenigen Anteil der Gesamtvarianz der Testscores, der ausschließlich auf den Generalfaktor zurückzuführen ist. Für die messtechnische Legitimation eines globalen Gesamttestwerts wird gemäß psychometrischer Praxis eine kritische Schranke von $\omega_h > 0{,}70$ festgelegt (vgl. Rodriguez et al., 2016).

3.3.15.4 Visualisierung mittels struktureller Pfadanalyse

Zur anschaulichen Beurteilung der dimensionalen Architektur und der Variablenbeziehungen wird die hierarchische Faktorstruktur des BRT-2 grafisch in einem Pfaddiagramm veranschaulicht. Die softwaregestützte Generierung der Grafik erfolgte über das R-Paket *psych* (Version 2.6.5; Matsunaga, 2010; Revelle, 2026) unter Verwendung der Funktion `omega.diagram()`. Als mathematische Grundlage wird hierbei die Schmid-Leiman-Transformation berechnet, welche die explorativen Ladungsstrukturen der 18 Aufgaben in einen übergeordneten Generalfaktor (G) und die spezifischen Gruppenfaktoren überführt.

Um eine visuelle Überladung im Textbett zu verhindern, wird ein algorithmischer Filter appliziert, welcher Pfade mit standardisierten Faktorladungen von < 0,20 automatisch ausblendet. Über das Argument `simple = TRUE` wird eine Struktur erzwungen, bei welcher jede der 18 Aufgaben primär an ihren dominanten Hauptfaktor gebunden wird, während die Messfehler über `errors = FALSE` unterdrückt werden. Die Ausgabe erfolgt als hochauflösende Vektorgrafik (SVG) unter Einbindung der serifenlosen Schriftart *Aptos*.

3.3.15.5 Extraktion multidimensionaler Informationsfunktionen

Nach erfolgreicher Modellselektion werden die multidimensionalen Informationsstrukturen analysiert, um die lokale diagnostische Messpräzision über das Fähigkeitskontinuum hinweg zu bestimmen. Hierzu werden die summierten Test-Informationskurven (TIC) des Generalfaktors jenen der spezifischen Faktoren gegenübergestellt. Auf Itemebene erfolgt eine komparative Richtungsanalyse anhand zweier mathematischer Projektionsverfahren:

1. Bedingte Item-Information (Methode 1): Berechnung der punktuellen Item-Information auf dem Generalfaktor unter der mathematischen Restriktion, dass die latenten Ausprägungen der spezifischen Gruppenfaktoren exakt auf Null fixiert werden ($\theta_{F_i} = 0$).
2. Gerichtete multidimensionale Information (Methode 2): Berechnung des Informationsertrags unter Berücksichtigung der multidimensionalen Achsenrotation im gemeinsamen Koordinatenraum.

Darüber hinaus werden zur differenzierten Bestimmung der inkrementellen Messkraft die Item-Informationskurven (IIC) der spezifischen Gruppenfaktoren (F1 bis F3) separat extrahiert und evaluiert. Um die spezifische Eigeninformation dieser Inhaltsdimensionen unabhängig vom übergeordneten Gesamtkonstrukt zu analysieren, wird das latente Fähigkeitsniveau des Generalfaktors im Schätzprozess mathematisch auf Null fixiert ($\theta_G = 0$). Dieses Vorgehen ermöglicht auf kategorialer Ebene eine präzise Identifikation von Aufgaben, deren psychometrische Substanz maßgeblich durch die spezifische Restvarianz einer einzelnen Inhaltsfacette getragen wird und sich dem vollständigen Informationsübergang in den Generalfaktor entzieht.

4 Ergebnisse

Das vorliegende Kapitel widmet sich der systematischen Darstellung, Analyse und inferenzstatistischen Absicherung der empirischen Befunde, die im Rahmen der großflächigen Datenerhebung mit dem Bayreuther Rechentest (BRT-2) generiert wurden. Das primäre Erkenntnisinteresse liegt darin, das neu entwickelte Testinstrument sowohl auf mikroanalytischer Ebene der einzelnen Aufgaben als auch auf makroanalytischer Ebene der Gesamtskalen umfassend zu evaluieren, um dessen diagnostische Präzision und psychometrischen Güte für den sonderpädagogischen sowie grundschuldidaktischen Einsatzbereich empirisch abzusichern. Um dem Leser eine transparente, wissenschaftliche stringente Validierungskette zu präsentieren, folgt die Strukturierung der Ergebnisse einer konsequenten methodischen Progression, die sich schrittweise von elementaren, variablenorientierten Kennwerten hin zu komplexen, modellbasierten Testanalysen entwickelt.

Zu diesem Zweck gliedert sich die empirische Berichterstattung in drei logisch aufeinander aufbauende Kernbereiche. Den Ausgangspunkt bildet das nachfolgende Kapitel 4.1, welches sich auf die deskriptiven und inferenzstatistischen Leistungsanalysen der Stichprobe konzentriert. In diesem Abschnitt wird das Datenmaterial auf manifester Ebene der Testscores seziert, um über robuste, nicht-parametrische Prüfverfahren fünf zentrale forschungsleitende Fragen hinsichtlich intra- und interindividueller Leistungsunterschiede, Merkmalsinterkorrelationen sowie spezifischer Heterogenitäts- und Extremgruppenstrukturen der Lernenden zu beantworten. Darauf aufbauend werden in Kapitel 4.2 die klassischen testtheoretischen Parameter konsolidiert. Dieser Abschnitt fokussiert die Bestimmung der internen Konsistenz unter kritischer Reflexion der τ-Äquivalenz mittels Cronbachs α und McDonalds ω, die deskriptive Evaluation der Aufgabenschwierigkeiten sowie die mathematische Modellierung von bedingten Fehlerwahrscheinlichkeiten zur Aufdeckung typischer Fehlernetzwerke im Antwortverhalten.

Den testtheoretischen und mathematischen Schwerpunkt des gesamten Ergebnisteils markiert schließlich Kapitel 4.3. In diesem umfassenden Modellkapitel wird der theoretische Rahmen der Klassischen Testtheorie verlassen und das Datenmaterial über fortgeschrittene, multivariate und probabilistische Verfahren tiefgehend seziert. Die Analysen spannen sich dabei in fünf aufeinander bezogenen Säulen von strukturprüfenden Dimensionsanalysen (EFA und CFA) über die probabilistische Skalierung mittels des Generalized Partial Credit Models (IRT) bis hin zur mathematischen Überprüfung der geschlechtsspezifischen Messinvarianz (DIF). Den Abschluss dieses Komplexes bilden die Verteilungsanalysen der latenten Personenparameter sowie die orthogonale Varianzzerlegung im Rahmen eines Bifaktor-Modells, wodurch ein lückenlos abgesichertes empirisches Fundament für die anschließende fachdidaktische Diskussion der Arbeit generiert wird.

4.1 Deskriptive und inferenzstatistische Leistungsanalysen der Stichprobe

Das vorliegende Kapitel widmet sich der systematischen Aufbereitung und inferenzstatistischen Absicherung der manifestierten Leistungsdaten, die über die Rohwerte des Bayreuther Rechentests (BRT-2) in der Stichprobe generiert wurden. Bevor das Datenmaterial in den nachfolgenden Abschnitten über komplexe, modellbasierte Verfahren der probabilistischen Testtheorie auf latenter Ebene abstrahiert wird, erfordert es einer fundierten Durchleuchtung der tatsächlich, manifesten Testperformanz. Die unkonditionierte Analyse der Rohwertstrukturen bildet das empirische Fundament der gesamten Arbeit und dient der gezielten Beantwortung von fünf zentralen Forschungsfragen. Da die Verteilungscharakteristika von Schulleistungstests im sonderpädagogischen Diagnostikbereich häufig von der klassischen Normalverteilungsannahme abweichen, werden in diesem gesamten Abschnitt konsequent robuste, nicht-parametrische Prüfverfahren implementiert, um mathematisch valide und interpretationssichere Befunde zu garantieren.

Um dem Leser eine transparente und inhaltlich stringente Beweisführung zu präsentieren, folgt die Darstellung einer logischen Sequenz, die sich von globalen Dimensionsvergleichen über relationale Analysen bis hin zu detaillierten Heterogenitätsbetrachtungen entwickelt. Den Ausgangspunkt markiert hierbei Kapitel 4.1.1, welches sich den Leistungsunterschieden zwischen den drei mathematischen Teilbereichen zuwendet. Mittels eines Friedman-Tests und nachgeschalteter Post-Hoc-Wilcoxon-Test wird hierbei geprüft, ob sich systematische Niveaudifferenzen zwischen den Inhaltsbereichen sichern lassen. Daran anschließend untersucht Kapitel 4.1.2 die Interkorrelation der mathematischen Teilbereiche über Spearman-Rangkorrelationen, um das messtheoretische Zusammenspiel und die Kohäsion der Subskalen zu beleuchten.

Der Fokus wechselt in Kapitel 4.1.3 von der skaleninternen Perspektive hin zu interindividuellen Differenzen, indem potenzielle geschlechtsspezifische Leistungsunterschiede über Wilcoxon-Tests evaluiert und hinsichtlich ihrer praktischen Relevanz über Effektstärken kritisch gewürdigt werden. Da zentrale Lageparameter die diagnostisch bedeutsame Streuung innerhalb einer Stichprobe jedoch nur unvollständig abbilden können, widmet sich Kapitel 4.1.4 der expliziten Analyse der Leistungsstreuung und Heterogenität auf Basis von Interquartilsabständen. Den methodischen Abschluss dieses manifesten Leistungskapitels bildet schließlich in Kapitel 4.1.5 ein Extremgruppenvergleich zwischen dem oberen und unteren Leistungsviertel. Durch die kontrastierende Gegenüberstellung der stärksten und schwächsten 25 % der Lernenden werden verborgene Homogenitätsunterschiede und Min/Max-Strukturen offengelegt, die für die schulpraktische und kriterienorientierte Screening-Diagnostik von fundamentaler Bedeutung sind.

4.1.1 Leistungsunterschiede zwischen den mathematischen Teilbereichen

Zuallererst soll die Frage geklärt werden, ob die Lernenden in den verschiedenen mathematischen Teilbereichen unterschiedlich abschneiden. Um diese Fragestellung untersuchen zu können, wurde zu Beginn der Auswertung die Verteilung der erreichten Punktzahlen sowohl über die drei einzelnen Kompetenzbereiche als auch für den Gesamtbereich visuell geprüft (siehe Abbildung 1). In allen vier Histogrammen zeigt sich eine ausgeprägte Linksschiefe, was auf deutliche Decken-Effekte (Ceiling-Effekte) hinweist. Ein Großteil der $N = 574$ Lernenden erzielte sehr hohe Punktzahlen, während das untere Leistungsspektrum nur schwach besetzt ist.

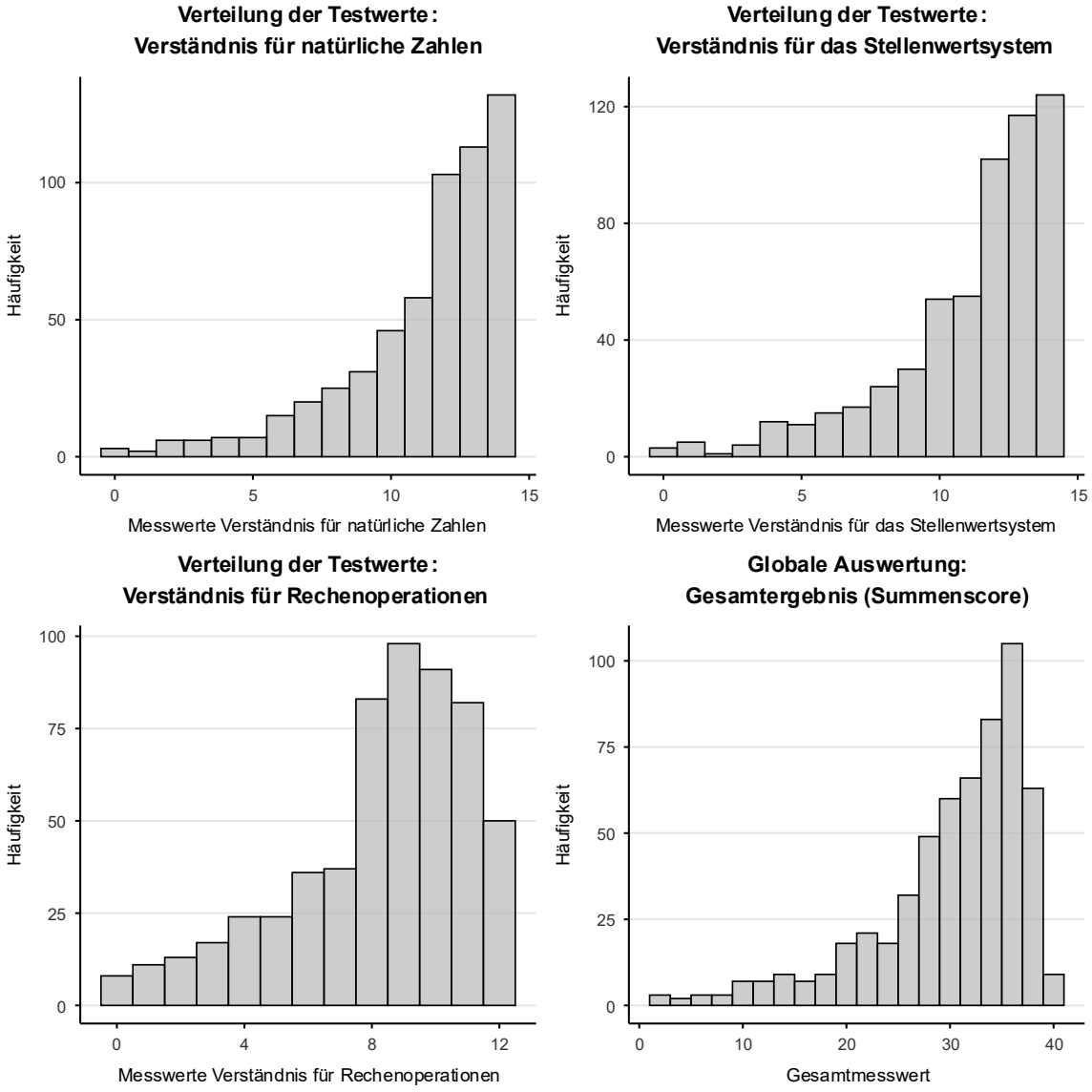


Abbildung 1 Häufigkeitsverteilungen der drei Kompetenzbereiche Verständnis für natürliche Zahlen, Verständnis für das Stellenwertsystem und Verständnis für Rechenoperationen des BRT-2, sowie die Verteilung der Gesamtpunktzahlen zur Veranschaulichung der Abweichung von der Normalverteilung (N = 574)

Bei den Verteilungen fällt auf, dass die Teilbereiche *Verständnis für natürliche Zahlen* und *Verständnis für das Stellenwertsystem* stark linksschief sind und die Mehrheit der Lernenden zwischen 12 und 14 Punkten erreicht. Niedrigere Punktzahlen kommen nur selten vor. Beim *Verständnis für Rechenoperationen* ist die Verteilung eher zentriert, weist

jedoch auch eine geringe Linksschiefe auf, wobei der Modus der Verteilung bei 9 Punkten liegt und die Ergebnisse deutlich stärker streuen als bei den anderen beiden Bereichen. Die Gesamtverteilung fasst die einzelnen Ergebnisse zusammen, wobei die meisten Lernenden zwischen 30 und 40 Punkten erreichen.

Diese visuelle Asymmetrie liefert die direkte mathematische Begründung dafür, warum die inferenzstatistischen Voraussetzungen für parametrische Verfahren verletzt sind. Die anschließenden Shapiro-Wilk-Tests bestätigen diesen Eindruck für alle drei Kompetenzbereiche hochsignifikant (*Verständnis für natürliche Zahlen*: $W = 0{,}83576$, $p < 2{,}2 \cdot 10^{-16}$; *Verständnis für das Stellenwertsystem*: $W = 0{,}83975$, $p < 2{,}2 \cdot 10^{-16}$; *Verständnis für Rechenoperationen*: $W = 0{,}91146$, $p < 2{,}2 \cdot 10^{-16}$) und ebenso für die Gesamtverteilung der Punkte der Lernenden ($W = 0{,}87231, p < 2{,}2 \cdot 10^{-16}$), weshalb im Folgenden konsequent auf nicht-parametrische Verfahren zurückgegriffen werden muss. Für den Vergleich der drei Kompetenzbereiche sind dies der Friedman-Test sowie als Post-hoc-Test der Wilcoxon-Test. Der Friedman-Test liefert hierbei hochsignifikante Unterschiede zwischen den drei Kompetenzbereichen ($\chi^2 = 591{,}36$, $df = 2$, $p < 2{,}2 \cdot 10^{-16}$). Da dieser Test jedoch keine Auskunft darüber gibt, zwischen welchen spezifischen Bereichen die Unterschiede vorliegen, wurde im Anschluss ein paarweiser Wilcoxon-Test durchgeführt. Die Ergebnisse dieses Tests (siehe Tabelle 1) zeigen hierbei, dass signifikante Unterschiede sowohl zwischen den Bereichen *Verständnis für natürliche Zahlen* und *Verständnis für Rechenoperationen*, als auch zwischen den Bereichen *Verständnis für das Stellenwertsystem* und *Verständnis für Rechenoperationen* vorliegen. Zwischen dem *Verständnis für natürliche Zahlen* und dem *Verständnis für das Stellenwertsystem* hingegen wurde kein signifikanter Unterschied festgestellt.

Tabelle 1 Ergebnisse des Wilcoxon-Tests. Für p-Werte kleiner 0,05 liefert er signifikante Ergebnisse. Es wurde die Bonferroni-Korrektur angewendet.

	Verständnis für natürliche Zahlen	Verständnis für Rechenoperationen
Verständnis für Rechenoperationen	$< 2 \cdot 10^{-16}$	
Verständnis für das Stellenwertsystem	1	$< 2 \cdot 10^{-16}$

Um zu untersuchen, wie sich der Unterschied zwischen den Bereichen auswirkt, wurden in Abbildung 2 die Boxplots zu den Kompetenzbereichen analysiert. Hierbei wurden die prozentual erreichten Punkte herangezogen, da der Bereich Verständnis für Rechenoperationen eine geringere Maximalpunktzahl aufweist als die anderen beiden Bereiche und somit die direkte Vergleichbarkeit hergestellt wird.

Anhand der Boxplots kann direkt abgelesen werden, dass der Bereich Verständnis für Rechenoperationen geringere Werte aufweist (schwächer ausgeprägt ist) als die anderen Bereiche, da hier der Median der prozentual erreichten Punkte bei einem Wert von 75 %

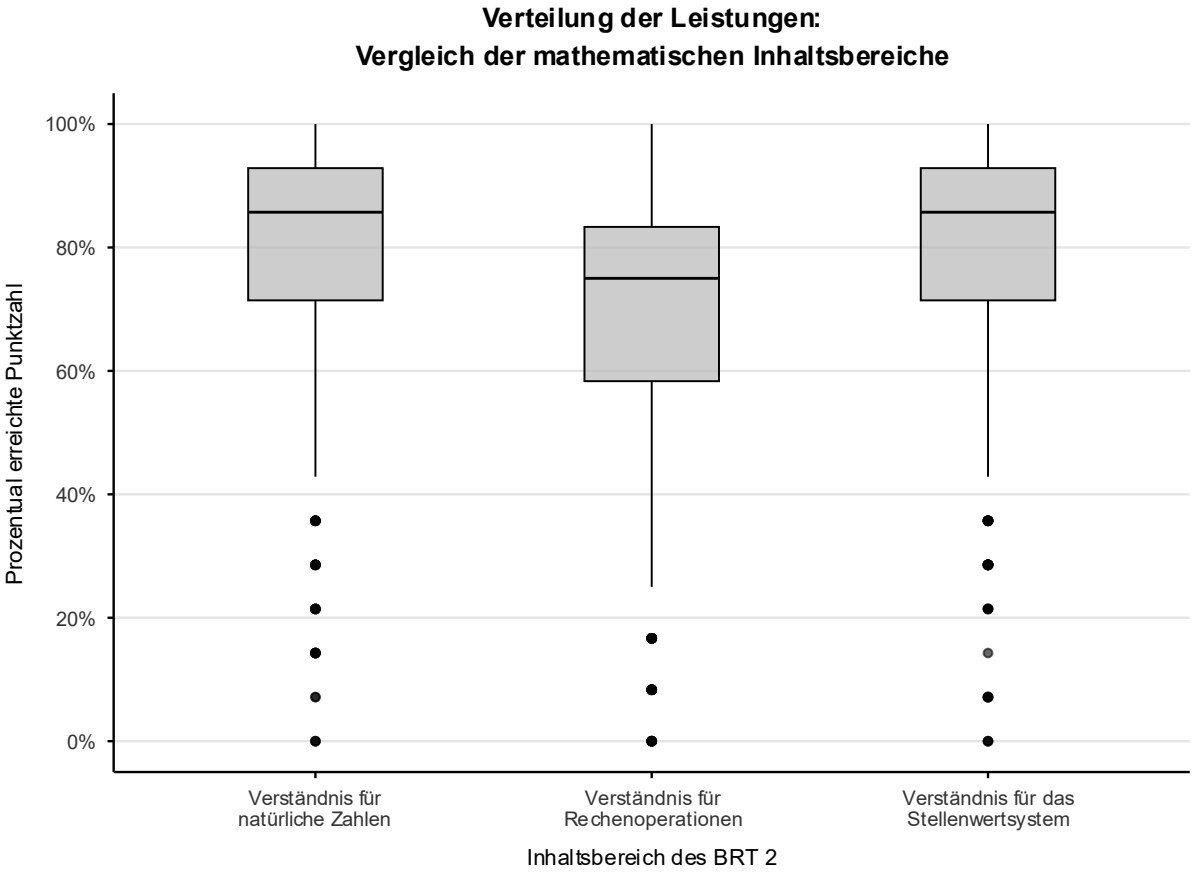


Abbildung 2 Boxplots der drei Kompetenzbereiche bezüglich der prozentual erreichten Punkte (N = 574).

der maximalen Punkte liegt, wohingegen die Mediane der anderen beiden Bereiche bei über 80 % der maximalen Punkte liegen. Hinsichtlich der absoluten Streuung unterscheiden sich die drei Teilbereiche nicht. Die genauen Punktwerte können der Tabelle 2 entnommen werden.

Weiterhin wurde die Effektstärke Kendall-W (*W*) berechnet. Hierbei ergibt sich ein Wert von $W = 0{,}514$, was für einen starken Effekt spricht. Dies deutet darauf hin, dass es sich hierbei um ein fundamentales Muster und nicht nur um einen unbedeutenden Nebeneffekt handelt.

Tabelle 2 Darstellung der wichtigsten Kennwerte für die drei Kompetenzbereiche und das Gesamtergebnis

	N	Median	IQR	Min	Max
Gesamtergebnis	574	33	8	1	40
Verständnis für natürliche Zahlen	574	12	3	0	14
Verständnis für das Stellenwertsystem	574	12	3	0	14
Verständnis für Rechenoperationen	574	9	3	0	12

Anmerkung. N = Stichprobengröße, IQR = Interquartilsabstand

4.1.2 Interkorrelation der mathematischen Teilbereiche

Zur Beantwortung der Fragestellung, inwiefern hohe Leistungen in einem der drei mathematischen Teilbereiche systematisch mit ebenso hohen Leistungen in den jeweils anderen Kompetenzbereichen einhergehen (Interkorrelation), werden nachfolgend die bivariaten Zusammenhänge analysiert. Hierzu wurden zunächst in Abbildung 3 die erreichten Punktzahlen der Lernenden von je zwei Teilbereichen gegeneinander aufgetragen. Um dem methodischen Problem der Überlagerung identischer Datenpunkte (sogenanntes Overplotting) bei diskreten Testwerten entgegenzuwirken, wurden die daraus resultierenden Punkte in Form einer gewichteten Darstellung visualisiert, welche die Häufigkeit gleicher Wertepaare grafisch berücksichtigt.

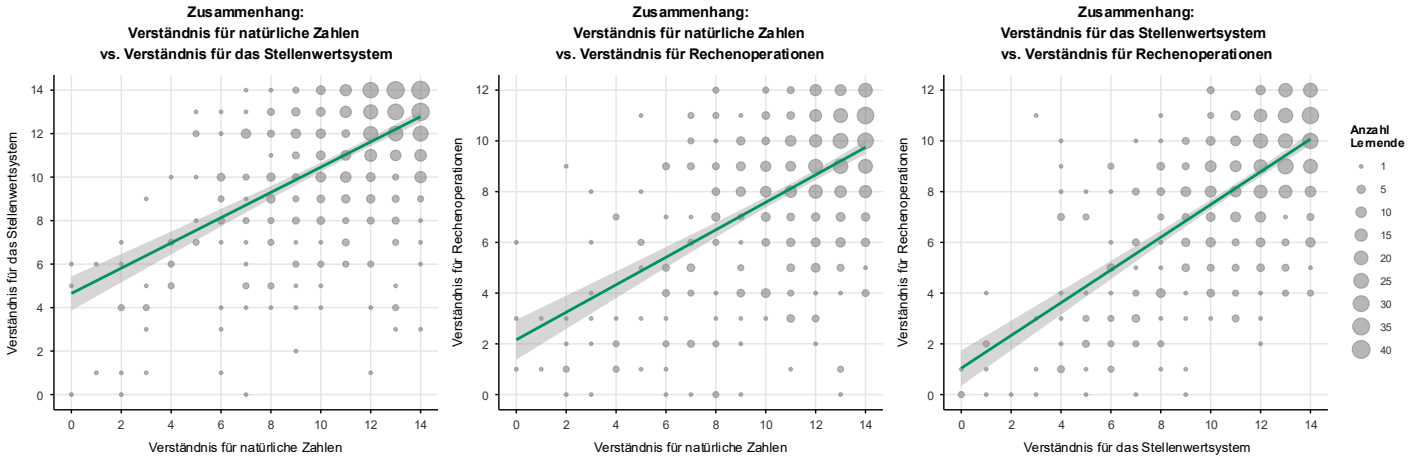


Abbildung 3 Darstellung der Zusammenhänge zwischen den Kompetenzbereichen. Mit der Regressionsgerade in grün und den Konfidenzbändern in grau dargestellt.

Anhand dieser grafischen Darstellung wird unmittelbar deutlich, dass ein positiver Zusammenhang zwischen den drei Bereichen besteht. Weiterhin ist zu erkennen, dass der stärkste visuelle Zusammenhang zwischen den Bereichen *Verständnis für das Stellenwertsystem* und *Verständnis für Rechenoperationen* besteht, da hier die Regressionsgerade die größte Steigung aufweist.

Dieser visuell explorierte Zusammenhang lässt sich auch rechnerisch durch die Bestimmung der bivariaten Korrelationen nachweisen. Aufgrund der in Kapitel 4.1.1 dokumentierten, ausgeprägten Linksschiefe der Daten und der damit einhergehenden Verletzung der Normalverteilungsannahme wird in diesem Fall auf das nicht-parametrische Verfahren der Spearman-Rang-Korrelation (r_s) zurückgegriffen (vgl. Kapitel 3.3). Dieses Verfahren erweist sich gegenüber den vorliegenden Decken-Effekten als robust, da es auf Rangplätzen statt auf den mathematischen Rohwerten operiert und somit monotone, nicht zwingend lineare Beziehungen präzise abbilden kann.

Die inferenzstatistische Überprüfung bestätigt die grafischen Trends vollumfänglich. Mit diesem Verfahren erhalten wir für die beiden Bereiche *Verständnis für natürliche Zahlen* und *Verständnis für das Stellenwertsystem* eine hochsignifikante Korrelation von $r_s \approx 0{,}448$ ($p < 2{,}2 \cdot 10^{-16}$), für die Bereiche *Verständnis für natürliche Zahlen* und *Ver-*

ständnis für Rechenoperationen erhalten wir eine Korrelation von $r_s \approx 0{,}469$ ($p < 2{,}2 \cdot 10^{-16}$) während für die Relation zwischen dem *Verständnis für das Stellenwertsystem* und *Verständnis für Rechenoperationen* eine Korrelation von $r_s \approx 0{,}568$ ($p < 2{,}2 \cdot 10^{-16}$) berechnet wird.

Somit lässt sich empirisch konstatieren, dass zwischen allen Bereichen statistisch signifikante, positive und gemessen an den gängigen sozialwissenschaftlichen Konventionen (Cohen, 1988) moderate bis mittlere Korrelationen vorliegen. Dies deutet darauf hin, dass die untersuchten mathematischen Konstrukte des BRT-2 zwar eigenständige Facetten abbilden, jedoch in einer engen wechselseitigen Abhängigkeit zueinanderstehen.

4.1.3 Geschlechtsspezifische Leistungsunterschiede

Um die Fragestellung zu klären, ob hinsichtlich der Gesamtleistung oder der Leistung in den drei Kompetenzbereichen geschlechterspezifische Unterschiede vorliegen, wurde auch hier zunächst überprüft, ob die Verteilungen der Punktzahlen in den drei Teilbereichen sowie der Gesamtpunktzahl bei den einzelnen Geschlechtern ebenfalls von der Normalverteilungsannahme abweichen. Hierzu wurden zunächst die Punktverteilungen in Abbildung 4 wieder als Histogramme dargestellt, um eine visuelle Prüfung vorzunehmen. Hierbei konnte festgestellt werden, dass die Verteilungen auch auf die Geschlechter getrennt weiterhin eine ausgeprägte Linksschiefe aufweisen. Diese visuell festgestellte Abweichung von der Normalverteilung wurde durch anschließend durchgeführte Shapiro-Wilk-Tests für alle Verteilungen anhand signifikanter Ergebnisse bestätigt (für die vollständigen Testwerte siehe Tabelle A in Anhang A). Die Shapiro-Wilk-Tests wurden hierbei nur für die geschlechterspezifischen Verteilungen berechnet, da ausschließlich diese für die weiterführenden Analysen genutzt wurden; für die Verteilungen der Lernenden ohne Geschlechterangabe wurde kein Test durchgeführt. Aufgrund der signifikanten Abweichung von der Normalverteilung wurde für die Untersuchung der geschlechtsspezifischen Leistungsunterschiede konsequent auf nicht-parametrische Verfahren zurückgegriffen.

Anhand der Wilcoxon-Rangsummentests konnte festgestellt werden, dass sowohl bei der Gesamtleistung als auch bei den Leistungen in den beiden Kompetenzbereichen *Verständnis für das Stellenwertsystem* und *Verständnis für Rechenoperationen* signifikante Unterschiede zwischen den Geschlechtern vorliegen. Die genauen Werte können Tabelle 3 entnommen werden.

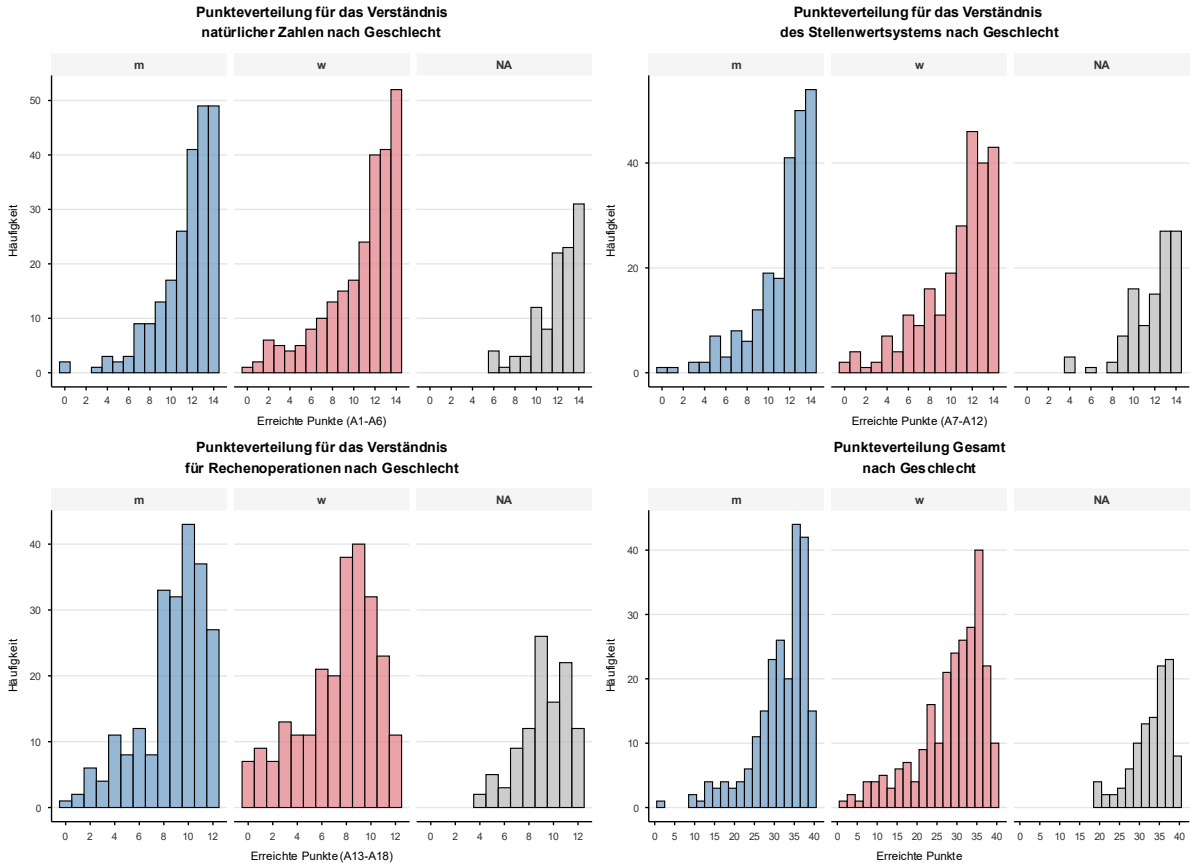


Abbildung 4 Histogramme der erreichten Punkte in den drei Kompetenzbereichen und in der Gesamtverteilung differenziert nach Geschlechtern. Die Jungen (N = 224) sind in Blau, die Mädchen (N = 243) sind in Rot und die Lernenden ohne Angabe zum Geschlecht (N = 107) sind in Grau dargestellt.

Tabelle 3 Ergebnisse der Wilcoxon-Rangsummentests zum Leistungsvergleich zwischen Jungen und Mädchen

	Median Jungen	*IQR* Jungen	Median Mädchen	*IQR* Mädchen	*W*	*p*	*r*
Gesamtergebnis	34	8	31	11	33045	$6{,}143 \cdot 10^{-5}$	0,185
Verständnis für natürliche Zahlen	12	3	12	4	29917	0,06062	0,0868
Verständnis für das Stellenwertsystem	12	3	12	4	31390	0,003731	0,134
Verständnis für Rechenoperationen	9	3	8	4	34460	$5{,}479 \cdot 10^{-7}$	0,232

Anmerkung. IQR = Interquartilsabstand, W = Testwertstatistik, p = p-Wert, r = Effektstärke. Die Stichprobengröße der Jungen beträgt N = 224 und die Stichprobengröße der Mädchen beträgt N = 243 und das effektive Signifikanzniveau ist auf α = 0,05 gesetzt.

Anhand der Ergebnisse bezüglich der Gesamtleistung kann festgestellt werden, dass Jungen einen höheren Median erzielen als Mädchen und somit im Gesamttest besser abschneiden. Anhand von Tabelle 3 und Abbildung 5 ist zudem erkennbar, dass die zentralen Leistungsbereiche der Mädchen eine höhere Heterogenität aufweisen als die der Jungen.

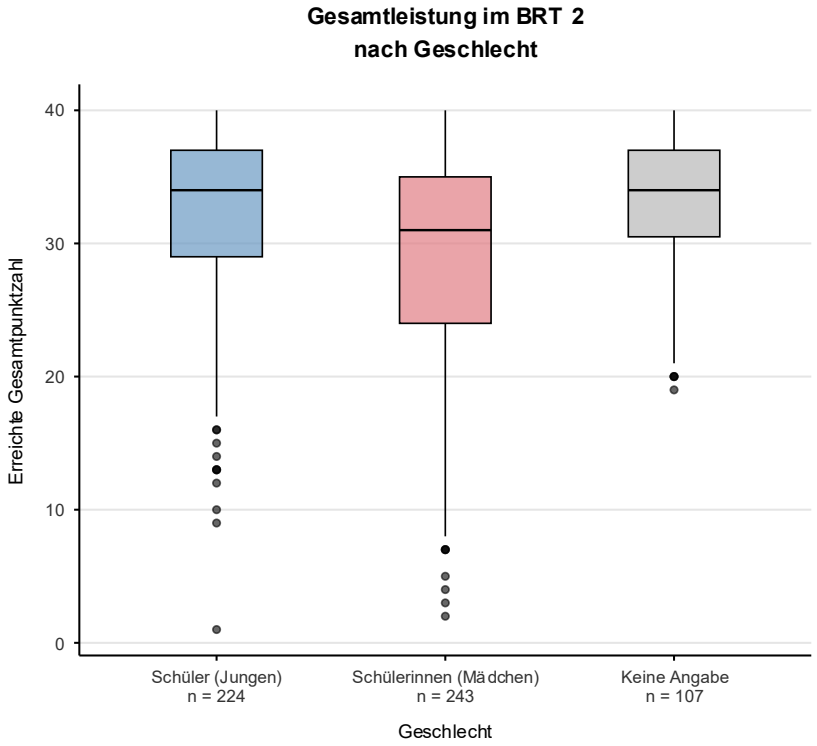


Abbildung 5 Darstellung der Gesamtergebnisse der Lernenden nach Geschlecht

Es ist jedoch hervorzuheben, dass es sich hierbei zwar um einen hochsignifikanten Unterschied zwischen den Geschlechtern handelt, die zugehörige Effektstärke nach Cohen (1988) mit $r = 0{,}185$ (siehe Tabelle 3) jedoch gering ausfällt.

Auch in dem Bereich Verständnis für Rechenoperationen gibt es einen hochsignifikanten Unterschied zwischen den weiblichen und männlichen Lernenden. Die Jungen weisen im Median einen Punkt mehr auf als die Mädchen, wobei bei den Mädchen analog dazu der Interquartilsabstand größer ausfällt (Tabelle 3). Dies lässt sich deskriptiv anhand der entsprechenden Boxplots in Abbildung 6 nachvollziehen, da hier die Box sowie die Whiskers der Mädchen eine deutlich größere visuelle Spanne zeigen. In diesem Bereich zeigt sich mit $r = 0{,}232$ auch die größte Effektstärke, was nach Cohen (1988) einem kleinen bis mittleren Effekt entspricht.

Obwohl die Mediane der beiden Geschlechter im Bereich *Verständnis für das Stellenwertsystem* identisch sind (jeweils $\tilde{x} = 12$), zeigt der Wilcoxon-Test einen signifikanten Unterschied ($p = 0{,}003731$). Dies lässt sich auf die Arbeitsweise des Tests als Rangsummentest zurückführen. Die Rangverteilung der Jungen fällt systematisch günstiger aus, da sie im unteren Leistungsbereich seltener vertreten sind als die Mädchen, deren erstes Quartil hier tiefer liegt.

Im Bereich *Verständnis für die natürlichen Zahlen* konnte kein signifikanter Geschlechtsunterschied festgestellt werden ($p = 0{,}06062$). Die Mediane der Mädchen und der Jungen sind hier identisch, jedoch ist der Interquartilsabstand bei den Mädchen etwas größer.

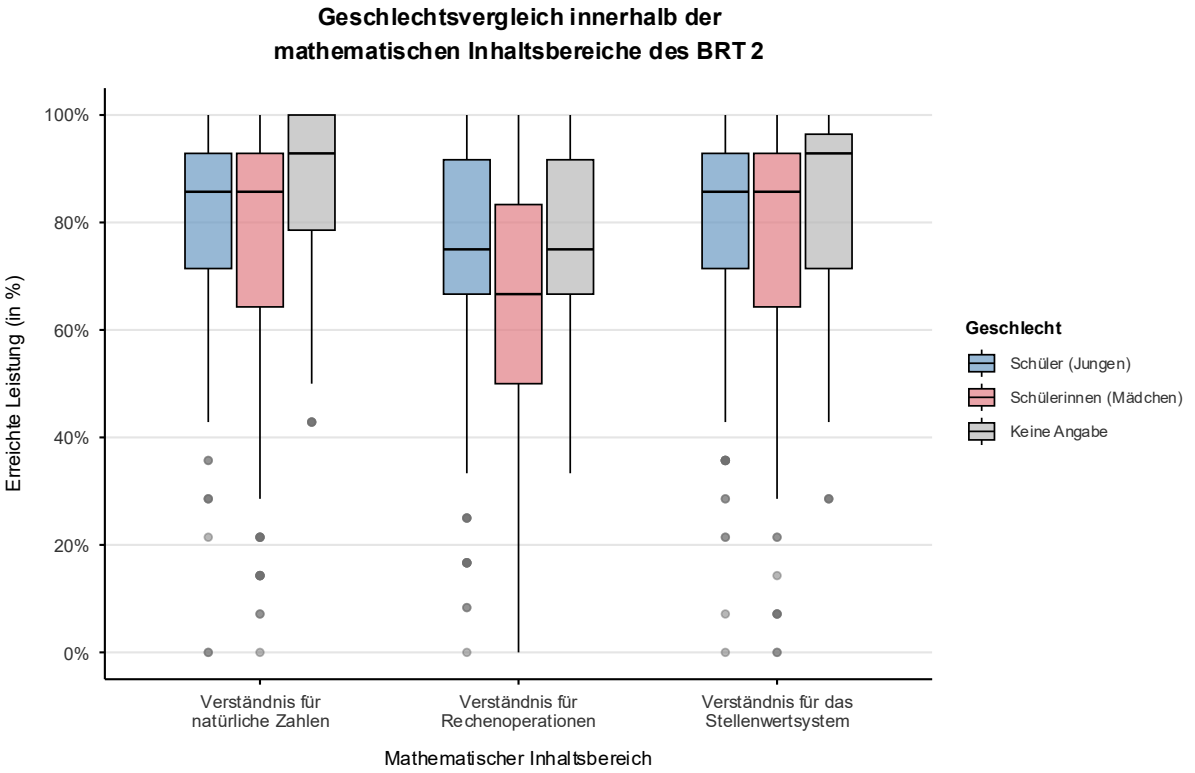


Abbildung 6 Boxplots zum Geschlechtsvergleich innerhalb der mathematischen Inhaltsbereiche des BRT-2

Zusammenfassend lässt sich festhalten, dass Jungen im Gesamtergebnis sowie in zwei der drei Teilbereiche signifikant höhere Leistungstendenzen aufweisen, wenngleich die Effektstärken klein ausfallen. Auffällig ist jedoch bereits in der deskriptiven Betrachtung (vgl. Tabelle 3), dass die Interquartilsabstände (IQR) der Mädchen über alle Skalen hinweg systematisch höher ausfallen als die der Jungen. Dies deutet darauf hin, dass Mädchen eine größere Inhomogenität in ihren mathematischen Leistungen aufweisen. In welchem Maße diese Untersuchungen belastbar sind, wird im nachfolgenden Kapitel 4.1.4 explizit geprüft.

4.1.4 Analyse der Leistungsstreuung und Heterogenität

Mediane zeigen nur die halbe Wahrheit, deshalb ist eine andere Frage für die Schulpraxis viel interessanter. Wie weit klafft die Schere in den Klassen auseinander?

Diese Leistungsprüfung wird auf zwei Ebenen analysiert. In Kapitel 4.1.4.1 erfolgt die inferenzstatistische Überprüfung der geschlechtsspezifischen Varianz, um die deskriptiv beobachteten Unterschiede der Streuung abzusichern. Danach, in Kapitel 4.1.4.2 richtet sich der Blick auf den direkten Vergleich der mathematischen Inhaltsbereiche untereinander. Da die Skalen ungleiche Punktzahlen besitzen, erfolgt dieser Vergleich über prozentual transformierte Werte, um Verzerrungen auszuschließen.

4.1.4.1 Geschlechtsspezifische Unterschiede in der Leistungsstreuung

Nachdem in Kapitel 4.1.3 anhand der absoluten Interquartilsabstände (vgl. Tabelle 3) sowie durch die visuelle Inspektion der Boxplots (vgl. Abbildung 5 und Abbildung 6) deskriptiv eine höhere Streuung der Schülerinnen im Gesamtergebnis sowie in allen drei Inhalts-

bereichen beobachtet wurde, erfolgt an dieser Stelle die inferenzstatistische Absicherung. Aufgrund der starken Linksschiefe der vorliegenden Daten des BRT-2 wird hierzu der nicht-parametrische Fligner-Killeen-Test herangezogen, welcher robust gegenüber Abweichungen von der Normalverteilungsannahme ist.

Für das Gesamtergebnis des BRT-2 bestätigt der Test einen signifikanten Unterschied in der Leistungsstreuung zwischen den Geschlechtern ($\chi^2 = 9.62, df = 1, p = 0{,}002$). Die in der Stichprobe beobachtete, höhere mathematische Streuung der Mädchen ist somit statistisch belastbar, sodass von einer vorliegenden Varianzheterogenität auszugehen ist.

Ein differenzierter Blick auf die drei mathematischen Inhaltsbereiche zeigt jedoch ein uneinheitliches Bild. Im Teilbereich *Verständnis für natürliche Zahlen* erwies sich der Unterschied in der Leistungsstreuung zwischen Jungen und Mädchen ebenfalls als hochsignifikant ($\chi^2 = 7{,}54, df = 1, p = 0{,}006$). Die deskriptiv größere Streuung der Schülerinnen (*IQR* = 4) im Vergleich zu den Schülern (*IQR* = 3) ist somit nicht auf zufällige Schwankungen zurückzuführen.

Für den Bereich *Verständnis für das Stellenwertsystem* konnte hingegen kein statistisch signifikanter Unterschied nachgewiesen werden ($\chi^2 = 2{,}84, df = 1, p = 0{,}092$). Trotz der numerischen Differenz in den absoluten Interquartilsabständen (Mädchen: *IQR* = 4; Jungen, *IQR* = 3) ist hier formell von einer homogenen Varianz zwischen den Geschlechtern auszugehen.

Im Kompetenzbereich *Verständnis für Rechenoperationen* verfehlt der Unterschied das kritische Signifikanzniveau von $\alpha = 0{,}05$ knapp ($\chi^2 = 3{,}57, df = 1, p = 0{,}059$). Auch für diesen Inhaltsbereich wird somit statistisch die Nullhypothese einer homogenen Varianz beibehalten.

Zusammenfassend lässt sich festhalten, dass sich das Bild einer systematisch größeren Leistungsstreuung der Schülerinnen nicht über alle Skalen hinweg inferenzstatistisch absichern lässt. Während Mädchen im Gesamtergebnis sowie im Bereich des *Verständnisses für natürlichen Zahlen* eine nachweislich größere Heterogenität in ihren Testleistungen aufweisen, sind die Varianzen beim Stellenwertsystem und den Rechenoperationen statistisch als homogen zu betrachten.

4.1.4.2 Vergleich der Leistungsstreuung zwischen den Inhaltsbereichen

Neben den geschlechtsspezifischen Differenzen ist ebenfalls von wissenschaftlichem Interesse, ob sich die Leistungsstreuungen zwischen den drei Kompetenzbereichen (*Verständnis für natürliche Zahlen*, *Verständnis für das Stellenwertsystem* und *Verständnis für Rechenoperationen*) innerhalb der Gesamtstichprobe unterscheidet. Da der Bereich *Verständnis für Rechenoperationen* eine geringere Maximalpunktzahl (12 Punkte) als die anderen beiden Bereiche (jeweils 14 Punkte) aufweist, ist ein direkter Vergleich der absoluten Interquartilsabstände methodisch nicht erlaubt. Um eine Verzerrung durch unterschiedliche Maximalpunktzahlen zu vermeiden, wurden die Rohpunkte in prozentuale

Tabelle 4 Deskriptive Statistiken der prozentualen Punktzahlen nach mathematischen Inhaltsbereichen sowie dem Gesamtergebnis (Gesamtstichprobe, N = 574).

	Maximalpunkte	Median (%)	*IQR* (%)	Min (%)	Max (%)
Gesamtergebnis	40	82,50 %	20,00 %	2,5 %	100 %
Verständnis für natürliche Zahlen	14	85,71 %	21,43 %	0 %	100 %
Verständnis für das Stellenwertsystem	14	85,71%	21,43 %	0 %	100 %
Verständnis für Rechenoperationen	12	75,00 %	25,00 %	0 %	100 %

Anmerkung. Die prozentualen Werte beziehen sich auf den jeweiligen Anteil der erreichten Punktzahl an der maximal möglichen Punktzahl des Teilbereichs.

Werte transformiert. Die resultierenden relativen Kennwerte für die Gesamtstichprobe sind in Tabelle 4 dargestellt.

Der systematische Vergleich der relativen Interquartilsabstände offenbart ein weitgehend stabiles Bild der Leistungsstreuung mit einer feinen Nuancierung zwischen den Skalen. Während die absoluten Streuungsmaße über alle drei Kompetenzbereiche hinweg vollkommen identisch ausfallen (*IQR* = 3; vgl. Tabelle 2), zeigt die relative Betrachtung, dass der Inhaltsbereich *Verständnis für Rechenoperationen* mit einem prozentualen *IQR* von 25 % die größte Leistungsstreuung aufweist. In diesem spezifischen Kompetenzbereich klafft die Schere zwischen mathematisch leistungsstarken und -schwachen Kindern in der zweiten Jahrgangsstufe somit relativ gesehen am weitesten auseinander.

Die beiden verbleibenden Inhaltsbereiche *Verständnis für natürliche Zahlen* und *Verständnis für das Stellenwertsystem* präsentieren sich hingegen als vollkommen deckungsgleich. Sowohl das *Verständnis für natürliche Zahlen*, wie auch das *Verständnis für das Stellenwertverständnis* weisen einen relativen Interquartilsabstand von 21,43 % auf und fallen somit geringfügig homogener aus.

Die paarweisen Vergleiche der absoluten Abweichung mittels Wilcoxon-Vorzeichenrangtests ergaben keinerlei signifikante Unterschiede in den Abweichungen zwischen dem *Verständnis für natürliche Zahlen*, *Verständnis für das Stellenwertsystem* und *Verständnis für Rechenoperationen*. Hierbei lagen nicht alle p-Werte über dem Niveau von 0,05 (VNZ vs. VSWS $p = 0{,}895$; VSWS vs. VRO $p < 0{,}001$; VNZ vs. VRO $p < 0{,}001$). Die hochsignifikante Abweichung unter Einbezug des *Verständnisses für Rechenoperationen* ist primär als methodisches Artefakt der veränderten Skalenmetrik zu interpretieren, da dieser Kompetenzbereich eine um zwei Punkte geringere Maximalpunktzahl aufweist, wodurch nach der prozentualen Transformation ungleiche diskrete Schrittweiten erhalten werden. Aufgrund der hohen statistischen Power der vorliegenden Stichprobe reagiert der rangbasierte Vorzeichentest hochgradig sensitiv auf die reinen gitterbedingten Unterschiede. Da die inhaltlich und strukturell direkt vergleichbaren Dimensionen (VNZ vs. VSWS) eine fundamentale Streuungsgleichheit aufweisen, wird die Annahme hinreichend homogener Varianzstrukturen für die nachfolgenden Analysen beibehalten.

Zusammenfassend lässt sich festhalten, dass die relative Bandbreite der Leistungen der Lernenden über die verschiedenen Inhaltsbereiche des BRT-2 hinweg ein hohes Maß an Konstanz aufweist. Die leicht erhöhte relative Varianz im Bereich des Verständnisses für Rechenoperationen resultiert primär aus der kompakten Testarchitektur dieses Teilbereichs, auf der sich individuelle Punktabweichungen prozentual stärker niederschlagen.

4.1.5 Extremgruppenvergleich – Oberes vs. unteres Leistungsviertel

Um die mathematischen Kompetenzprofile in den extremen Leistungsspektren detaillierter zu analysieren, wurden die Teilstichproben der leistungsstärksten 25 % (obere Leistungsgruppe, N = 177) und der leistungsschwächsten 25 % (untere Leistungsgruppe, N = 164) gegenübergestellt. Die Zuweisung der Gruppen erfolgte über die Quartile des erreichten Gesamtwerts beim BRT-2. Wie in Kapitel 3.3.6 methodisch begründet, erfolgt diese Kontrastierung innerhalb der Kompetenzbereiche rein deskriptiv, da inferenzstatistische Tests aufgrund des zirkulären Gruppendesigns keine empirische Aussagekraft aufweisen würde. Die visuelle Darstellung der prozentualen Leistungswerte innerhalb der drei Bereiche ist in Abbildung 7 dargestellt, die dazugehörigen deskriptiven Werte können Tabelle 5 entnommen werden.

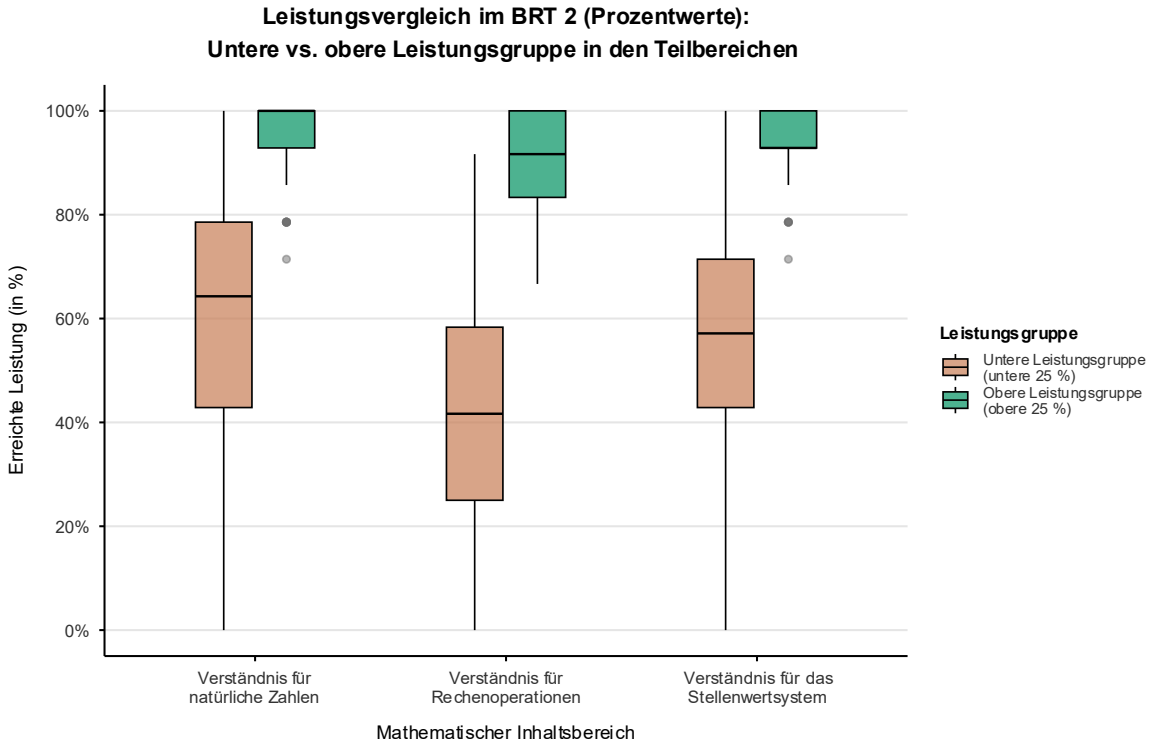


Abbildung 7 Leistungsvergleich im BRT-2 (Prozentwerte) zwischen der unteren und der oberen Leistungsgruppe innerhalb der drei Kompetenzbereiche

Tabelle 5 Deskriptive Kennwerte der Teilleistungen im Extremgruppenvergleich

	Maximalpunkte	Median	IQR	Min	Max
Untere Leistungsgruppe					
Verständnis für natürliche Zahlen	14	9 (64,3 %)	5 (35,7 %)	0 (0,0 %)	14 (100 %)
Verständnis für das Stellenwertsystem	14	8 (57,1 %)	4 (28,6 %)	0 (0,0 %)	14 (100 %)
Verständnis für Rechenoperationen	12	5 (41,7 %)	4 (33,3 %)	0 (0,0 %)	11 (91,7 %)
Obere Leistungsgruppe					
Verständnis für natürliche Zahlen	14	14 (100 %)	1 (7,1 %)	10 (71,4 %)	14 (100 %)
Verständnis für das Stellenwertsystem	14	13 (92,9 %)	1 (7,1 %)	10 (71,4 %)	14 (100 %)
Verständnis für Rechenoperationen	12	11 (91,7 %)	2 (16,7 %)	8 (66,7 %)	12 (100 %)

Anmerkung. Die prozentualen Werte in Klammern geben den relativen Anteil an der maximal erreichbaren Punktzahl des jeweiligen Kompetenzbereichs an.

Wie aus Abbildung 7 ersichtlich wird, zeigen sich fundamentale Unterschiede in den Verteilungen der beiden Extremgruppen, die über alle drei Kompetenzbereiche hinweg konsistent mathematische Diskrepanzen offenbaren.

Die obere Leistungsgruppe (obere 25 %) zeichnet sich durch einen ausgeprägten Deckeneffekt aus. In den Bereichen *Verständnis für natürliche Zahlen* und *Verständnis für das Stellenwertsystem* weisen die Lernenden homogene Höchstleistungen auf. Das drückt sich in dem geringen Interquartilsabstand von je 7,1 % (1 Punkt) und den hohen Medianen von 100 % beim *Verständnis für natürliche Zahlen* und 92,9 % (13 Punkte) beim *Verständnis für das Stellenwertsystem* aus. Einzig im Bereich *Verständnis für Rechenoperationen* zeigt sich eine geringfügig breitere Streuung mit einem relativen Interquartilsabstand von 16,7 %, während der relative Median von 91,7 % auf einem dauerhaft hohen Niveau verbleibt.

Im Gegensatz dazu präsentiert sich die untere Leistungsgruppe (untere 25 %) als mathematisch hochgradig heterogen, was an den stark gedehnte Interquartilsabständen aller drei Skalen ersichtlich wird. Das Leistungsniveau variiert innerhalb dieser Gruppe erheblich und reicht von 0,0 % erreichter Punkte bis weit über die 50,0-%-Marke hinaus.

Ein profilbezogener Vergleich macht weiterhin deutlich, dass die Defizite der unteren Leistungsgruppe nicht gleichmäßig verteilt sind. Während die Lernenden in den Bereichen *Verständnis für natürliche Zahlen* (Median = 64,3 %) und *Verständnis für das Stellenwertsystem* (Median = 57,1 %) noch moderate Leistungen erzielen, zeigt sich im Bereich *Verständnis für Rechenoperationen* ein deutlicher Leistungseinbruch. Der Median liegt hier mit 41,7 % nur knapp über der 40-%-Marke. Außerdem erreicht nicht einmal das dritte Quartil die Marke von 60 % der maximal zu erreichenden Punkte.

Zusammenfassend dokumentiert der Extremgruppenvergleich, dass mathematische Leistungsschwäche in der untersuchten zweiten Jahrgangsstufe mit einer massiven Streuung der individuellen Kompetenzen einhergeht. Die Schere zwischen den beiden Leistungsspektren klafft hierbei im Bereich der Rechenoperationen deutlich auseinander.

4.2 Psychometrische Skalen- und Itemanalyse des BRT-2

Das nachfolgende Kapitel widmet sich der klassischen testtheoretischen Fundierung des Bayreuther Rechentests (BRT-2) auf Skalen- und Itemebene. Nachdem im vorangegangenen Abschnitt die manifesten Rohwerte und globalen Gruppenunterschiede der Stichprobe im Fokus standen, erfolgt nun der Übergang zur psychometrischen Detailanalyse des Instrumentariums im Rahmen der Klassischen Testtheorie (KTT). Bevor die Datenstruktur den mathematisch hochgradig restriktiven Modellannahmen der probabilistischen Testtheorie (IRT) ausgesetzt wird, gilt es, die Homogenität, fundamentale Messgenauigkeit und aufgabenmetrische Verteilung über traditionelle, etablierte Kennwerte abzusichern. Diese klassische Itemanalyse liefert unverzichtbare deskriptive Indikatoren für die diagnostische Eignung des Aufgabenpools und erlaubt es, die mathematische Güte des Tests schrittweise von groben Skaleneigenschaften bis hin zu spezifischen Fehlerkopplungen feingliedrig zu sezieren.

Um eine logische und inhaltlich kohärente Argumentationskette zu präsentieren, gliedert sich dieses Kapitel in eine dreistufige analytische Sequenz, die sich sukzessive vom Abstrakten zum Konkreten bewegt. Den Ausgangspunkt markiert Kapitel 4.2.1, welches sich der internen Konsistenz und Reliabilität der mathematischen Kompetenzbereiche zuwendet. Hierbei wird über den kontrastierenden Vergleich von Cronbachs α und McDonalds ω die methodisch kritische Annahme der essenziellen τ-Äquivalenz reflektiert. Darauf aufbauend werden in Kapitel 4.2.2 die klassischen itemmetrischen Kennwerte in Form der Aufgabenschwierigkeit konsolidiert. Die deskriptive Betrachtung der mittleren Punkte und Schwierigkeitsscores für die Items A1 bis A18 erlaubt die Rekonstruktion eines empirischen Schwierigkeitsgradienten über das gesamte Testinstrument hinweg.

Den diagnostisch innovativen Abschluss dieser Sektion bildet schließlich Kapitel 4.2.3, welches sich von der reinen Richtig-Falsch-Metrik löst und die bedingten Fehlerwahrscheinlichkeiten sowie assoziierte Fehlernetzwerke in den Mittelpunkt rückt. Über die mathematische Bestimmung bedingter Wahrscheinlichkeiten der Form $P(Aufgabe\ B\ falsch\ |Aufgabe\ A\ falsch)$ werden typische, inhaltlich gekoppelte Fehlerstrukturen und Bearbeitungsbarrieren offengelegt.

4.2.1 Interne Konsistenz und Reliabilität der Kompetenzbereiche

Unter Anwendung der in Kapitel 3.3.7 definierten statistischen Verfahren wurden die Reliabilitätskoeffizienten für die vorliegende Stichprobe berechnet. Die empirischen Kennwerte sowie deren resultierende psychometrische Bewertung sind in Tabelle 6 zusammengefasst.

Tabelle 6 Reliabilitätskoeffizienten der Gesamtskala und der Inhaltsbereiche des BRT-2 (N = 574)

Itembereich	k	α	ω_t	ω_h	Bewertung
Gesamt	18	0,86	0,92	0,72	Exzellent
Verständnis für natürliche Zahlen	6	0,55	0,73	-	Akzeptabel (Solide)
Verständnis für das Stellenwertsystem	6	0,69	0,82	-	Gut
Verständnis für Rechenoperationen	6	0,76	0,83	-	Gut

Anmerkung. k = Anzahl der Items in der jeweiligen Skala, α = Cronbachs Alpha, ω_t = McDonalds Omega total, ω_h = McDonalds Omega Hierarchisch

Für die Gesamtskala des BRT-2 zeigt sich eine hervorragende globale Messgenauigkeit, die durch ein Omega Total von $\omega_t = 0{,}92$ als exzellent eingestuft wird. Das für die hierarchische Struktur entscheidende hierarchische Omega liegt mit $\omega_h = 0{,}72$ über der kritischen methodischen Schwelle. Damit ist statistisch nachgewiesen, dass der übergeordnete Generalfaktor den Großteil der Varianz erklärt. Die Aggregation der Einzelbereiche zu einem interpretierbaren Gesamtwert ist somit vollkommen legitimiert.

Die differenzierte Betrachtung der einzelnen Kompetenzbereiche offenbart die in Kapitel 3.3.7 prognostizierten Diskrepanzen zwischen den Schätzverfahren. Der Inhaltsbereich des *Verständnisses für natürliche Zahlen* erzielt ein formal unzureichendes Cronbachs Alpha von $\alpha = 0{,}55$. Unter Verwendung von McDonalds Omega resultiert jedoch eine akzeptable interne Konsistenz von $\omega_t = 0{,}73$.

Beim Bereich *Verständnis für das Stellenwertsystem* bleibt Cronbachs Alpha mit $\alpha = 0{,}69$ knapp unter der traditionellen Schranke. Jedoch weist das robuste McDonalds Omega mit einem Wert von $\omega_t = 0{,}82$ eine gute Konsistenz auf.

Beim *Verständnis für Rechenoperationen* liegen sowohl Cronbachs Alpha ($\alpha = 0{,}76$) als auch McDonalds Omega ($\omega_t = 0{,}83$) in einem guten Bereich hinsichtlich der internen Konsistenz.

Die Ergebnisse belegen, dass alle drei Teilbereiche des BRT-2 über eine ausreichend hohe psychometrische Qualität verfügen, um für die inhaltlichen Profilanalysen herangezogen zu werden.

4.2.2 Itemmetrische Kennwerte: Aufgabenschwierigkeiten

Zur detaillierteren Analyse der psychometrischen Eigenschaften des BRT-2 werden nachfolgend die itemmetrischen Kennzahlen der einzelnen Testaufgaben untersucht. Im Fokus stehen hierbei die Identifikation von Aufgaben, die für die Stichprobe eine besonders hohe oder geringe Hürde darstellten. Die empirischen Ergebnisse der hierzu in Kapitel 3.3.8 hergeleiteten Schwierigkeitsindizes (P_i) sind in Tabelle 7 zusammengefasst.

Bei der deskriptiven Betrachtung der generierten Schwierigkeitsindizes stechen im Testprofil insbesondere sechs Aufgaben durch ausgeprägte Kennwerte heraus. Zum einen identifiziert die Analyse die Aufgaben A1, A3 und A11 als überdurchschnittlich leicht. Hier wurden die höchsten Schwierigkeitsindizes ermittelt ($P_i \geq 0{,}88$), was methodisch dafürspricht, dass ein überwiegender Großteil der Lernende diese Aufgaben problemlos bearbeiten und korrekt lösen konnten. Wenn man die Aufgaben inhaltlich den Kompetenzbereichen zuordnet, wird ersichtlich, dass zwei der drei Aufgaben im dem Bereich *Verständnis für natürliche Zahlen* (A1: $P_i = 0{,}89$; A3 : $P_i = 0{,}88$) und eine in dem Bereich *Verständnis für das Stellenwertsystem* (A11: $P_i = 0{,}91$) zu verorten sind.

Bei den Aufgaben A2, A17 und A18 wiederum hatten die Lernenden die größten Lösungsschwierigkeiten, was sich in den geringsten berechneten Schwierigkeitsindizes widerspiegelt. Da diese Indizes jedoch im Bereich von 0,54 – 0,61 liegen, lässt sich konstatieren, dass im Durchschnitt dennoch mindestens 50 % der maximalen Punktzahl von der Stichprobe erzielt wurde. Es liegen somit selbst bei den anspruchsvolleren Items

Tabelle 7 Mittlere Punktzahlen und Schwierigkeitsindizes der einzelnen Aufgaben des BRT-2 (N = 574)

Aufgabe	Bereich	Maximalpunkte	Mittlere Punkte	Schwierigkeitsindex (P_i)
A1	VNZ	1	0,89	0,89
A2	VNZ	1	0,61	0,61
A3	VNZ	2	1,77	0,88
A4	VNZ	2	1,57	0,78
A5	VNZ	4	3,27	0,82
A6	VNZ	4	3,14	0,79
A7	VSWS	1	0,84	0,84
A8	VSWS	1	0,73	0,73
A9	VSWS	2	1,42	0,71
A10	VSWS	2	1,61	0,80
A11	VSWS	4	3,65	0,91
A12	VSWS	4	2,94	0,73
A13	VRO	2	1,55	0,78
A14	VRO	2	1,58	0,79
A15	VRO	2	1,41	0,70
A16	VRO	2	1,44	0,72
A17	VRO	2	1,09	0,54
A18	VRO	2	1,19	0,60

Anmerkung. VNZ = Verständnis für natürliche Zahlen, VSWS = Verständnis für das Stellenwertsystem, VRO = Verständnis für Rechenoperationen

keine extremen Ausreißer nach unten vor. Hier fällt strukturanalytisch auf, dass zwei der drei Aufgaben aus dem Bereich *Verständnis für Rechenoperationen* stammen (A17: $P_i = 0{,}54$; A18 : $P_i = 0{,}60$) und eine im Bereich *Verständnis für natürliche Zahlen* liegt (A2: $P_i = 0{,}61$).

Die restlichen zwölf Aufgaben sind im psychometrisch soliden Mittelfeld bei Indizes von 0,71 bis 0,84 zu finden, was auf eine ausgewogene Verteilung der verbleibenden Itemarchitektur hindeutet.

4.2.3 Bedingte Fehlerwahrscheinlichkeiten und Fehlernetze

Unter Anwendung des in Kapitel 3.3.9 beschriebenen methodischen Verfahrens wurden die bedingten Fehlerwahrscheinlichkeiten für die 18 Items des BRT-2 berechnet. Die nachfolgende Abbildung 8 zeigt die vollständige Matrix der paarweise bedingten Fehlerwahrscheinlichkeiten.

4.2.3.1 Numerische Analyse der Fehlermatrix

Die systematische Inspektion von Abbildung 8 verdeutlicht, dass die Fehlermuster im BRT-2 in hochgradig strukturierten Formen auftreten. Besonders prägnant sind die Zeilen der Aufgaben A2 und A17, welche über fast alle Spalten hinweg eine intensive rote Färbung aufweisen. Unabhängig von der Wahl des spezifischen Skalenelements, das von einem Lernenden falsch beantwortet wird, ist die Wahrscheinlichkeit für einen simultanen bzw. gekoppelten Fehler bei den Aufgaben A2 und A17 miteinander verknüpft.

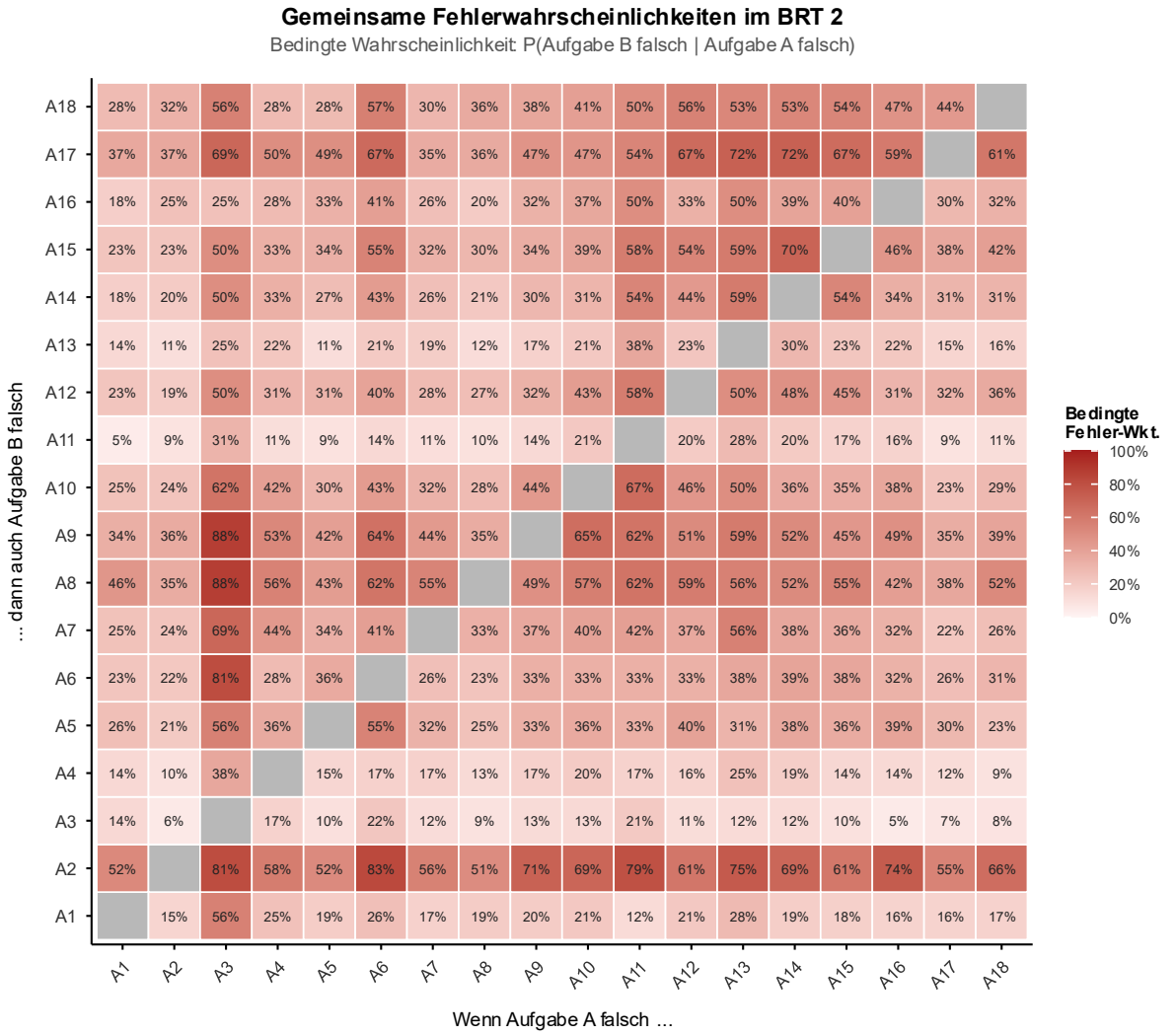


Abbildung 8 Matrix der bedingten Fehlerwahrscheinlichkeiten im BRT-2.

Weiterhin fallen die Zeilen der Aufgaben A3, A4 und A11 auf, da diese über alle Spalten hinweg durchgehend sehr hell gefärbt sind. Diese Items weisen eine deutlich geringere empirische Lösungshürde auf. Unabhängig von der Wahl einer Aufgabe, die durch den Lernenden falsch beantwortet wurde, ist die Wahrscheinlichkeit für einen simultanen Fehler bei den Aufgaben A3, A4 und A11 durchgehend sehr gering. Ein Fehler in einem anderen Testbereich induziert hier folglich kaum systematische Folgefehler auf diesem robusten Skalenniveau.

Auch die theoretisch beschriebene Asymmetrie bedingter Wahrscheinlichkeiten (vgl. Kapitel 2.3.6) lässt sich lehrbuchhaft in der Matrix ablesen. So beträgt zum Beispiel die Wahrscheinlichkeit dafür, dass ein Lernender Aufgabe 8 falsch beantwortet, wenn er bereits Aufgabe 3 falsch beantwortet hat, 88 %. Demgegenüber beträgt die inverse bedingte Wahrscheinlichkeit, bei einem Fehler in Aufgabe 8 auch ein unzureichendes Ergebnis in Aufgabe 3 zu erzielen, lediglich 55 %. Diese Richtungsabhängigkeit unterstreicht empirisch, dass mathematische Fehlermuster hierarchischen Pfaden folgen und nicht zwingend reziprok verlaufen.

4.2.3.2 Topologische Netzwerkanalyse und Identifikation von Fehler-Hubs

In Abbildung 9 ist die topologische Netzwerkstruktur der systematischen Abhängigkeiten ab einem Schwellenwert von $P > 60\,\%$ und im Fruchterman-Reingold-Layout visualisiert. Im Zentrum dieses Graphen aggregieren sich die Aufgaben A2 und A17 unmissverständlich als zentrale Fehler-Hubs (sogenannte Sinks).

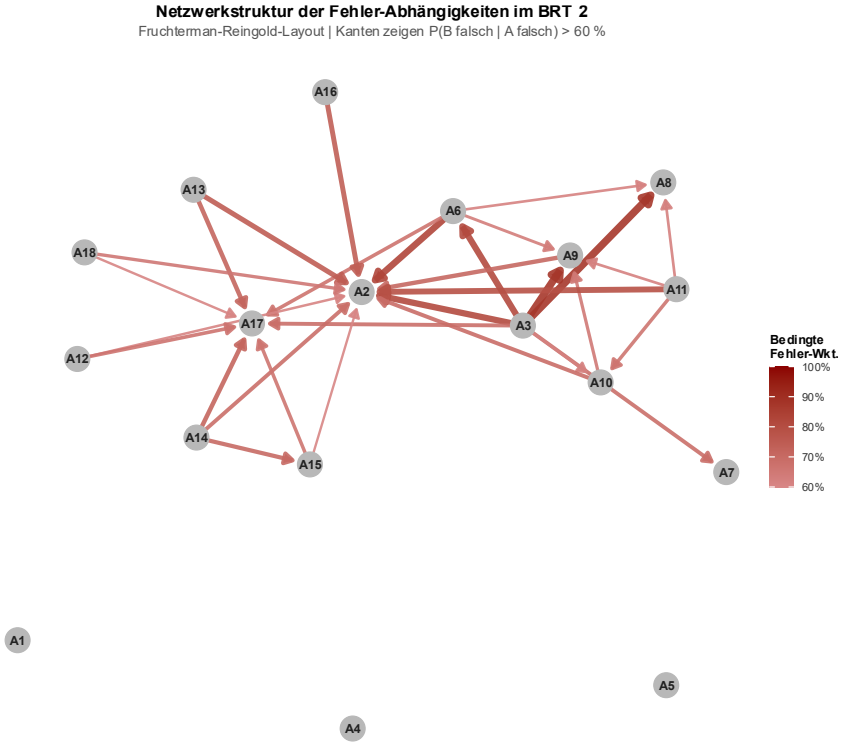


Abbildung 9 Topologische Netzwerkstruktur der Fehler-Abhängigkeit im BRT-2 (Fruchterman-Reingold-Layout, Kanten bei P > 60 %).

Die hohe Dichte an eingehenden Pfeilen bei A2 und A17 visualisiert deren Rolle als „Fehler-Staubecken“. Fast alle peripheren Items des Netzwerks senden gerichtete Kanten in dieses Zentrum aus. Ein Fehler auf den peripheren Skalenelementen zieht somit kaskadenartig einen Folgefehler bei diesen Kernitems nach sich. Aus testdiagnostischer Sicht fungieren diese Hubs als universelle Indikatoren für tieferliegende gemeinsame mathematische Verständnisschwierigkeiten.

Im prägnanten Kontrast dazu stehen die am unteren Rand isolierten Items A1, A4 und A5, welche im Netzwerk keinerlei Kantenverbindung aufweisen. Ein Rückverweis auf die Matrix aus Abbildung 8 verifiziert diese topologische Isolation. Die bedingten Wahrscheinlichkeiten in den zugehörigen Spalten bleiben durchweg auf einem niedrigen bis moderaten Niveau. Das Scheitern an den Aufgaben A1, A4 oder A5 besitzt somit keinerlei systematische Vorhersagekraft für das restliche Testprofil und deutet auf isolierte, itemspezifische Hürden hin, die von den globalen Fehlermustern der Stichprobe entkoppelt sind.

4.3 Personenzentrierte und strukturprüfende Modellanalysen

Nachdem in den vorausgegangenen Abschnitten die variablenorientierten Kennwerte sowie die elementaren Verteilungsstrukturen des BRT-2 detailliert analysiert wurden, widmet sich das vorliegende Kapitel den komplexeren, modellbasierten Testanalysen. Um ein ganzheitliches und methodisch fundiertes Bild der empirischen Testperformanz zu erhalten, wird das Datenmaterial nachfolgend aus zwei komplementären, testtheoretischen Perspektiven beleuchtet.

Die typischen Leistungsprofile der Lernenden werden in Kapitel 4.3.1 mittels einer Cluster Analyse empirisch untersucht, nachdem zuvor die optimale Clusteranzahl mittels vier verschiedener Methoden ermittelt wurde.

Daran schließen sich die strukturprüfenden Dimensionsanalysen (Kapitel 4.3.2) an. Mittels einer explorativen Faktorenanalyse (EFA) wird primär geprüft, ob sich die postulierten mathematischen Kompetenzbereiche empirisch in einer stabilen Faktorenstruktur widerspiegeln. Die dort identifizierte Ladungsarchitektur wird im nächsten Schritt über eine konfirmatorische Faktorenanalyse (CFA) am empirischen Modell auf ihre formale Güte (Model-Fit) getestet.

Im dritten Schritt folgt der Übergang zur probabilistischen Testtheorie (Kapitel 4.3.3). Mittels Item-Response-Theorie (IRT) werden die mathematischen Items des Tests probabilistisch skaliert. Über die Schätzung von Personen- und Itemparametern sowie die softwaregestützte Modellierung von Item-Informationskurven (IIC) und Test-Informationskurven (TIC) wird exakt analysiert, in welchen Fähigkeitsbereichen die einzelnen Aufgaben die höchste Messgenauigkeit aufweisen.

Als vierte Säule wird die Prüfung der Messinvarianz mittels einer Analyse des differenziellen Itemfunktionierens (DIF; Kapitel 4.3.4) durchgeführt. Diese stellt sicher, dass die Items des BRT-2 für die unterschiedlichen Geschlechtsgruppen der Stichprobe bei iden-

tischer latenter Fähigkeit keine systematische Verzerrung aufweisen, was die empirische Fairness des Testverfahrens absichert.

Als fünfte Säule rückt in Kapitel 4.3.5 der Fokus von der rein itemzentrierten Perspektive auf die Makro-Ebene der Lernenden. Durch die Extraktion der latenten Personenparameter (θ) aus dem zuvor etablierten probabilistischen Gesamtmodell werden globale Verteilungsanalysen und robuste, nicht-parametrische Gruppen- und Dimensionsvergleiche durchgeführt. Mittels Wilcoxon-Rangsummentests wird hierbei überprüft, ob sich geschlechtsspezifische Kompetenzprofile auf globaler Skalenebene statistisch absichern lassen, während ein Friedman-Rangsummentest untersucht, ob intraindividuell systematische Leistungsdiskrepanzen zwischen den drei mathematischen Faktoren vorliegen.

Den methodischen und testtheoretischen Abschluss des gesamten Modellkapitels bildet schließlich die Untersuchung auf die Existenz eines Generalfaktors mittels eines Bifaktor-Modells (Kapitel 4.3.6). Im Gegensatz zu klassischen höhergeordneten Faktorstrukturen erlaubt das Bifaktor-Modell eine simultane und orthogonale Zerlegung der Itemvarianz in einen globalen mathematischen Generalfaktor und die jeweiligen spezifischen Gruppenfaktoren. Dies ermöglicht eine präzise Bestimmung der modellbasierten Reliabilitätskoeffizienten, um die Eindimensionalität versus Multidimensionalität des BRT-2 final zu bewerten.

4.3.1 Identifikation typischer Leistungsprofile

Die Aufdeckung der latenten Heterogenität innerhalb der Stichprobe sowie die Identifikation typischer mathematischer Stärken- und Schwächenprofile basieren auf einem partitionsbasierten *k*-Means-Clustering-Algorithmus. Als Clustering-Variablen dienen die prozentualen Leistungswerte (d. h. die prozentual erreichte Punktzahl in den einzelnen Kompetenzbereichen) der bereinigten Gesamtstichprobe ($N = 574$) in den drei mathematischen Kernbereichen des BRT-2.

4.3.1.1 Evidenzbasierte Bestimmung der optimalen Clusteranzahl

Vor der detaillierten profilanalytischen Beschreibung der mathematischen Leistungstypen muss die Wahl der endgültigen Clusteranzahl statistisch legitimiert werden. Die Bestimmung der optimalen Clusterzahl *k* erfolgte über eine simultane Triangulation verschiedener mathematisch-statistischer Gütekriterien. Hierzu wurden die bivariate 2D-Partitionierung, das Ellbogen-Kriterium (WSS), die durchschnittliche Silhouettenbreite sowie die Gap-Statistik nach Tibshirani herangezogen. Die Ergebnisse dieser statistischen Überprüfungen sind in Abbildung 10 zusammengefasst.

Die 2D-Partitionierung der Cluster (Abbildung 10, Panel A) verdeutlicht die geometrische Verteilung der Gruppen im zweidimensionalen Raum, welcher durch die beiden stärksten Hauptkomponenten aufgespannt wird. Die erste Dimension (Dimension 1) fungiert als

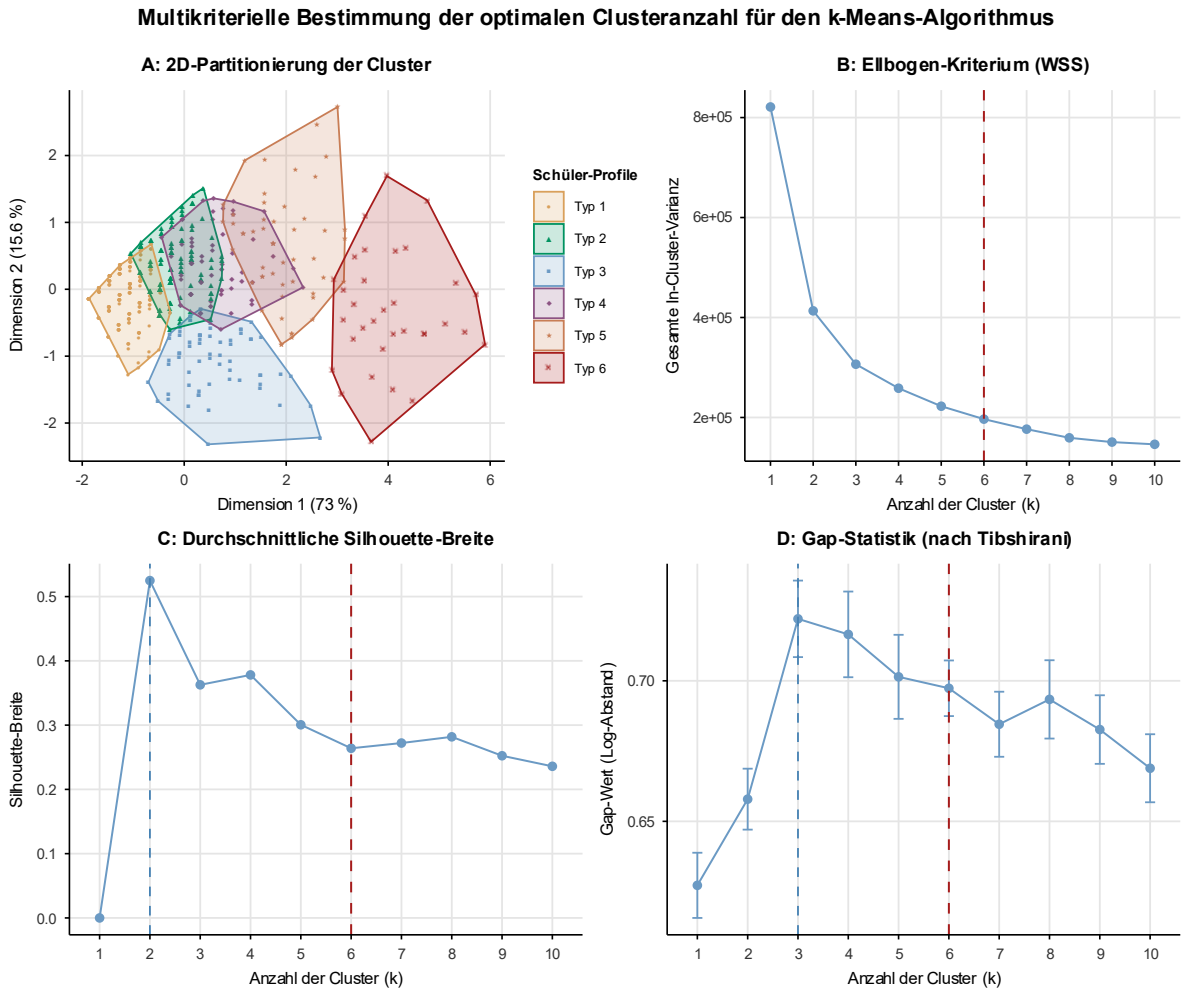


Abbildung 10 Multikriterielle Bestimmung der optimalen Clusterzahl für den k-means-Algorithmus

quantitativer Leistungsfaktor und erklärt 73 % der Gesamtvarianz, während die zweite Dimension (Dimension 2) mit 15,6 % die qualitativen Verlaufsmuster abbildet.

In dieser Darstellung wird ersichtlich, dass sich die Cluster des oberen und mittleren Leistungsspektrums (die Typen 1, 2, 4 und 5) im geometrischen Hauptkomponentenraum teilweise stark überschneiden und keine isolierten Punkthaufen bilden. Dieser Befund lässt sich testtheoretisch primär auf einen ausgeprägten Decken-Effekt zurückführen, welcher sich in einer starken Linksschiefe der mathematischen Primärdaten manifestiert.

Die Differenzierung dieser mathematischen Leistungstypen erfolgt hierbei über eine schichtenartige Anordnung entlang der vertikalen Achse (Dimension 2), welche die qualitativen Diskrepanzen in den Aufgabenprofilen abbildet. Im Gegensatz dazu präsentiert sich am rechten Rand des Leistungsspektrums (aufgrund der niedrigen Fähigkeit auf Dimension 1) das globale Risikoprofil (Typ 6) nahezu vollständig isoliert.

Das Ellbogen-Kriterium (Abbildung 10, Panel B) zeigt den Verlauf der totalen Varianz innerhalb der Cluster (Within-Cluster Sum of Squares, WSS) für die Lösungen $k = 1$ bis $k = 10$. Der steilste Informationsgewinn vollzieht sich beim Übergang von der Ein- zur Zwei-Cluster-Lösung. Exakt ab dem Knickpunkt bei $k = 2$ oder $k = 3$ stabilisiert sich der

Kurvenverlauf und geht in eine flachere Kurve über. Eine weitere Erhöhung der Clusterzahl darüber hinaus liefert somit keine ökonomische Reduktion der verbleibenden Restvarianz mehr.

Die durchschnittliche Silhouettenbreite (Abbildung 10, Panel C) weist ein globales Maximum bei *k* = 2 auf (Silhouetten-Koeffizient $> 0{,}50$). Dieser Peak resultiert aus dem dominanten Leistungs-Generalfaktor der Stichprobe (Dimension 1 erklärt 73 % der Varianz), welcher in der unüberwachten Klassifikation eine strikte Dichotomisierung in leistungsschwache und leistungsstarke Lernende mathematisch begünstigt. Unter den feingliedrigeren Partitionen zeigt sich bei *k* = 6 jedoch ein stabiles lokales Niveau, das für die vorliegenden empirische Daten eine psychometrisch valide Repräsentation bildet.

Als ergänzendes Absicherungsverfahren wurde die Gap-Statistik nach Tibshirani (Abbildung 10, Panel D) berechnet, welche die empirische In-Cluster-Varianz mit einer synthetischen Null-Referenzverteilung vergleicht. Der globale Peak der Kurve liegt bei $k = 3$ (Gap-Wert $> 0{,}7$). Beim Übergang zu $k = 6$ erfährt die Kurve jedoch innerhalb der Standardfehlergrenzen eine deutliche Stabilisierung und bildet ein lokales Plateau, bevor die Werte bei $k = 7$ wieder degressiv einbrechen.

Aus testtheoretischer Sicht wird die mathematisch induzierte 6-Cluster-Lösung somit gestützt. Eine Restriktion auf eine Lösung von $k = 2$ oder $k = 3$ würde zu einer informationstheoretischen Maskierung der Datenstruktur führen: Diagnostisch relevante, qualitative Diskrepanzprofile würden dabei statistisch nivelliert und mit den rein linearen Verläufen vermengt. Erst die Extraktion von sechs Clustern erlaubt es, die qualitativen Unterschiede im Umgang mit den drei Inhaltsbereichen trennscharf freizulegen.

4.3.1.2 Deskriptive Analyse der extrahierten Leistungsprofile

Die finale 6-Cluster-Lösung partitioniert die Gesamtstichprobe ($N = 574$) in sechs distinkte mathematische Leistungsprofile. Die empirischen Verläufe der sechs extrahierten Lernenden-Typen sind in Abbildung 11dargestellt und die exakten Clustergrößen sowie die spezifischen Profilcharakteristika sind in Tabelle 8 zu finden.

Wie aus Abbildung 11 ersichtlich wird, unterscheiden sich die identifizierten Profile fundamental in ihrer geometrischen Form und weisen teilweisekreuzende Verläufe auf. Die einzelnen Cluster stellen sich rein deskriptiv wie folgt dar.

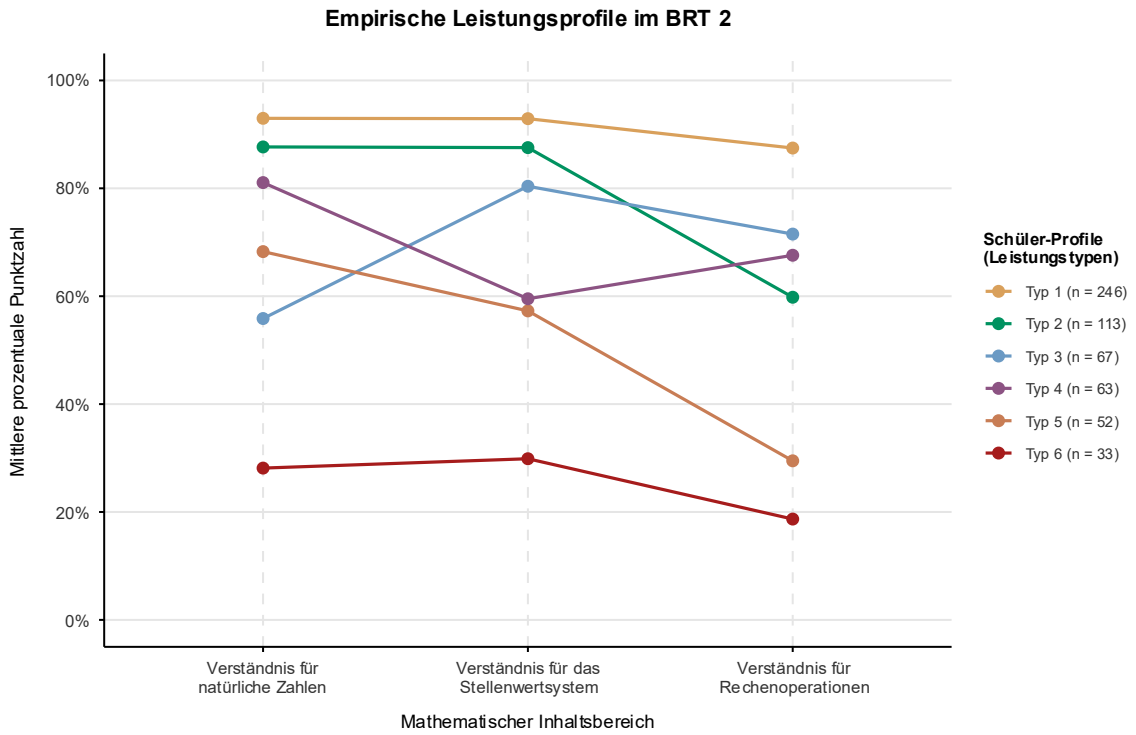


Abbildung 11 Empirische Leistungsprofile im BRT-2 über die drei mathematischen Kompetenzbereiche hinweg für die extrahierte 6-Cluster-Lösung.

Tabelle 8 Deskriptive Kennwerte und Charakteristika der sechs mathematischen Leistungstypen (N = 574)

Leistungstyp	n	Prozentualer Anteil	Bezeichnung des Profils	Charakteristischer Verlauf
Typ 1	246	42,9 %	Kompetenzstarke Allrounder	Homogen hohes Maximum über alle drei Skalen hinweg.
Typ 2	113	19,7 %	Konzeptuell Solide mit Operationslücken	Hohes Niveau beim VNZ und VSWS. Einbruch bei VRO.
Typ 3	67	11,7 %	Kompensatorischer System-Typ	Nicht-linearer Verlauf: Niedriger Start bei VNZ, ausgeprägter Peak bei VSWS und VRO.
Typ 4	63	11,0 %	Mechanisch-prozeduraler Rechentyp	Nicht-linearer Verlauf: Moderate Werte bei VNZ und VOR, selektiver Einbruch bei VSWS.
Typ 5	52	9,1 %	Basal-numerischer Typ	Solide Basiswerte bei VNZ, kontinuierlicher Absturz bei steigender Skalenkomplexität.
Typ 6	33	5,7 %	Globales Risikoprofil	Homogen kritisches Minimum über alle mathematischen Bereiche hinweg.

Anmerkung. Die Prozentangaben wurden auf eine Dezimale gerundet und beziehen sich auf den Anteil an der bereinigten Gesamtstichprobe. n = Stichprobengröße.

4.3.1.2.1 Typ 1 – Kompetenzstarke Allrounder

Dieses größte Cluster umfasst mit 42,9 % knapp die Hälfte der Gesamtstichprobe (n = 246). Die Lernenden dieses Typs zeigen über alle drei Inhaltsbereiche hinweg eine konstant hohe Performanz. In den Bereichen *Verständnis für natürliche Zahlen* und *Verständnis für das Stellenwertsystem* liegt die mittlere prozentuale Punktzahl geringfügig über den Werten im Bereich *Verständnis für Rechenoperationen*. Insgesamt werden in allen drei Bereichen weit über 80 % der erreichbaren Punkte erzielt. Die Gruppe repräsentiert das quantitativ leistungsstärkste Profil der Erhebung.

4.3.1.2.2 Typ 2 – Konzeptuell Solide mit Operationslücken

Das zweitgrößte Cluster repräsentiert knapp ein Fünftel der Gesamtstichprobe (n = 113, 19,7 %). Lernende dieses Typs weisen beim basalen *Verständnis für natürliche Zahlen* sowie beim *Verständnis für das Stellenwertsystem* hohe Kompetenzwerte auf und bewegen sich auf dem Niveau der Spitzengruppe (Typ 1). Sobald jedoch das Ausführen von Rechenoperationen gefordert wird, bricht die Leistungskurve signifikant auf einen prozentualen Mittelwert von ca. 60 % ein.

4.3.1.2.3 Typ 3 – Kompensatorischer System-Typ

Dieses Cluster umfasst 11,7 % der Stichprobe (n = 67) und zeigt einen ausgeprägten nicht-linearen Verlauf. Die Lernenden starten im Kompetenzbereich *Verständnis für natürliche Zahlen* mit einem niedrigen Mittelwert (< 60 %). Im Bereich des *Verständnisses für das Stellenwertsystem* folgt jedoch ein deutlicher Leistungssprung auf einen Mittelwert von etwa 80 %. Dieses Niveau kann auch im Bereich des *Verständnisses für Rechenoperationen* mit etwas unter 80 % erfolgreich gehalten werden, sodass die Kurve nach dem initialen Tief einen stabilen, überdurchschnittlichen Verlauf nimmt.

4.3.1.2.4 Typ 4 – Mechanisch-prozeduraler Rechentyp

In diesem Profil befinden sich 11,0 % der Lernenden (n = 63) und es ist durch einen wellenförmigen, nicht-linearen Kurvenverlauf charakterisiert. Die Lernenden zeigen ein stabiles Fundament bei den natürlichen Zahlen (knapp über 80 %) und erzielen auch beim *Verständnis für Rechenoperationen* einen absolut moderaten Mittelwert von knapp 70 %. Im statistischen Kontrast dazu bricht die Leistungskurve jedoch exakt im dazwischenliegenden Kompetenzbereich *Verständnis für das Stellenwertsystem* signifikant auf knapp unter 60 % ein.

4.3.1.2.5 Typ 5 – Basal-numerischer Typ

Dieses Cluster zeigt mit einem Anteil von 9,1 % (n = 52) einen kontinuierlichen, fast linearen Leistungsabfall bei ansteigender Komplexität der Testanforderungen. Während die Lernenden beim *Verständnis für natürliche Zahlen* mit knapp unter 70 % noch ein stabiles Ergebnis erzielen, sinken die Werte beim *Verständnis für das Stellenwertsystem* bereits auf unter 60 % ab. Beim *Verständnis für Rechenoperationen* dekompensiert das Profil weiter und erreicht einen kritischen Mittelwert von rund 30 %.

4.3.1.2.6 Typ 6 – Globales Risikoprofil

Dieses statistisch kleinste Cluster umfasst 5,7 % der Gesamtstichprobe (n = 33). Die Lernenden weisen in allen drei Teilbereichen des BRT-2 kritische Leistungswerte auf. Bereits beim *Verständnis für natürliche Zahlen* und beim Stellenwertsystem verbleiben die Leistungen im unteren Drittel der Skala. Im Kompetenzbereich *Verständnis für Rechenoperationen* wird mit einem mittleren Leistungswert von weniger als 20 % das absolute psychometrische Minimum dieses Profils verzeichnet.

Zusammenfassend dokumentieren die profilzentrierten Analysen eine ausgeprägte qualitative Vielfalt der Leistungsdaten. Insbesondere das zeitgleiche Vorliegen der beiden entgegengesetzten Diskrepanzprofile Typ 3 und Typ 4 belegt empirisch, dass identische Testgesamtwerte auf völlig unterschiedlichen mathematischen Profilarchitekturen beruhen können.

4.3.2 Strukturaufklärung und faktorielle Validität des BRT-2

Zur empirischen Überprüfung der angenommenen mathematischen Kompetenzbereiche und zur Aufklärung der dreidimensionalen Struktur des Bayreuther Rechentests (BRT-2) wurde eine explorative Faktorenanalyse (EFA) durchgeführt. Die methodisch-mathematischen Entscheidungskriterien folgen dem in Kapitel 3.3.11 definierten Vorgehen.

4.3.2.1 Überprüfung der Faktorabilität

Vor der Extraktion der Faktoren wurde die Spearman-Korrelationsmatrix der 18 Testitems (siehe Anhang B) auf ihre faktorielle Eignung hin untersucht.

Der Bartlett-Test auf Sphärizität erwies sich als hochsignifikant, $\chi^2 = 2079{,}07$; $df = 153$; $p < 0{,}001$. Damit kann die Nullhypothese einer Identitätsmatrix und somit einer vollständigen Unkorreliertheit der Items, verworfen werden. Es liegen statistisch bedeutsame Linearzusammenhänge vor.

Das globale Kaiser-Meyer-Olkin-Kriterium (KMO) lieferte mit einem Overall MSA von 0,91 einen exzellenten Wert („marvelous“ nach Kaiser & Rice, 1974). Auf Ebene der Einzelvariablen zeigten sie durchweg homogene Werte, die von $MSA = 0{,}87$ (A17) bis $MSA = 0{,}93$ (A3, A9, A11) reichten und damit die konservative Hürde von 0,60 deutlich überschritten. Die Datenmatrix ist folglich in hohem Maße faktorisierbar.

4.3.2.2 Faktorenextraktion und das Kriterium der Einfachstruktur

Zunächst wurde die Anzahl der Faktoren mittels einer Parallelanalyse bestimmt. Hierbei wurde eine initiale Struktur von sechs Faktoren erhalten.

Die Betrachtung des 6-Faktoren-Modells (siehe Anhang C) zeigte weiterhin, dass das Kaiser-Guttman-Kriterium (Eigenwert > 1) lediglich für die Faktoren 1, 2, 3 und 6 erfüllt ist (Faktoren 4 und 5 weisen SS-Loadings < 1 auf). Weiterhin erklären alle sechs Faktoren

zusammen 45,8 % der Gesamtvarianz. Hierbei entfallen auf die ersten drei Faktoren 30,4 % der Gesamtvarianz, während die Faktoren 4 bis 6 lediglich 15,4 % aufklären.

Des Weiteren lädt auf Faktor 4 nur ein einziges Item (A5) mit 0,697 sauber und auf Faktor 5 lediglich zwei Elemente (A1: 0,535 und A3: 0,471). Da derartige singuläre oder schwach besetzte Faktoren testtheoretisch kaum interpretierbar sind, sprechen diese Befunde deutlich für eine Drei-Faktor-Lösung und gegen eine Sechs-Faktoren-Lösung.

In Konsequenz wurde eine restriktive Drei-Faktor-Lösung berechnet. Entsprechend des in Kapitel 3.3.11 definierten theoretisch begründeten Vorgehens wurde hierbei eine oblique Promax-Rotation appliziert. Nach der Rotation erklären die drei extrahierten Dimensionen zusammen 40,7 % der Gesamtvarianz. Diese Varianzaufklärung setzt sich aus 14,4 % durch Faktor 1, 14,1 % durch Faktor 2 und 12,2 % durch Faktor 3 zusammen. Angesichts der ökonomischen Reduktion von sechs auf drei Dimensionen ist der marginale Verlust an Varianzaufklärung zugunsten einer stabilen Einfachstruktur methodisch geboten.

Die Zuordnung der Aufgaben zu den Dimensionen erfolgt primär über die Mustermatrix (Pattern Matrix), da diese die um Interfaktorkorrelation bereinigten, reinen standardisierten Regressionskoeffizienten abbildet. Tabelle 9 zeigt die Ladungsstruktur im Detail.

Tabelle 9 Rotierte Pattern-Matrix der 3-Faktor Lösung mit den wichtigsten Kennwerten der Faktoren und der Faktorkorrelationsmatrix. (Ladungen < 0,30 wurden ausgeblendet)

Item	Faktor 1	Faktor 2	Faktor 3	h^2
A1	0,395			0,158
A2			0,672	0,436
A3	0,689			0,485
A4	0,483			0,183
A5	0,404			0,353
A6			0,420	0,432
A7	0,679			0,416
A8	0,646			0,427
A9	0,432		0,418	0,616
A10	0,453			0,469
A11		0,376		0,420
A12		0,568		0,477
A13			0,696	0,506
A14		0,732		0,525
A15		0,662		0,579
A16			0,826	0,601
A17		0,872		0,556
A18		0,448	0,308	0,475
SS loadings	2,601	2,538	2,196	
Proportion Var	0,144	0,141	0,122	
Cumulative Var	0,144	0,285	0,407	

Item	Faktor 1	Faktor 2	Faktor 3	h^2
Faktorkorrelationen				
Faktor 1	-			
Faktor 2	0,674	-		
Faktor 3	0,705	0,658	-	

Anmerkung. N = 574. Faktorenextraktion mittels Minimum-Residual-Methode mit Promax-Rotation. h^2 = Kommunalität, SS loadings = Quadratsumme der Faktorladungen

Die empirische Strukturierung deckt sich in weiten Teilen mit den theoretischen Erwartungen, weist jedoch spezifische Besonderheiten auf. Faktor 1 wird primär durch die Items A1, A3, A4, A5, A7, A8, A9 und A10 konstruiert, während unter Faktor 2 die Items A11, A12, A14, A15, A17 und A18 fallen. Die verbleibenden Aufgaben A2, A6, A13 und A16 bilden den Faktor 3. Diese Aufgaben bilden, wie auch die Korrelationsmatrix (vgl. Anhang B) visuell stützt, ein hochkorreliertes, eigenständiges Cluster.

Weiterhin ist hier erkennbar, dass die Aufgaben A9 und A18 signifikante Doppelladungen aufweisen. Während Aufgabe 9 auf Faktor 1 (0,432) und fast gleichstark auf Faktor 3 (0,418) lädt, lässt sich Aufgabe 18 auf den Faktoren Faktor 2 (0,448) und 3 (0,308) verorten.

Zusätzlich liefert die Faktorkorrelationsmatrix am Ende von Tabelle 9 wertvolle testtheoretische Erkenntnisse. Die drei extrahierten Faktoren weisen untereinander hohe positive Zusammenhänge auf, die von r = 0,658 (Faktor 2 und 3) bis zu r = 0,705 (Faktor 1 und 3) reichen. Diese ausgeprägten Interkorrelationen rechtfertigen nicht nur methodisch die Wahl einer obliquen Rotation, sondern liefern zudem eine starke empirische Evidenz für das Vorliegen eines übergeordneten Generalfaktors, welcher nachfolgend in Kapitel 4.3.6 über ein Bifaktor-Modell explizit geprüft wird.

4.3.2.3 Konfirmatorische Absicherung und Modellvergleich

Zur Überprüfung der aus der EFA gewonnen Drei-Faktor-Struktur wurde eine konfirmatorische Faktorenanalyse (CFA) berechnet und gegen das theoretische angenommene Ausgangsmodell sowie ein unifaktorielles Basismodell (Einfaktormodell) getestet. Tabelle 10 fasst die robusten Fit-Indizes zusammen.

Tabelle 10 Robuster Fit Vergleich der konkurrierenden Strukturmodelle

Modell	Robust CFI	Robust TLI	Robust RMSEA	SRMR
Einfaktormodell	0,879	0,862	0,079	0,064
Theoretisches Modell	0,898	0,882	0,073	0,060
EFA-3-Faktor-Modell	0,943	0,934	0,055	0,051

Anmerkung. N = 574. CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual.

Der Modellvergleich zeigt eine eindeutige Überlegenheit des datengeleiteten EFA-3-Faktor-Modells. Während das rein theoretische Modell und das Einfaktormodell die konventionellen Schwellenwerte von $> 0{,}90$ (bzw. die restriktiven Kriterien nach Hu & Bentler, 1999) für den CFI und TLI verfehlen, erzielt das EFA-Modell mit einem $CFI = 0{,}943$ und einem $TLI = 0{,}934$ eine gute Anpassung. Der $RMSEA = 0{,}055$ (wünschenswert: $<$

$0{,}06$) sowie der $SRMR = 0{,}051$ (wünschenswert: $< 0{,}08$) bestätigen die exzellente Modellpassung des modifizierten Drei-Faktor-Gefüges.

4.3.3 Probabilistische Analysen und diagnostische Eignung

Entsprechend dem in Kapitel 3.3.12 definierten Protokolls wird in diesem Kapitel die empirische Parameterschätzung des Generalized Partial Credit Models (GPCM) für die drei Faktoren des BRT-2 dokumentiert. Die Analysen basieren auf der Gesamtstichprobe von *N* = 574 Lernenden.

4.3.3.1 Empirische Prüfung der IRT-Modellvoraussetzungen

Die mathematischen Voraussetzungen für die probabilistische Skalierung wurden für alle drei Faktoren konsequent bestätigt. Hinsichtlich des globalen Modell-Fits und der damit verbundenen essenziellen Eindimensionalität weisen die informationstheoretischen Fit-Indizes für alle drei Faktoren eine hervorragende funktionale Anpassung auf. Für Faktor 1 zeigt der globale M_2-Test keine statistisch signifikante Abweichung vom unifaktoriellen Modell ($M_2 = 17{,}198,\ df = 20,\ p = 0{,}64$), was durch einen $RMSEA = 0{,}00$ und einen $CFI = 1{,}00$ gestützt wird. Für Faktor 2 ($M_2 = 21{,}785,\ df = 9,\ p = 0{,}01$) und Faktor 3 ($M_2 = 3{,}327,\ df = 2,\ p = 0{,}19$) belegen die Werte der Fit-Indizes ($RMSEA_{F2} = 0{,}050$, $CFI_{F2} = 0{,}989$; $RMSEA_{F3} = 0{,}034,\ CFI_{F3} = 0{,}997$) ebenfalls eine uneingeschränkte Modellpassung. Die vollständige Ergebnismatrix der globalen Passung ist in Tabelle C im Anhang D hinterlegt.

Die statistische Analyse der Residualkorrelationen zur Überprüfung der lokalen Unabhängigkeit schließt kritische lokale Abhängigkeiten über alle Skalen hinweg aus (siehe die vollständigen Q_3-Matrizen im Anhang E). Der maximale positive Korrelationswert liegt bei unbedenklichen 0,025. Die über den algorithmischen Filter extrahierten Paarungen für Faktor 1, namentlich die Itemkombinationen A10-A3 mit $Q_3 = -0{,}205$ und A9-A8 mit $Q_3 = -0{,}268$, weisen ausnahmslos negative Vorzeichen auf. Da testtheoretisch ausschließlich positive Residuenkorrelationen eine Verletzung der lokalen Unabhängigkeit signalisieren, sind diese negativen Werte psychometrisch irrelevant und bedürfen keiner Skalenbereinigung.

Für die sechs Items des zweiten Faktors bestätigt sich dieses exzellente Bild weiter. Der maximale empirische Q_3-Wert beträgt hier lediglich -0,042. Dies bedeutet, dass sämtliche verbliebenen Residuenkorrelationen innerhalb dieses Faktors im negativen Bereich verortet sind, womit jegliche lokalen Abhängigkeiten empirisch ausgeschlossen werden können. Die vier Items des dritten Faktors zeigen ein analoges, hochgradig stabiles Ergebnis. Mit einem maximalen Q_3-Wert von -0,085 verbleiben auch hier alle Werte strikt unterhalb der kritischen Relevanzgrenze, was die Annahme lokaler Unabhängigkeit für die finale Modellierung uneingeschränkt legitimiert.

4.3.3.2 Lokaler Item-Fit und Analyse pathologischer Schwellenparameter

Die Überprüfung des lokalen Item-Fits via $S - X^2$-Statistik offenbart eine weitgehend homogene Modellpassung für die Faktoren 1 und 3, da die empirischen Werte fast ausnahmslos weit außerhalb der kritischen Signifikanzgrenze verbleiben ($p \geq 0{,}05$). Auf Ebene des zweiten Faktors zeigt sich jedoch ein Modellverstoß, die Aufgabe A11 fällt mit einem Wert von $S - X^2 = 36{,}278$ ($df = 17,\ p = 0{,}004$) eklatant aus dem gemeinsamen psychometrischen Modellgefüge. Ebenso weist die Aufgabe A14 mit $p = 0{,}039$ eine statistisch signifikante Passungsunschärfe auf dem 5 %-Niveau auf.

Die testtheoretischen Ursachen dieser Passungsdefizite lassen sich visuell über die spezifischen Verläufe der Kategorie-Antwortkurven (Trace Lines) sowie für den ersten Faktor über die numerischen Schwellenparameter der Tabelle 11 nachvollziehen. Die vollständigen Parameterschätzungen und Fit-Tabellen aller 18 Testitems des BRT-2 sind zur lückenlosen Dokumentation im Anhang F hinterlegt.

Tabelle 11 Parameterstruktur des GPCM für ausgewählte Items aus Faktor 1

Item	Diskrimination (a)	b1	b2	b3	b4
A1	0,785	-2,910	NA	NA	NA
A3	1,403	-2,596	-1,355	NA	NA
A4	0,649	-3,070	-1,132	NA	NA
A5	0,353	3,948	-1,784	-1,532	-7,020

Anmerkung. N = 574. Auszug der empirischen Kennwerte aus der globalen IRT-Schätzung für Faktor 1. NA indiziert ein von Natur aus dichotom oder niedrigstufig kodiertes Aufgabenformat.

Wie die Kennwerte in Tabelle 11 ausweisen, liegt bei der Aufgabe A5 eine schwere Inversion der Schwellenparameter vor. Der mathematische Schwellenwert sinkt von einem extrem hohen initialen Übergangspunkt ($b_1 = 3{,}949$) über die nachfolgenden Zwischenstufen drastisch ab und kulminiert in einem stark negativen finalen Wert ($b_4 = -7{,}019$). Die kategorialen Konsequenzen dieser ordnungswidrigen Schwellenstruktur werden im zugehörigen Trace-Line-Plot in Abbildung 12 visuell fassbar.

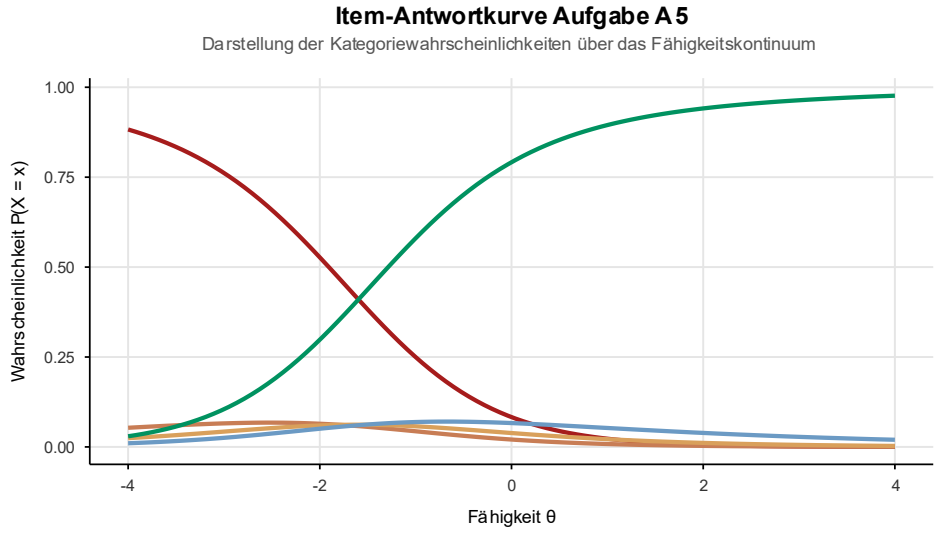


Abbildung 12 Kategorie-Antwortkurven (Trace Lines) für das Item A5 mit vollständig komprimierten mittleren Antwortstufen

Die mittleren Antwortkategorien für 1 Punkt, 2 Punkte und 3 Punkte werden im mathematischen Schätzprozess völlig flach komprimiert. Sie erreichen zu keinem Zeitpunkt des latenten Fähigkeitsspektrums θ die höchste Antwortwahrscheinlichkeit. Das Item misst in der diagnostischen Realität folglich rein dichotom, da die Lernenden entweder direkt in der niedrigsten Kategorie verbleiben (0 Punkte) oder das Item vollständig bis zur Maximalkategorie (4 Punkte) lösen. Die Zwischenstufen der Punktevergabe erweisen sich als psychometrisch dysfunktional.

Die empirischen Kurvenverläufe für den zweiten Faktor offenbaren für das defizitäre Item A11 eine exakt identische strukturelle Anomalie, welche in Abbildung 13 dokumentiert ist.

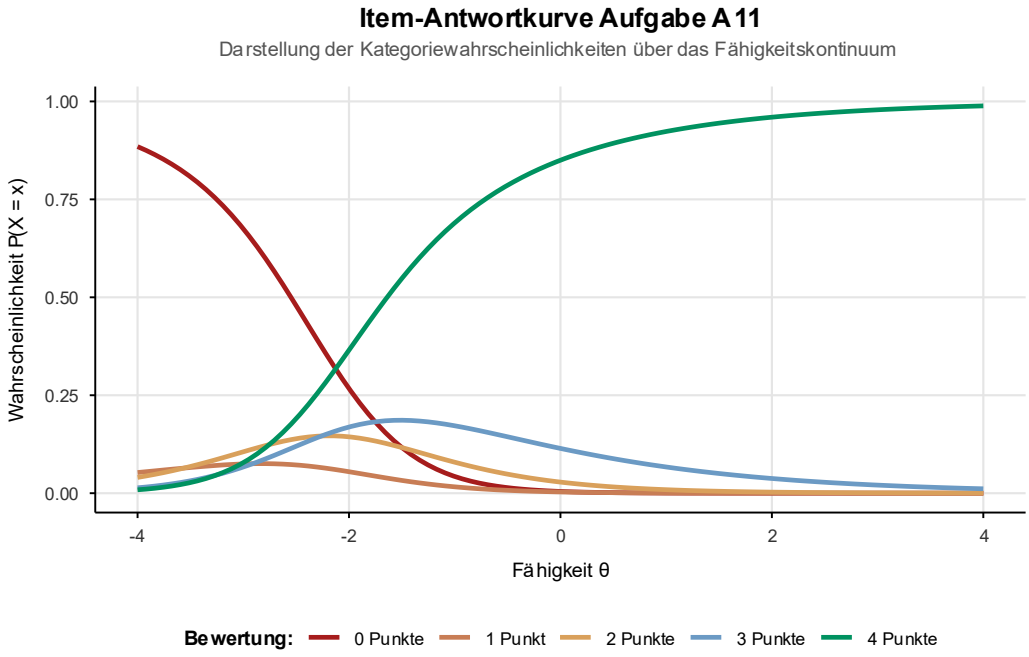


Abbildung 13 Kategorie-Antwortkurven (Trace Lines) für das fehlende Item A11 mit implodierten Teilpunkte-Kategorien

Auch ohne den expliziten Einbezug der numerischen Schwellenwerte zeigt der visuelle Befund des Trace-Line-Plots, dass die mittleren Teilpunkte-Kategorien im Schätzprozess vollständig kollabieren. Dies erklärt den massiven statistischen Misfit im $S - X^2$-Test ($p = 0{,}004$) unmittelbar. Das mathematische Modell des GPCM versucht krampfhaft, ein polytomes Stufenschema über eine Aufgabe zu zwingen, welche von den Lernenden in der Testpraxis faktisch rein binär beantwortet wird. Das Item verweigert sich somit der theoretisch intendierten messtheoretischen Differenzierung.

Im Gegensatz zur totalen Kategorieimplosion bei Aufgabe A11 zeigt der Kurvenverlauf von Aufgabe A14 einen klassischen psychometrischen Grenzfall auf (die zugehörige Grafik ist im globalen Panel für Faktor 2 im Anhang G verortet). Die mittlere Antwortkategorie besitzt hier zwar einen mathematisch identifizierbaren Peak, dieser fällt jedoch im direkten Vergleich extrem schmal aus und wird von den dominanten Außenkategorien stark eingeengt. Diese graduelle Kategoriekompression führt im Zusammenspiel mit der hohen Test-

stärke der vorliegenden Stichprobengröße (N = 574) zum signifikanten Anschlag des lokalen Item-Fits ($S - X^2 = 31{,}126,\ p = 0{,}039$).

Im Kurvenprofil des dritten Faktors zeigt sich für die Aufgabe A6 ein analoges, wenn auch im Vergleich zu Faktor 2 deutlich abgeschwächtes Kompressionsmuster der mittleren Antwortstufen. Infolge dieser moderaten Verengung behauptet das Item mit einem Wert von $S - X^2 = 7{,}108$ ($p = 0{,}069$) die formale Signifikanzgrenze des Item-Fits, weshalb es als ausreichend modellkonform eingestuft wird. Die vollständigen Übersichten aller Kategorie-Antwortkurven, getrennt nach den drei Faktoren, sind zur umfassenden Dokumentation im Anhang G hinterlegt.

4.3.3.3 Lokale Item- und globale Testinformationserträge

Die Auswertung der Item-Informationskurven (IIC; Abbildung 14) und Testinformationskurven (TIC; Abbildung 15) liefert ein eindeutiges empirisches Profil bezüglich des diagnostischen Einsatzzwecks des BRT-2.

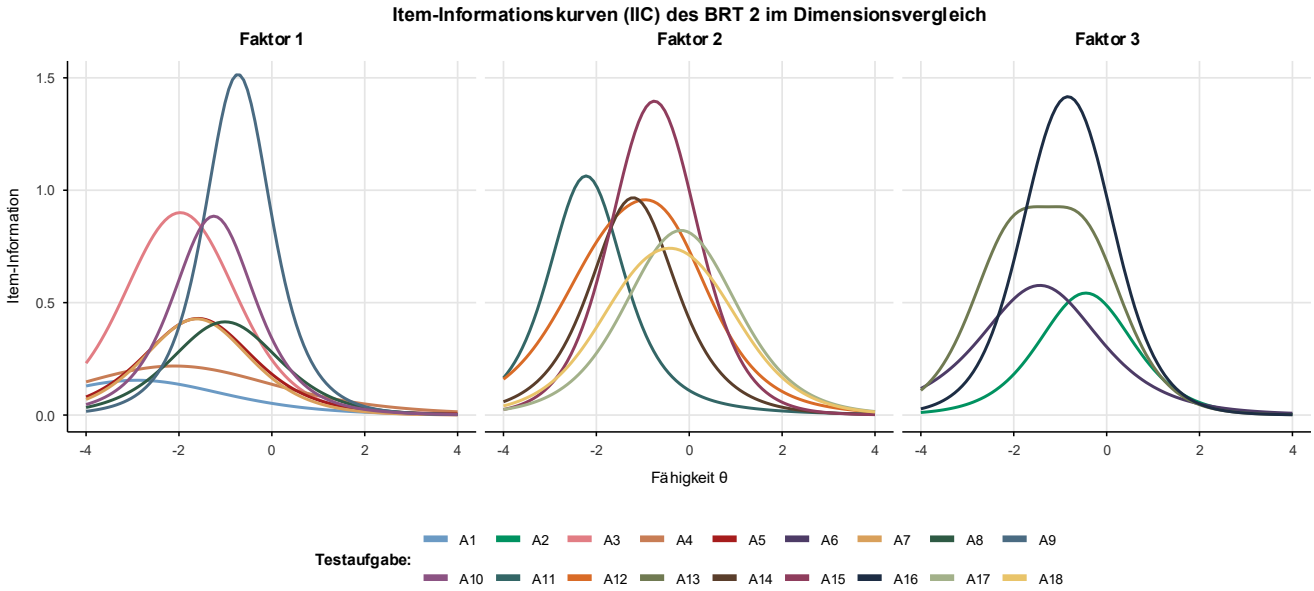


Abbildung 14 Item-Informationskurven (IIC) getrennt nach Faktor 1, Faktor 2 und Faktor 3.

Hinsichtlich der asymmetrischen Verortung der Maxima zeigt Abbildung 14, dass die Peaks der Informationskurven über alle drei Faktoren hinweg geschlossen im negativen Fähigkeitsbereich zwischen $\theta = 0$ und $\theta = -2$ liegen. Oberhalb des Mittelwerts ($\theta > 0$) konvergieren die Informationserträge aller Aufgaben rasch gegen Null.

Bezüglich der differenziellen Messkraft auf Itemebene lassen sich innerhalb der Faktoren deutliche Diskrepanzen in der Diskriminationskraft identifizieren. Als herausragend trennscharfe Indikatoren im kritischen θ-Raum erweisen sich die Aufgaben A10 (Faktor 1) mit einem Maximum von $I(\theta) \approx 1{,}5$, Aufgabe A15 (Faktor 2) und Aufgabe 16 (Faktor 3) jeweils mit einem Peak von $I(\theta) \approx 1{,}4$. Demgegenüber weisen die Aufgaben A1, A4 und das komprimierte Item A5 extrem flache, informationsarme Kurvenverläufe auf und tragen nur marginal zur Verringerung des Standardmessfehlers bei.

Die Zusammenführung der Item-Informationen zu faktorenbezogenen Test-Informationskurven (Tabelle 12, visualisiert in Abbildung 15) bestätigt die psychometrische Stabilität der dimensionalen Architektur unter Berücksichtigung der Skalenlängen.

Tabelle 12 Globale IRT-Reliabilitätskennwerte der drei Faktoren

Faktor	Itemzahl	Maximaler Informationsertrag	Empirische Reliabilität
Faktor 1	8	$I(\theta) \approx 4{,}3$	0,647
Faktor 2	6	$I(\theta) \approx 5{,}0$	0,746
Faktor 3	4	$I(\theta) \approx 3{,}3$	0,645

Anmerkung. Globale Kennwerte aggregiert aus den summierten Item-Informationsfunktionen.

Faktor 2 generiert mit einem Informationsmaximum von $I(\theta) \approx 5{,}0$ bei $\theta = -1$ den höchsten globalen Informationsertrag, was konsistent mit der höchsten empirischen Reliabilität von 0,746 korrespondiert. Für eine ökonomische Skala von nur sechs Items stellt dies eine psychometrisch sehr stabile Teststärke dar.

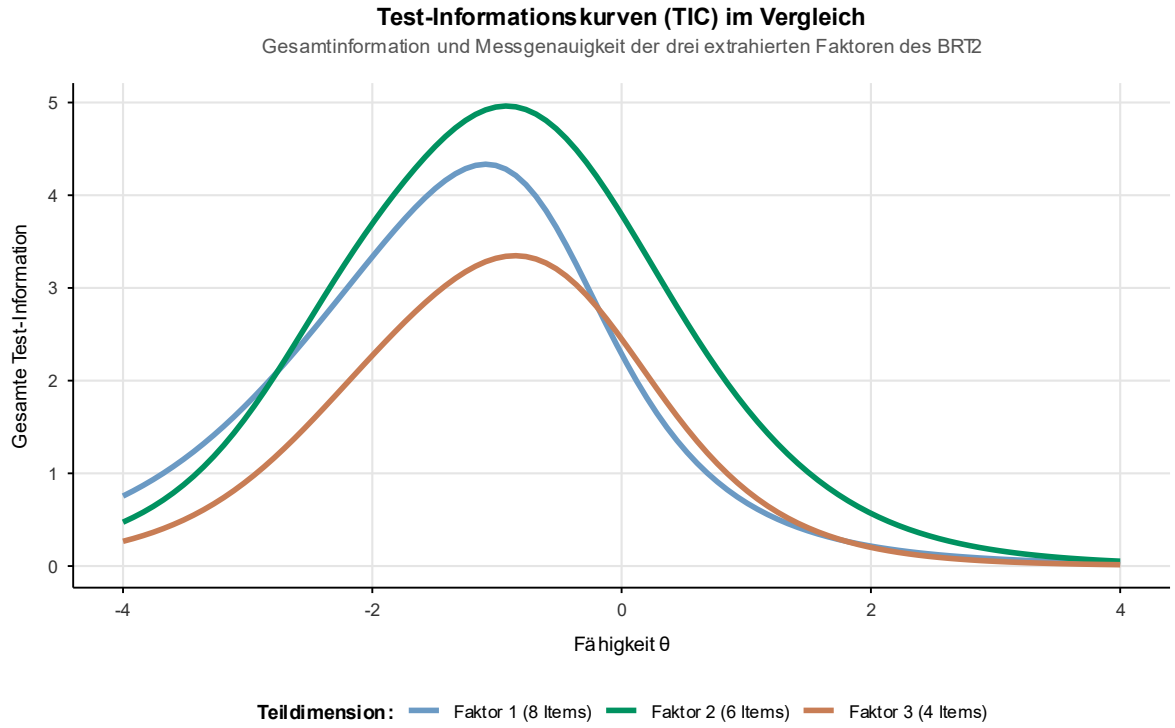


Abbildung 15 Test-Informationskurven (TIC) der drei Faktoren im direkten Vergleich

Demgegenüber bleibt Faktor 1 trotz einer maximalen Skalenlänge von acht Items in der Gesamtinformation darunter (maximaler Informationsertrag $I(\theta) \approx 4{,}3$, empirische Reliabilität von 0,647). Diese relative Einbuße ist direkt auf den mathematischen Dämpfungseffekt der flachen, informationsarmen Aufgaben A1, A4 und A5 zurückzuführen, welche die akkumulierte Skalenpräzision mindern.

Faktor 3 bündelt seine im Vergleich geringe Gesamtinformation (maximaler Informationsertrag $I(\theta) \approx 3{,}3$, empirische Reliabilität von 0,645) infolge der Kürze von nur vier Items dennoch hochgradig fokussiert und spitz im Bereich um $\theta = -1$, was auf eine sehr homogene Schwierigkeitsabstimmung der verbliebenen Aufgaben hindeutet.

4.3.4 Analyse ungleicher Aufgabenfunktionen (Differential Item Functioning)

Zur Überprüfung der geschlechtsspezifischen Messinvarianz des BRT-2 wurde eine Analyse des Differential Item Functioning (DIF) für die geschlechtsbereinigte Stichprobe ($N = 467$) durchgeführt. Die Parameterschätzung im Rahmen des erweiterten Generalized Partial Credit Models (GPCM) erreichte für alle 18 Testaufgaben eine vollständige numerische Konvergenz.

4.3.4.1 Tabellarische Übersicht des statistischen DIF-Screenings

Das inferenzstatistische Screening über die gesamte Itemmatrix identifiziert unter konsequenter Berücksichtigung der Benjamini-Hochberg-Adjustierung exakt 4 der 18 Testaufgaben mit einem manifesten geschlechtsspezifischen DIF ($p_{adj} < 0{,}05$). Die verbleibenden 14 Aufgaben erweisen sich nach der mathematischen Kontrolle der Alpha-Fehler-Inflation als vollständig messinvariant. Die statistischen Kennwerte der auffälligen Indikatoren sind in Tabelle 13 zusammengefasst.

Tabelle 13 Statistische Kennwerte der Items mit signifikantem Differential Item Functioning

Item	Gruppe	Modellkonvergenz	ΔAIC	ΔBIC	p-Wert	Adj. p-Wert
A1	Jungen, Mädchen	TRUE	-8,861	-0,568	0,002	0,015
A8	Jungen, Mädchen	TRUE	-6,544	1,749	0,005	0,031
A9	Jungen, Mädchen	TRUE	-5,619	6,820	0,009	0,040
A17	Jungen, Mädchen	TRUE	-17,024	-4,585	0,000	0,001

Anmerkung. N = 467. Signifikanzprüfung auf Basis der adjustierten Wahrscheinlichkeitswerte (Adj. p-Wert) nach Benjamini & Hochberg. Die vollständige Tabelle aller 18 Items befindet sich im Anhang H

Die mathematische Relevanzprüfung verdeutlicht den massiven Effekt der Signifikanzadjustierung. Aufgaben, die im unkorrigierten Zustand eine scheinbare Asymmetrie aufweisen, namentlich die Aufgaben A10 (p = 0,020), A13 (p=0,036) und A14 (p=0,044), überschreiten nach der prozeduralen Kontrolle die kritische Irrtumswahrscheinlichkeit deutlich ($p_{adj} > 0{,}05$). Diese Aufgaben werden somit als methodisch fair eingestuft. Als hochgradig veränderliche Kernindikatoren verbleiben hingegen die Aufgaben A1, A8, A9 und das Item A17 ($p_{adj} < 0{,}05$).

4.3.4.2 Qualitative Richtungsanalyse über erwartete Aufgaben-Scores

Die qualitative Inspektion der erwarteten Aufgaben-Scores (Abbildung 16) erlaubt eine Differenzierung der detektierten Effekte in uniforme und nicht-uniforme Verlaufsformen.

Ein deutlicher, über weite Teile des Leistungsspektrums stabiler Vorteil für die weiblichen Lernenden zeigt sich bei Aufgabe A8. Hier liegt die Leistungskurve der Mädchen trennscharf über der Kurve der Jungen. Bei der von Grund auf dichotomen Aufgabe A1 übertreffen die Schülerinnen die Schüler ab einem Fähigkeitswert von $\theta \approx -2{,}3$ ebenfalls und dominieren den gesamten durchschnittlichen sowie überdurchschnittlichen Leistungsbereich.

Ein mathematisch besonders stark ausgeprägter, uniformer Effekt zugunsten der männlichen Lernenden zeigt sich bei Aufgabe A17. Hier verläuft die Leistungskurve der Jungen

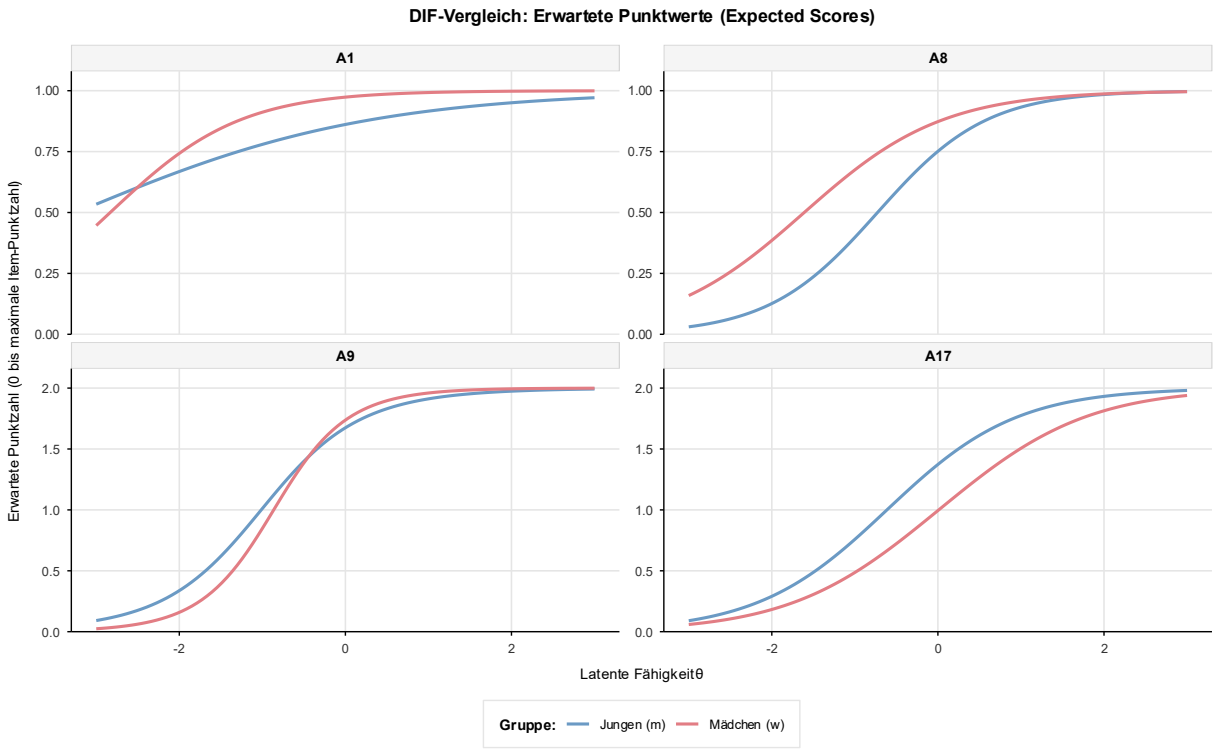


Abbildung 16 Erwartete Aufgaben-Scores im Geschlechtervergleich für die Items mit signifikantem DIF.

über das gesamte latente Fähigkeitsspektrum hinweg signifikant oberhalb der Kurve der Mädchen. Jungen erzielen hier bei völlig identischer Merkmalsausprägung eine systematisch höhere Punktwertserwartung.

Ein manifestes nicht-uniformes Verhalten weist ausschließlich die Aufgabe A9 auf. Bei diesem Item besitzen Jungen im unterdurchschnittlichen Leistungsbereich ($\theta < -0{,}5$) einen deutlichen Vorteil, bevor die geschlechtsspezifischen Funktionsgraphen sich schneiden und sich der Effekt im überdurchschnittlichen Fähigkeitssegment zugunsten der Mädchen invertiert.

4.3.4.3 Kategoriale Tiefenanalyse über Antwort-Wahrscheinlichkeiten

Die kategoriale Zerlegung der Antwortwahrscheinlichkeiten für das Erreichen der einzelnen Teilpunkte (Abbildung 17) legt die mathematischen Ursachen der gefundenen Leistungsdriften offen.

Da die Aufgaben A1 und A8 ein dichotomes Format aufweisen, existieren hier definitionsgemäß nur die Antwortstufen für 0 Punkte und 1 Punkt. Bei Aufgabe A8 wird das uniforme DIF dadurch verursacht, dass Jungen (durchgezogene rote Linie) über das gesamte Spektrum eine deutlich höhere Wahrscheinlichkeit aufweisen, 0 Punkte zu erzielen, während die Wahrscheinlichkeit für den korrekten Lösungsschritt (1 Punkt, gelbe Linie) bei den Mädchen (gestrichelte Linie) signifikant früher und steiler ansteigt.

Bei Aufgabe A17 zeigt sich die kategoriale Ursache des Effekts zugunsten der Jungen. Die Wahrscheinlichkeitskurve für das Erreichen der maximalen 2 Punkte (grüne Linie) ist bei den Jungen (durchgezogen) massiv nach links in Richtung geringer Fähigkeitsausprägun-

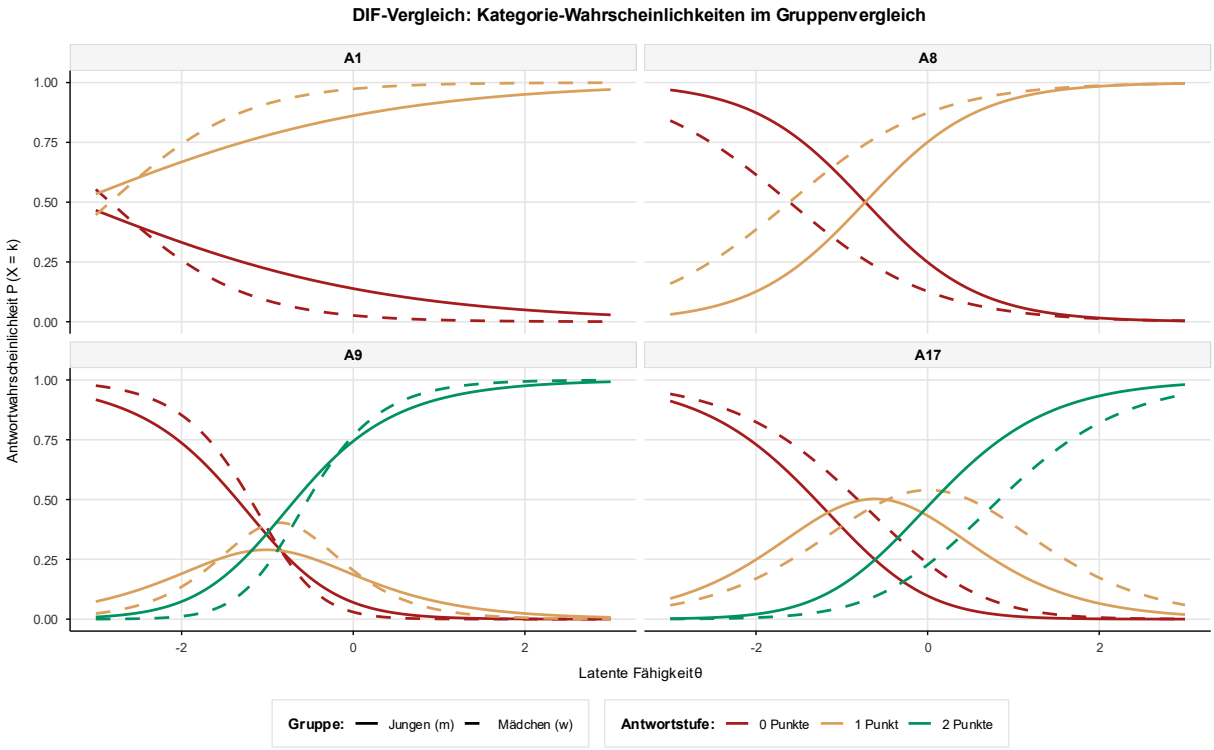


Abbildung 17 Kategorie-Antwortwahrscheinlichkeiten im Geschlechtervergleich für die auffälligen DIF-Items.

gen verschoben. Mädchen verbleiben im direkten Gruppenvergleich über einen breiten Fähigkeitsbereich in der niedrigsten Bepunktungskategorie (gestrichelte rote Linie für 0 Punkte).

Der Achsenkreuzungspunkt bei Aufgabe A9 wird auf kategorialer Ebene dadurch generiert, dass Jungen im unteren Segment ($\theta < -1{,}0$) eine deutlich stabilere Wahrscheinlichkeit aufweisen, die mittlere Teilpunktekategorie von 1 Punkt (durchgezogene gelbe Linie) zu erreichen. Erst im positiven Fähigkeitsbereich ziehen die Mädchen bei der Wahrscheinlichkeit für die vollen 2 Punkte (gestrichelte grüne Linie) gleich und übertreffen die Jungen im weiteren Kurvenverlauf hinsichtlich der Wahrscheinlichkeit des maximalen Punktgewinns. Die vollständigen Übersichten aller Kategorie-Antwortwahrscheinlichkeiten der übrigen, messinvarianten Items sind in Anhang H dokumentiert.

4.3.5 Verteilungsanalysen und Gruppenvergleich der latente Personenparameter

Die aus den Skalierungen des Generalized Partial Credit Models (GPCM) extrahierten Personenparameter (θ-Werte) bilden die Grundlage für die nachfolgenden Analysen der empirischen Kompetenzprofile. Die Berechnungen beschränken sich auf die geschlechtsbereinigte Teilstichprobe (N = 467).

4.3.5.1 Interindividuelle Differenzen: Geschlechtsspezifische Kompetenzprofile

Die inferenzstatistischen Kennwerte der geschlechtsspezifischen Rangsummenvergleiche über die drei mathematischen Dimensionen des BRT-2 sind in Tabelle 14 zusammen-

gefasst. Die korrespondierende Streuung und Verteilung der extrahierten Personenfähigkeiten wird parallel in Abbildung 18 grafisch visualisiert.

Tabelle 14 Kennwerte der Rangsummenvergleiche (Wilcoxon-Rangsummentest) der latenten Fähigkeiten nach Geschlecht.

Dimension	z-Wert	Effektstärke (*r*)	*p*-Wert
Faktor 1	2,059	0,095	0,0396
Faktor 2	6,226	0,288	< 0,001
Faktor 3	1,255	0,058	0,2096

Anmerkung. N = 467. Die latente Personenfähigkeit (θ) beziehen sich auf die EAP-Schätzung aus dem GPCM. Die Faktoren entsprechen den Dimensionen des extrahierten EFA-Modells.

Die synoptische Betrachtung der inferenzstatistischen Parameter und der grafischen Verteilungscharakteristika offenbart ein hochgradig differenziertes Bild über die drei Kompetenzbereiche des Testinstruments.

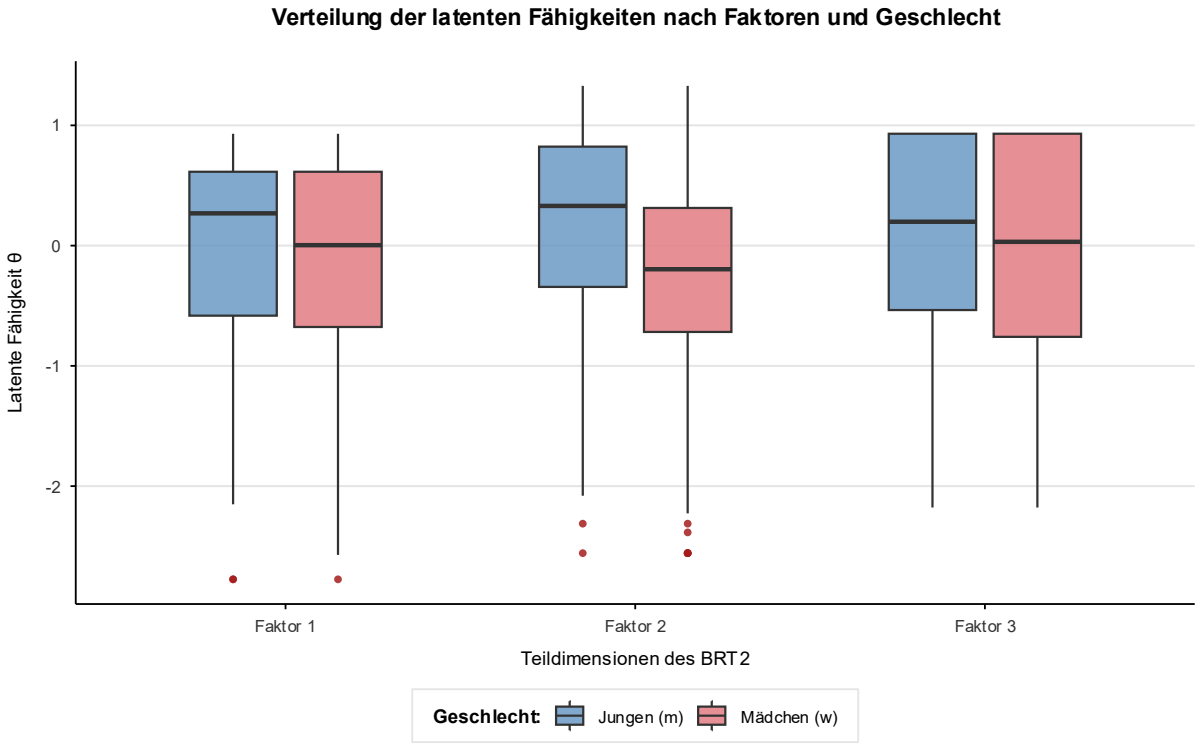


Abbildung 18 Verteilung der latenten Fähigkeiten nach Faktoren und Geschlecht.

Für den ersten Faktor zeigt sich ein statistisch signifikanter Unterschied im mittleren Leistungsniveau ($z = 2{,}059$, $p = 0{,}0396$). Die korrespondierende Effektstärke fällt mit $r = 0{,}095$ jedoch gering aus, was auf eine minimale praktische Relevanz im diagnostischen Alltag hindeutet, wobei männliche Lernende geringfügig höhere Rangplätze einnehmen. Diese kleine Effektdifferenz wird durch die visuelle Struktur in Abbildung 18 direkt untermauert. Die Boxplots der Jungen und Mädchen zeigen stark überlappende Interquartilsabstände (IQR), wobei der Median bei den Jungen etwas höher liegt als der der Mädchen. Insgesamt ist der gesamte Verteilungsblock der Jungen somit etwas höher als jener der Mädchen angesiedelt. Am unteren Ende der Verteilung (unterhalb von $\theta = -2{,}5$)

ist für beide Geschlechter jeweils nur ein einzelner, isolierter Extremwert als Ausreißerpunkt identifizierbar, was die hohe Verteilungshomogenität auf dieser Dimension bestätigt.

Im Gegensatz dazu steht ein hochgradig signifikanter und substanzieller Geschlechtseffekt auf dem zweiten Faktor ($z = 6{,}226,\ p < 0{,}001$). Mit einer mittleren Effektstärke von $r = 0{,}288$ zeigt sich auf dieser Dimension ein manifester Kompetenzunterschied zugunsten der männlichen Probanden, welcher sich auch in den numerischen Gruppenmittelwerten widerspiegelt ($\overline{\theta_m} = 0{,}186$ und $\overline{\theta_w} = -0{,}295$). Diese ausgeprägte Lagedifferenz tritt im zweiten Panel der Abbildung 18 plastisch hervor. Der gesamte Verteilungsblock der Jungen ist sichtbar nach oben in den positiven Fähigkeitsbereich verschoben, wobei ihr Gruppenmedian knapp unterhalb von $\theta = 0{,}5$ verortet ist. Demgegenüber ist der Median der Mädchen deutlich im negativen Bereich (nach $\theta = -0{,}3$) abgesenkt. Die identische Struktur der beiden Teilstichproben dokumentiert sich lediglich in den unteren Extrembereichen, wo beide Gruppen im niedrigsten Leistungssegment (unterhalb von $\theta = -2{,}0$) vereinzelte, parallel verlaufende statistische Ausreißer aufweisen.

Für den dritten Faktor lässt sich weder inferenzstatistisch ($z = 1{,}255,\ p = 0{,}2096$) noch visuell ein relevanter Unterschied zwischen den Geschlechtergruppen nachweisen. Die marginale Effektstärke von r = 0,058 belegt eine vollständige psychometrische Äquivalenz. Die Whisker-Variationen und die oberen Quartile der beiden Boxplots sind über beide Geschlechtergruppen hinweg nahezu deckungsgleich. Die Mediane sowie die unteren Quartile der Jungen liegen minimal höher als die der Mädchen. Im Gegensatz zu den Faktoren 1 und 2 treten im sichtbaren Bereich dieses Faktors keine markanten Ausreißerpunkte im unteren Fähigkeitsspektrum auf, was auf eine harmonische und geschlechtsunabhängige Erfassung der Probandenleistung hindeutet.

4.3.5.2 Intraindividuelle Differenzen: Dimensionsübergreifender Leistungsvergleich

Der dimensionsübergreifende Vergleich der individuellen Fähigkeitsausprägungen über die drei mathematischen Faktoren hinweg liefert Kennwerte zur relativen Homogenität der Skalierung.

Der globale Rangsummenvergleich über die unbereinigte Gesamtstichprobe hinweg verfehlt die kritische Signifikanzgrenze ($\chi^2 = 5{,}7003,\ df = 2,\ p = 0{,}0578$). Zur Kontrolle potenzieller Verzerrungen durch unvollständige Datensätze zeigt auch der analoge Vergleich innerhalb der geschlechtsbereinigten Teilstichprobe kein statistisch signifikantes Ergebnis ($\chi^2 = 4{,}8951,\ df = 2,\ p = 0{,}0865$).

Es liegen somit keine systematischen Rangplatzverschiebungen der individuellen Fähigkeiten zwischen den drei mathematischen Teilbereichen vor. Die Ausprägungen der Personenfähigkeiten verhalten sich innerhalb der Lernenden über die Faktoren hinweg homogen, was den drei Faktoren des BRT-2-EFA-Modells ein ausgewogenes Skalenniveau attestiert.

4.3.6 Untersuchung der dimensionalen Architektur mittels Bifaktor-Modellierung

Den testtheoretischen und mathematischen Abschluss der modellbasierten Analysen bildet die empirische Bestimmung von globalen und spezifischen Varianzanteilen des Bayreuther Rechentests (BRT-2) über eine explizite Strukturprüfung mittels Bifaktor-Modellen. Diese abschließende Modellierung schließt an zentrale Befunde der vorangegangenen Abschnitte an, welche kumulativ auf die Existenz eines dominanten Generalfaktors hindeuten. Zum einen dokumentierten sowohl die im Vorfeld durchgeführte Korrelationsanalyse der manifesten Subskalen (Kapitel 4.1.2), deren Koeffizienten sich in einem Korridor von $r = 0{,}45$ bis $r = 0{,}57$ bewegen, als auch die Interkorrelationen der Faktoren aus der EFA-Modellierung (Kapitel 4.3.2), die sich im Bereich zwischen $r = 0{,}658$ bis $r = 0{,}705$ bewegen, moderate bis starke positive Zusammenhänge. Beide Analyseebenen verweisen somit auf eine beträchtliche gemeinsame Varianzbasis. Zum anderen signalisierte der im Rahmen der Reliabilitätsanalyse (Kapitel 4.2.1) ermittelte hohe Ausprägungswert von McDonald's hierarchischem Omega ($\omega_h = 0{,}72$) eine ausgeprägte relationale Kohäsion über alle 18 Aufgaben hinweg. Die Bifaktor-Modellierung erlaubt es vor diesem empirischen Hintergrund, diese gemeinsame Varianzbasis auf Itemebene orthogonal in einen übergreifenden Generalfaktor (G-Faktor) und drei spezifische Gruppenfaktoren (Inhaltsdimensionen) zu zerlegen, um den diagnostischen Mehrwert der einzelnen Facetten komparativ zu bewerten.

4.3.6.1 Inferenzstatistischer Modellvergleich und globale Modellgüte

Für die Evaluation der optimalen dimensionalen Architektur des BRT-2 wurden fünf komplementäre Modellspezifikationen des Generalized Partial Credit Models (GPCM) berechnet. Tabelle 15 fasst die zugehörigen Informationskriterien (AIC, BIC), die Log-Likelihood-Werte sowie die C_2-basierten absoluten und inkrementellen Fit-Indizes ($RMSEA$, TLI, CFI) synoptisch zusammen.

Tabelle 15 Synoptischer Modellvergleich und globale Fit-Indizes der GPCM- und Bifaktor-Strukturen (N = 574).

Modell	AIC	BIC	Log.-Lik.	RMSEA	TLI	CFI
GPCM-1F	15374,81	15627,27	-7629,407	0,047	0,968	0,972
GPCM-Theorie	15889,37	16141,82	-7886,684	0,071	0,927	0,936
GPCM-EFA	15728,91	15981,37	-7806,457	0,062	0,945	0,952
Bifaktor-Theorie	15350,06	15680,86	-7599,032	0,038	0,979	0,984
Bifaktor-EFA	15302,98	15633,78	-7575,490	0,022	0,993	0,995

Anmerkung. N = 574. Sämtliche Fit-Indizes (RMSEA, TLI, CFI) wurden einheitlich über die Limited-Information-M_2-Statistik (Typ C2) extrahiert. Die GPCM-Modelle wurden mit dem Standard-EM-Schätzer und die Bifaktor-Modelle mit dem QMCEM-Schätzer geschätzt.

Die in Tabelle 15 dokumentierten Kennwerte verweisen auf eine eindeutige empirische Rangordnung zwischen den konkurrierenden Modellstrukturen. Das theoriegeleitete Dreifaktormodell liefert mit einem AIC von 15889,37 und einem BIC von 16141,82 die ungünstigsten Anpassungsmaße des gesamten Vergleichs. Eine graduelle Modellverbesse-

rung zeigt sich beim dreidimensionalen, explorativ spezifizierten Modell. Eine psychometrische Besonderheit des Datensatzes offenbart sich jedoch beim Blick auf das rein eindimensionale Basismodell (GPCM-1F). Dieses übertrifft die beiden mehrdimensionalen Standardmodelle ohne Generalfaktor in allen Informationskriterien und Fit-Indizes deutlich ($AIC = 15374{,}81$, $BIC = 15627{,}27$, $RMSEA = 0{,}047$). Dass eine restriktive, eindimensionale Struktur die Daten besser repräsentiert als die theoretisch postulierten Einzelfaktoren, liefert ein starkes empirisches Argument für die massive Dominanz einer übergreifenden, allgemeinen Rechenkompetenz im BRT-2.

Die beiden hierarchischen Bifaktor-Strukturen, welche diese allgemeine Kompetenz explizit modellieren, weisen eine nochmals gesteigerte Erklärungskraft auf. Das theoriegeleitete Bifaktormodell (Bifaktor-Theorie) reduziert die Informationskriterien im Vergleich zur eindimensionalen Baseline weiter ($AIC = 15350{,}06$, $BIC = 15680{,}86$). Diese Modellarchitektur stößt jedoch an ihre Grenzen, sobald sie einer zusätzlichen, strengeren Strukturprüfung mittels einer konfirmatorischen Faktorenanalyse (CFA) ausgesetzt wird. Die rein theoriegeleiteten Null-Restriktionen widersprechen den realen Antwortmustern so stark, dass die CFA in einen vollständigen numerischen Konvergenzabbruch läuft.

Die über alle Indizes hinweg überlegene und auch im CFA-Raum stabil konvergierende Datenpassung wird folglich durch das Bifaktor-EFA-Modell erzielt. Dieses Modell behält die restriktive Grundstruktur bei, ordnet die Items jedoch auf Basis einer vorab durchgeführten explorativen Faktorenanalyse (Kapitel 4.3.2) datengetrieben den spezifischen Dimensionen zu. Es generiert das absolute globale Minimum des Informationsraums ($AIC = 15302{,}98$, $BIC = 15633{,}78$) und maximiert die Plausibilität ($Log\text{-}Likelihood = -7575{,}490$) bei einer makellosen Approximationsgüte ($RMSEA = 0{,}022$, $TLI = 0{,}993$, $CFI = 0{,}995$).

4.3.6.2 Faktorladungsarchitektur und Varianzzerlegung

Die nachfolgende Analyse kontrastiert die Parameter- und Varianzstrukturen der beiden komplementären Bifaktor-Ansätze. Die standardisierten Faktorladungen auf dem globalen Generalfaktor (G), die Ladungen auf den drei spezifischen Gruppenfaktoren (F1 bis F3) sowie die resultierenden Kommunalitäten (h^2) der 18 Testaufgaben sind für das theoriegeleitete Bifaktormodell in Tabelle 16 und für das empirisch hergeleitete Modell in Tabelle 17 dargelegt.

Der Vergleich zeigt, dass beide Modellierungen eine erhebliche Dominanz des Generalfaktors (G) abbilden. Im theoriegeleiteten Modell (Tabelle 16) verbucht dieser mit einem Sum-of-Square-Loadings (SS-Loading) von 6,036 einen Anteil von 33,5 % der Gesamtvarianz für sich. Aufgrund des hohen Explained-Common-Variance-Werts (ECV) von 0,76 dominiert der Generalfaktor die gemeinsame Varianz, sodass die Daten als essenziell eindimensional angenommen werden können. Das datengetriebene Bifaktormodell (EFA)

Tabelle 16 Standardisierte Faktorladungen und Kommunalitäten im Bifaktormodell (Theorie)

Item	Generalfaktor (G)	Spezifisch F1	Spezifisch F2	Spezifisch F3	h^2
A1	0,331	0,275			0,185
A2	0,584	0,043			0,343
A3	0,548	0,307			0,394
A4	0,289	0,535			0,369
A5	0,502	0,353			0,377
A6	0,599	0,161			0,385
A7	0,469		0,174		0,250
A8	0,520		0,187		0,306
A9	0,695		0,226		0,534
A10	0,569		0,627		0,717
A11	0,586		0,232		0,397
A12	0,614		0,147		0,398
A13	0,652			-0,290	0,509
A14	0,566			0,607	0,689
A15	0,702			0,312	0,590
A16	0,699			-0,221	0,537
A17	0,585			0,336	0,455
A18	0,709			0,050	0,505
SS loadings	6,036	0,608	0,584	0,714	7,942
Proportion Var	0,335	0,034	0,032	0,040	0,441

Tabelle 17 Standardisierte Faktorladungen und Kommunalitäten im Bifaktormodell (EFA)

Item	Generalfaktor (G)	Spezifisch F1	Spezifisch F2	Spezifisch F3	h^2
A1	0,329	0,340			0,224
A2	0,568			0,254	0,388
A3	0,553	0,472			0,529
A4	0,315	0,343			0,217
A5	0,520	0,187			0,305
A6	0,656			-0,002	0,431
A7	0,504	0,251			0,317
A8	0,535	0,280			0,365
A9	0,765	0,144			0,607
A10	0,658	0,083			0,439
A11	0,585		0,198		0,381
A12	0,576		0,347		0,452
A13	0,607			0,320	0,471
A14	0,468		0,559		0,531
A15	0,614		0,461		0,590
A16	0,634			0,621	0,788
A17	0,476		0,550		0,529
A18	0,660		0,271		0,509
SS loadings	5,8	0,661	1,059	0,552	8,072
Proportion Var	0,322	0,037	0,059	0,031	0,449

(Tabelle 17) repliziert diese Struktur auf einem ähnlich hohen Niveau mit einem SS-Loading von 5,8, was einem Gesamtvarianzanteil von 32,2 % und einem ECV von 0,7185 entspricht. Es kann also auch hier angenommen werden, dass die Daten essenziell eindimensional sind. Die verbleibenden spezifischen Gruppenfaktoren weisen in beiden Modellen ausnahmslos geringe Varianzen im einstelligen Prozentbereich auf.

Die vertiefte Betrachtung der Tabelle 16 und Tabelle 17 demaskiert jedoch die Schwachstelle der rein theoretischen Blockbildung. Während das Modell in Tabelle 16 die Items A1 bis A6 starr auf dem ersten Faktor fixiert, bricht die empirische Zuordnung in Tabelle 17 diese Struktur auf. Das Item A2 wandert mit einer nennenswerten Ladung von 0,254 auf den dritten spezifischen Faktor, während das Item A6 mit einer Ladung von -0,002 fast vollständig auf dem dritten Inhaltsfaktor verschwindet. Ebenso zeigt sich eine veränderte Clusterung im Bereich der hinteren Aufgabenblöcke, bei denen sich insbesondere die Items A14 (0,559) und A17 (0,550) anstelle des theoretisch postulierten dritten Faktors mathematisch dem zweiten spezifischen Faktor zuordnen. Dass das Modell mit dieser datengetriebenen Zuordnung (Tabelle 17) problemlos in einer CFA konvergiert, belegt, dass die gesetzten Null-Restriktionen hier im Einklang mit den tatsächlichen Kovarianzstrukturen der Stichprobe stehen, während die rein theoretischen Restriktionen an der empirischen Realität scheitern.

Diese hierarchische Verteilung der Erklärungskraft wird im strukturellen Pfaddiagramm (Abbildung 19) exemplarisch veranschaulicht.

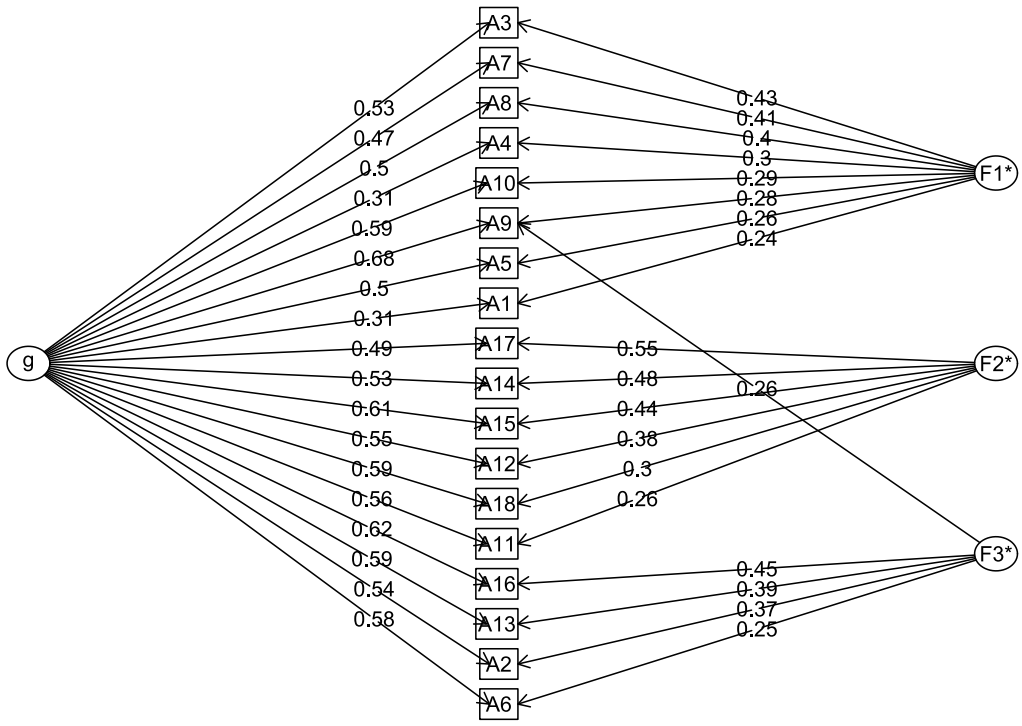


Abbildung 19 Strukturelles Pfaddiagramm der orthogonalen Varianzzerlegung

Darin zeigt sich prägnant, dass der übergeordnete Faktor G als zentraler Prädiktor simultan auf den gesamten Item-Pool einwirkt und die primäre mathematische Kohäsion des Tests abbildet, während die drei Gruppenfaktoren (F1* bis F3*) plangemäß als voneinander isolierte Restvarianzen agieren.

Die Gesamtvarianzaufklärung kumuliert sich im theoriegeleiteten Modell auf 44,1 % und im datengetriebenen Modell auf 44,9 %.

4.3.6.3 Modellbasierte Informations- und Präzisionsanalysen

Die Messgenauigkeit des Generalfaktors sowie der spezifischen Faktoren über das gesamte latente Fähigkeitskontinuum hinweg wird nachfolgend anhand der softwaregestützten Test- und Item-Informationskurven ausgewertet. Hierbei sind in Abbildung 20 die Testinformationskurven des Generalfaktors und der spezifischen Gruppenfaktoren dargestellt.

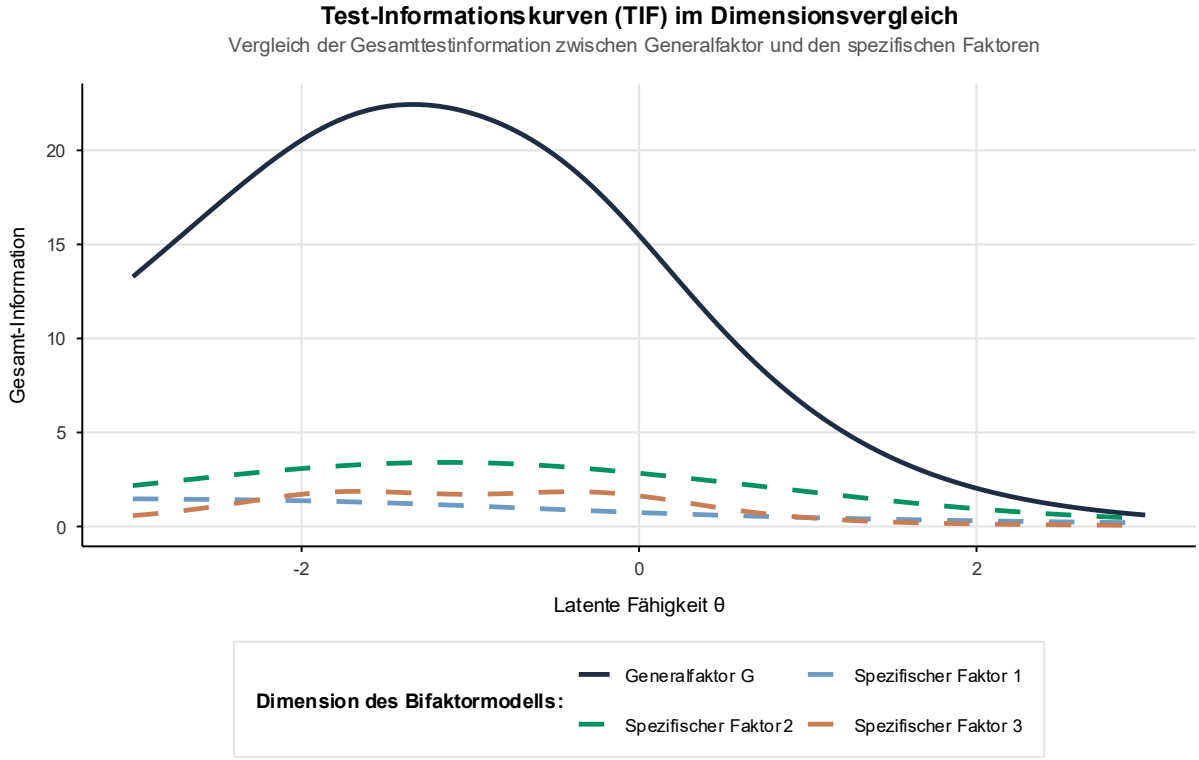


Abbildung 20 Test-Informationskurven (TIF) im Dimensionsvergleich.

Die Analyse der globalen Test-Informationskurven (Abbildung 20) bestätigt die messtechnische Dominanz des Generalfaktors über den gesamten Fähigkeitsverlauf hinweg. Die Test-Informationskurve des G-Faktors (durchgezogene Linie) zeigt einen steilen, glockenförmigen Verlauf, der im unterdurchschnittlichen Leistungssegment des Kontinuums zwischen $\theta = -2{,}5$ und $\theta = -0{,}5$ ein massives Informationsmaximum von über 22 Einheiten erreicht. Dies attestiert dem BRT-2 eine prägnante Messgenauigkeit exakt in dem Skalenbereich, der für die zuverlässige Screening-Diagnose von Rechenschwäche und mathematischen Entwicklungsverzögerungen im Grundschulalter gefordert wird. Die spezifi-

schen Inhaltsfaktoren 1, 2 und 3 (gestrichelte Linien) bewegen sich im direkten Vergleich konstant auf einem niedrigen Niveau zwischen 0 und 5 Informationseinheiten.

Abbildung 21 stellt die Item-Informationskurven (IIC) auf dem Generalfaktor anhand zweier unterschiedlicher mathematischer Darstellungsvarianten komparativ gegenüber.

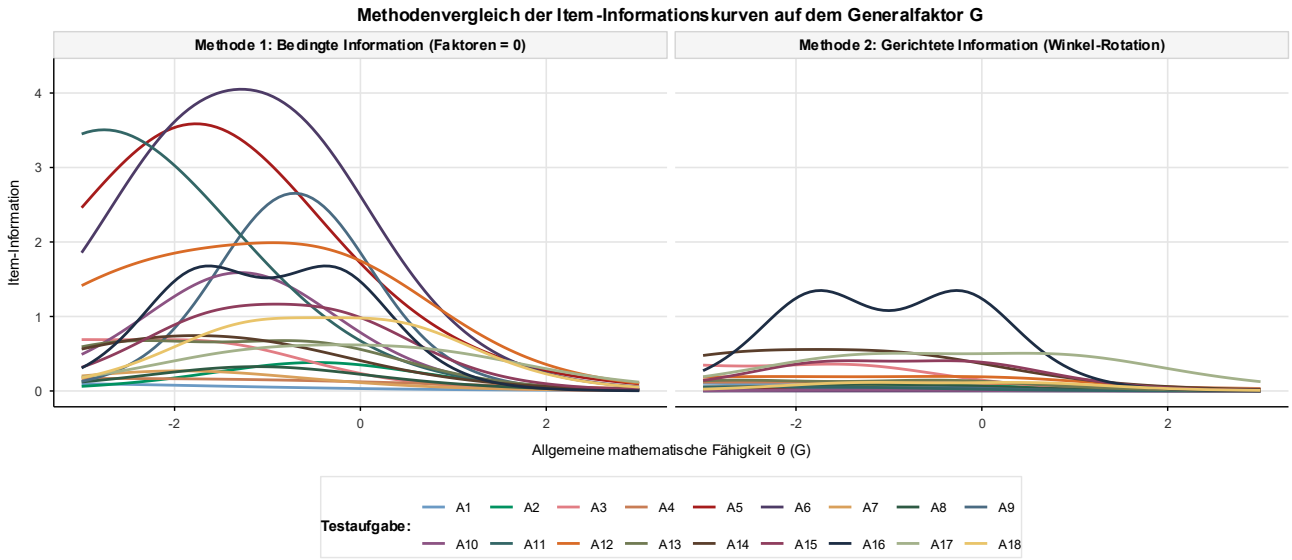


Abbildung 21 Methodenvergleich der Item-Informationskurven auf dem Generalfaktor G.

Im linken Panel (Methode 1) wird die bedingte Item-Information abgebildet, bei welcher die Ausprägung der spezifischen Gruppenfaktoren mathematisch auf Null fixiert wurden. Unter dieser Bedingung zeigt sich eine dichte Kumulation ausgeprägter Kurvenstrukturen im negativen Merkmalsraum. Die Aufgaben mit den höchsten Diskriminationsparametern, insbesondere das Item A9 sowie die Items A6, A10 und A18, bilden deutliche Maxima im Bereich zwischen $\theta = -2{,}5$ und $\theta = 0{,}0$ ab und steuern ein hohes Maß an punktueller Information bei.

Demgegenüber illustriert das rechte Panel (Methode 2) die gerichtete multidimensionale Information unter Berücksichtigung der multidimensionalen Achsenrotation. Die Informationskurven der überwiegenden Mehrheit der 18 Testaufgaben flachen unter dieser Projektionsbedingung vollständig ab und kollabieren nahe an die empirische Nulllinie, da ihre Messpräzision vollständig in den Generalfaktor übergeht. Die markante Ausnahme in diesem Panel ist das Item A16 (Dunkelblau), welches sich dem Kollaps entzieht und ein stabiles, bivariates Doppelhöcker-Profil mit Informationsmaxima bei $\theta \approx -1{,}7$ und $\theta \approx -0{,}3$ konserviert.

Die isolierte spezifische Eigeninformation der verbliebenen Gruppen wird schließlich in Abbildung 22 visualisiert. Deren dreiteilige Panelstruktur spiegelt die Proportionen auf Itemebene präzise wider.

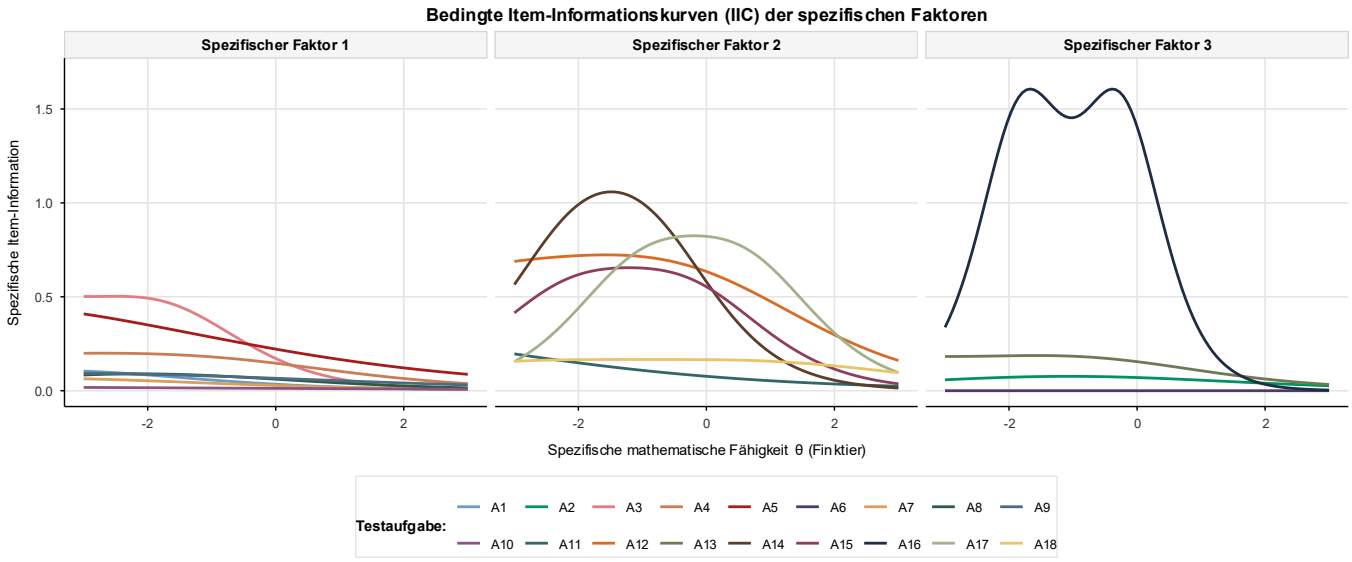


Abbildung 22 Bedingte Item-Informationskurven (IIC) der spezifischen Faktoren.

Im Bereich des spezifischen Faktors 1 (linkes Panel) verharren alle Itemkurven im Einklang mit der geringen Proportionalvarianz in einem psychometrisch unbedeutenden Korridor unterhalb von 0,5 Informationseinheiten, wobei lediglich die Items A3 und A5 flache Verläufe im unterdurchschnittlichen Leistungssegment zeigen. Der spezifische Faktor 2 (mittleres Panel) absorbiert eine moderate, lokal begrenzte Aufgabeninformation. Die Items A14, A15 und A17 bilden eigenständige Kurvenbögen aus, deren Scheitelpunkte im Bereich zwischen $\theta = -2{,}0$ und $\theta = 0{,}0$ Werte zwischen 0,5 und 1,1 Informationseinheiten erzielen und die Existenz einer inhaltlich interpretierbaren Restdimension stützen.

Ein extremes Verteilungsbild zeigt sich schließlich auf der Dimension des spezifischen Faktors 3 (rechtes Panel). Während fast alle Aufgaben eine sehr geringe Information aufweisen, dominiert das Item A16 (dunkelblaue Linie) diese spezifische Dimension singulär mit einer massiven, doppelgipfligen Informationskurve, die in der Spitze Werte von über 1,6 Einheiten erreicht. Dieser mathematische Befund korrespondiert exakt mit der hohen spezifischen Faktorladung von 0,621 aus Tabelle 17 und belegt im Zusammenspiel mit dem Richtungsvergleich, dass die psychometrische Substanz des dritten Faktors primär durch die spezifische Eigenvarianz dieser einzelnen Aufgabe getragen wird.

5 Diskussion

In diesem Kapitel sollen die Ergebnisse zusammengeführt werden und hierbei die forschungsleitenden Fragen beantwortet werden. Zunächst liegt hierbei, in Kapitel 5.1, der Fokus auf der Beantwortung der forschungsleitenden Fragen sowie der beiden Forschungsfragen bezüglich der Struktur und der diagnostischen Eignung. Im Anschluss soll in Kapitel 5.2 auf die methodischen Limitationen dieser Arbeit eingegangen werden. Zum Abschluss dieses Kapitels (Kapitel 5.3) werden anhand der gewonnenen Ergebnisse noch Möglichkeiten zur Weiterentwicklung des Bayreuther-Rechentests ausgeführt.

5.1 Integration und Interpretation der zentralen Befunde

In diesem Kapitel sollen die beiden Forschungsfragen behandelt werden. Dabei wird in Kapitel 5.1.1 zunächst auf die forschungsleitenden Fragen, die mit der Strukturaufklärung zu tun haben, eingegangen und die Frage, ob sich die theoretische Dreiteilung der Kompetenzbereiche in den empirischen Daten widerspiegelt, beantwortet werden. Danach soll in Kapitel 5.1.2 die Frage, ob sich der Test als Instrument zur Diagnostik von Rechenschwäche eignet, mithilfe der forschungsleitenden Fragen beantwortet werden.

5.1.1 Strukturelle Validität und dimensionale Architektur

Anhand der Ergebnisse konnte ein starker Unterschied in den Leistungen festgestellt werden. Während es in den Bereichen *Verständnis für natürliche Zahlen* und *Verständnis für das Stellenwertsystem* keine signifikanten Unterschiede gab, waren die Ergebnisse im Bereich *Verständnis für Rechenoperationen* signifikant schwächer. Weiterhin finden sich mittlere bis starke positive Korrelationen zwischen den Kompetenzbereichen. Diese Ergebnisse lassen sich auch im Extremgruppenvergleich wiederfinden. So sind die besten 25 % zwar sehr homogen in ihrer Leistung, jedoch ist die Streuung im Kompetenzbereich *Verständnis für Rechenoperationen* etwas höher als bei den anderen beiden Teilbereichen. Bei den schwächsten 25 % liegt der Median dieses Kompetenzbereichs am tiefsten. Weiterhin sind ihre Leistungen deutlich heterogener als die der besten 25 %.

Bei einer explorativen Faktorenanalyse konnten festgestellt werden, dass der Test empirisch drei Faktoren aufweist, jedoch unterscheiden sich diese von den theoretisch angenommenen. Während ein Großteil der Aufgaben aus den Bereichen *Verständnis für natürliche Zahlen* und *Verständnis für das Stellenwertsystem* auf einen gemeinsamen Faktor laden, laden die rechenlastigeren Aufgaben aus den *Bereichen Verständnis für das Stellenwertsystem* und *Verständnis für Rechenoperationen* auf einem zweiten Faktor. Auf dem dritten Faktor laden lediglich vier Aufgaben.

Die Verschmelzung der Kompetenzbereiche für das *Verständnis von natürlichen Zahlen* und das Stellenwertsystem zu dem Faktor Zahlbegriffsentwicklung und dezimale Strukturkompetenz ist dadurch erklärbar, dass es sich bei der Entwicklung vom *Verständnis für natürliche Zahlen* und dem *Verständnis für das Stellenwertsystem* um integrative Ent-

wicklungsprozesse handelt und nicht um strikt getrennte Kompetenzen (Gerster & Schultz, 2007; Siegler et al., 2011). Der Faktor zwei ist weiterhin das *Verständnis für Rechenoperationen* und der dritte Faktor stellt ein relationales Zahl- und Operationsverständnis dar. Die Aufgaben auf diesem Faktor prüfen, wie gut Lernenden Beziehungen erfassen. Während bei Aufgabe 2 und bei Aufgabe 16 a) um quantitative relationale Vergleiche geht stehen in Aufgabe 6 die relationalen Zahlbeziehungen im Vordergrund. Bei den Aufgaben 13 und 16 steht weiterhin die relationale Übersetzung von Sachkontexten in Rechenoperationen im Vordergrund.

Aufgrund der hohen Korrelationen zwischen den Bereichen und des in Kapitel 4.3.6 erwähnten hohen McDonalds hierarchischem Omega liegt es nahe zu prüfen, ob es sich um ein Ein-Faktor-Modell handelt. Hierbei ergab der Modellvergleich der GPCM-Modelle einen besseren Fit für das Ein-Faktor-Modell als für die dreidimensionalen Modelle. Jedoch hatte das Ein-Faktor-Modell in der strengeren konfirmatorischen Faktorenanalyse die schlechtesten Fit-Indizes. Aufgrund der genannten Hinweise wurde dann noch geprüft, ob eine Bifaktor-Modellierung die empirische Struktur besser beschreibt. Hierbei konnte gezeigt werden, dass ein Bifaktor-Modell mit den spezifischen Faktoren aus der EFA die Struktur am besten beschreibt. Somit kann das Modell als essenziell eindimensional angenommen werden, wodurch die Bildung der Gesamtwerte erst sinnvoll ermöglicht wird.

5.1.2 Diagnostische Eignung und probabilistische Modellierung

Um die Frage zu klären, ob der Test ein gutes Instrument zur Diagnostik von Rechenschwäche ist, wurden zunächst die Gesamtleistung sowie die Leistungen in den einzelnen theoretischen Teilbereichen hinsichtlich Geschlechterunterschieden untersucht. Hierbei konnten sowohl bei der Gesamtleistung als auch beim *Verständnis für das Stellenwertsystem* schwache Effekte und im Bereich *Verständnis für Rechenoperationen* ein mittlerer Effekt gefunden werden. Hierbei haben Mädchen geringere Mediane oder stellen eine heterogenere Leistungsgruppe dar. Lediglich beim Verständnis für natürliche Zahlen wurde kein Geschlechterunterschied gefunden.

Bei Betrachtung der internen Konsistenz wird ersichtlich, dass diese nach McDonalds Omega über alle Kompetenzbereiche hinweg eine akzeptable bis gute Reliabilität aufweisen. Aufgrund der essenziellen Eindimensionalität des Tests wurde auch die Konsistenz des gesamten Testverfahrens ermittelt. Diese stellt eine exzellente Reliabilität für das Testverfahren dar. Anhand der Prüfung der Aufgabenschwierigkeit mittels des Schwierigkeitsindex konnten die besonders leichten und die schwereren Aufgaben ermittelt werden. Hierbei sind die drei besonders leichten Aufgaben die Aufgaben A1, A3 und A11, wobei letztere die leichteste Aufgabe des Tests ist. Die schwersten Aufgaben sind A2, A17 und A18, wobei A17 die schwerste Aufgabe des Tests darstellt.

Daraufhin wurden die bedingten Wahrscheinlichkeiten berechnet und als Netzwerkgrafik dargestellt. Anhand der Grafik kann man erkennen, dass die Aufgaben A2 und A17 am häufigsten falsch beantwortet werden unter der Voraussetzung, dass eine andere

Aufgabe zuvor falsch beantwortet wurde. Wenn Aufgabe A3 falsch beantwortet wurde, ist die Wahrscheinlichkeit groß, dass dann auch eine weitere Aufgabe falsch beantwortet wird. Die Aufgaben A1, A4 und A6 stehen hierbei isoliert.

Die Clusteranalyse ergibt sechs verschiedene Leistungstypen. Hierbei gibt es Lernende die konstant gute Leistungen erbringen, einen Teil der konstant schlechte Leistungen erbringen und 4 weitere Typen, die in einem oder zwei der theoretischen Kompetenzbereiche Schwächen zeigen.

Mittels der IRT-Analyse wird die Eignung der Aufgaben zur Diagnostik von Rechenschwäche untersucht. Hierbei wurden einige interessante Informationen erhalten. Zunächst wurde ersichtlich, dass bei den Aufgaben A5, A6 und A11 eine Inversion der Schwellenparameter vorliegt. Woraus resultiert, dass die mittleren Punktestufen stark komprimiert werden, wodurch die Aufgaben eher dichotom anstatt polytom funktionieren. Außerdem ist zu erkennen, dass die Aufgaben A1, A4 und A5 die geringste Information tragen.

Die Peaks der Informationskurven lassen sich alle im Bereich von $-3 \leq \theta \leq 0$ verorten, also im negativen Fähigkeitsbereich. Das bedeutet, dass dort der Standardmessfehler $SE(\theta)$ am kleinsten ist, was perfekt für ein Instrument zur Bestimmung von Rechenschwäche ist. Es liefert somit verlässlichere Daten bei leistungsschwächeren Lernenden.

Bei der Prüfung der Fairness des Tests offenbart sich ein massives Problem. Die DIF-Analyse bringt unter dem strengen Benjamini-Hochberg-Verfahren zur Signifikanzkorrektur ein bedenkliches Ergebnis zu tage. Vier Items weisen ein signifikantes Differential Item Functioning auf. Bei diesen vier Items besitzen Jungen und Mädchen bei identischer latenter Fähigkeit nicht dieselbe Wahrscheinlichkeit, diese Aufgaben zu lösen. In Aufgabe 17 ist hierbei ein uniformes DIF zugunsten der Jungen zu finden. In dieser Aufgabe geht es um die Subtraktion mit Zehnerübergang und sehr nahe beieinander liegenden Zahlen. Bei diesen Aufgaben haben die Jungen einen Vorteil, da sie eine signifikant höhere Tendenz aufweisen, intuitive, abkürzende und mathematisch flexiblere Strategien spontan anzuwenden (Gallagher et al., 2000). Neben diesem Vorgehen nutzen sie visuo-räumliche Repräsentationen, wodurch weniger kognitive Ressourcen verbraucht werden (Geary, 1996; Krinzinger, 2010). Mädchen hingegen neigen dazu, die im Unterricht erlernten, formalen Algorithmen regelkonform und schrittweise abzuarbeiten (Carr & Jessup, 1997; Gallagher & De Lisi, 1994).

Bei Aufgabe 8 liegt das uniforme DIF-Muster in umgekehrter Form vor. Hier haben die Mädchen den Jungen gegenüber einen Vorteil. Da die Aufgabe im Bereich des *Verständnisses für das Stellenwertsystem* liegt, ist dies zunächst verwunderlich, da hier nach wissenschaftlichen Erkenntnissen die Jungen besser abschneiden (Krinzinger, 2010). Der Vorteil in dieser Form lässt sich jedoch durch Unterschiede in der sequenziellen Verarbeitungsgenauigkeit sowie der Impulsivitätskontrolle bei Mädchen erklären. Sie besitzen im Bereich des strukturellen Zählens und der prozeduralen Ausführung nachweislich hocheffektive Kompetenzen (Barel & Tzischinsky, 2022; Carr &

Jessup, 1997), während Jungen dazu neigen, die grafische Struktur ganzheitlich und oberflächlich zu erfassen. Durch den zusätzlichen Zeitdruck übersehen sie Würfel häufiger oder zählen diese doppelt (Barel & Tzischinsky, 2022; Riley et al., 2016).

Eine ähnliche Begründung gilt auch für die nicht uniforme Aufgabe A1, bei der die Lernenden Finger abzählen müssen. Im niedrigen Fähigkeitsspektrum und darüber liegt hier der Vorteil zu Gunsten der Mädchen. Hierbei ist die Erklärung identisch zu der Erklärung von Aufgabe 8. Jedoch weist hier im sehr niedrigen Fähigkeitsbereich das DIF auf einen Vorteil zugunsten der Jungen hin. Hier verfügen diese Lernende meist über kein stabiles Verständnis des Eins-zu-Eins-Prinzips beim Zählen. Sie können die abgebildeten Finger nicht abzählen, ohne sich dabei zu verheddern oder Zahlen auszulassen (Gaidoschik et al., 2021; Gerster & Schultz, 2007). In diesem Fähigkeitsbereich profitieren die Jungen aufgrund ihrer stärker ausgeprägten non-verbalen Mengenerfassung und dem Subitizing (Carr & Jessup, 1997; Gerster & Schultz, 2007). Sie können Mengen oft als ganzheitlich visuelle Gestalt (z. B. geöffnete Hand als Fünfergruppe) erfassen, wodurch weniger kognitive Ressourcen verbraucht werden als beim sequenziellen Abzählen (Gallagher & Lisi, 1994; Krinzinger, 2010; Kuhl, 2015).

Auch Aufgabe 9 weist ein ähnliches nicht-uniformes DIF-Muster wie Aufgabe 1 auf. In dieser Aufgabe müssen die Lernenden aus der Anzahl der Zehner und der Einer die daraus resultierende Zahl benennen. Bei Lernenden im sehr niedrigen Fähigkeitsbereich ist das dezimale Stellenwertsystem noch nicht kognitiv gefestigt. Bei der ersten Teilaufgabe werden hierbei zuerst die Einer, dann die Zehner genannt. Hier weisen leistungsschwache Mädchen eine starke Fixierung auf die verbale und auditive Sequenzierung auf. Sie verarbeiten die Information also rein sprachlich-chronologisch (Kuhl, 2015). Die leistungsschwachen Jungen hingegen weisen eine geringere sprachliche Fixierung auf und nutzen vermehrt räumlich-schematische Repräsentationen (Krinzinger, 2010; Zuber et al., 2009). Für die zweite Teilaufgabe, die im höheren Fähigkeitsbereich angesiedelt ist, ist die Anzahl der Einer zweistellig, wodurch hier erst einmal entbündelt und neu gebündelt werden muss. Diese hochkomplexe Operation erfordert eine exzellente Kapazität des sprachlich-phonologischen Arbeitsgedächtnisses sowie eine präzise Exekutivkontrolle und die fehlerfreie Einhaltung mathematischer Transformationsregeln (Seitz & Schumann-Hengsteler, 2000). Hier weisen die Mädchen im niedrigeren bis hohen Fähigkeitsspektrum signifikant reifere sprachanalytische und metakognitive Kontrollmechanismen auf (Barel & Tzischinsky, 2022; Riley et al., 2016). Die Jungen in diesem Fähigkeitsbereich neigen hingegen unter dem herrschenden Zeitdruck vermehrt zu kognitiven Abkürzungen und verfallen dadurch häufiger in assoziative Fehler (Gallagher et al., 2000; Riley et al., 2016).

Da für den Test das Bifaktor-Modell die beste Passung darstellt, sollen nun noch die IICs sowie die TIC betrachtet werden, um zu prüfen, ob auch hier eine Eignung zur Diagnose von Rechenschwäche vorliegt. Anhand der TIC ist direkt zu erkennen, dass der Generalfaktor die meiste Information trägt. Die spezifischen Gruppenfaktoren weisen deutlich niedrigere Kurven auf. Hierbei liegen alle Peaks der Kurven im Bereich von $-2 \leq \theta \leq 0$,

also im negativen Fähigkeitsbereich. Somit kann auch hier die Eignung zur Diagnostik von Rechenschwäche festgestellt werden. Sowohl die IICs entlang des Generalfaktors als auch die IICs der spezifischen Faktoren weisen ihre Peaks im negativen Fähigkeitsspektrum auf und flachen im positiven Fähigkeitsspektrum stark ab. Anhand dieser Ergebnisse kann auch ein Deckeneffekt im oberen Leistungsbereich festgestellt werden. Dies stellt jedoch kein Problem dar, da der Test lediglich zur Diagnostik von Rechenschwäche entwickelt wurde und im oberen Leistungsbereich nicht differenzieren muss. Somit kann die Eignung für die Diagnose von Rechenschwäche aufgrund der dargestellten Ergebnisse als teilweise angesehen werden, aufgrund der Ergebnisse der DIF-Analyse. Jedoch kann in Kombination mit der Bayreuther-Förderdiagnostik eine gute Eignung zur Diagnose von Rechenschwäche angenommen werden.

5.2 Methodische Limitationen

Die Arbeit weist jedoch auch einige methodische Limitationen auf. So wurden etwa die explorative Faktorenanalyse und die konfirmatorische Faktorenanalyse auf demselben Datensatz durchgeführt, wodurch ein Teil der guten Passung darauf zurückgeführt werden kann. Die Arbeit auf demselben Datensatz war hier unausweichlich, da der Datensatz zu klein ist, um ihn zu splitten und saubere Ergebnisse zu erhalten und ein weiterer Datensatz lag hierzu nicht vor. Daher besaß die mathematische Konvergenzsicherung zwingend Vorrang vor einer reinen Kreuzvalidierung.

Weiterhin wurde die DIF-Analyse auf Basis des klassischen EFA-Modells durchgeführt, da eine Multigruppenschätzung unter dem EFA-Bifaktor-Modell zu einer computationellen Überlastung geführt hätte. Die geschätzte Rechenzeit von über 36 Stunden pro Modelllauf steht in keinem angemessenen Verhältnis zur pragmatischen Machbarkeit der Arbeit, insbesondere vor dem Hintergrund akuter numerischer Instabilitäten und drohender Nicht-Konvergenz bei solch komplexen Parameterschätzungen.

Die Validität der geschlechterspezifischen DIF-Analysen wird durch eine weitere Stichprobenreduktion auf $N = 467$ auswertbare Fälle limitiert. Dieser Fallzahlverlust gegenüber der Gesamtaufbereitung birgt das Risiko von Selektionseffekten. Durch die Aufteilung in die beiden Vergleichsgruppen schrumpft die effektive Datenbasis pro Geschlecht auf ein kritisches Minimum von knapp 230 Personen pro Gruppe. Multigruppenschätzungen polytomer Items stoßen in diesen Größenordnungen an mathematische Grenzen. Einzelne Antwortstufen sind in den Extrembereichen des Fähigkeitsspektrums nur noch schwach besetzt. Das provoziert inflationierte Standardfehler bei den Itemparametern. Die Stabilität der identifizierten DIF-Effekte ist folglich mit methodischer Vorsicht zu interpretieren.

Über diese numerischen Grenzen hinaus schränkt die regionale Herkunft der Stichprobe die Generalisierbarkeit der Ergebnisse ein. Die Daten stammen aus dem bayrischen Raum. Da die Bildung in Deutschland föderalisiert ist, keine einheitlichen Lehrpläne existieren und die sozioökonomischen Faktoren sowie Migrationsquote regional stark variieren, lassen sich die Befunde nicht uneingeschränkt übertragen. Das Diagnoseinstrument

BRT-2 funktioniert in anderen Bundesländern aufgrund anderer didaktischer Schwerpunkte im Unterricht anders. Die psychometrischen Befunde sind regional gebunden.

Das Querschnittsdesign setzt eine weitere methodische Grenze, da es sich hierbei um eine Momentaufnahme handelt und die Entwicklungsprozesse im zeitlichen Verlauf unsichtbar bleiben. So bleibt auch völlig unklar, ob die gefundenen DIF-Muster stabil bleiben oder ob sie sich im Zuge der weiteren mathematischen Beschulung verflüchtigen.

5.3 Implikationen für die Weiterentwicklung des Instruments

Die empirischen Befunde dieser Arbeit liefern konkrete Ansatzpunkte für eine grundlegende Überarbeitung des BRT-2. Diese sollen in den folgenden Teilkapiteln ausgeführt werden.

5.3.1 Itemmetrische Optimierung und Schwellenbereinigung

Die Aufgaben A1, A4 und A5 weisen eine marginale Informationsleistung auf. Sie blähen die Testzeit auf, haben jedoch keinen diagnostischen Mehrwert für die Erfassung von Rechenschwäche. Diese Items sollten in einer überarbeiteten Nachfolgeversion restlos gestrichen werden (Moosbrugger & Kelava, 2020).

Ein mathematisches Problem zeigt sich bei den Aufgaben A5, A6 und A11. Hier liegt eine Inversion der Schwellenparameter vor. Die mittleren Punktstufen differenzieren empirisch nicht, wodurch das Antwortverhalten komprimiert wird. Für die Testpraxis bedeutet dies, entweder eine strikt dichotome Codierung im Manual zu verankern, oder die Antwortformate didaktisch neu konzipieren. Die Zwischenstufen müssen für Lernende im niedrigen Fähigkeitsbereich psychometrisch unterscheidbar werden.

5.3.2 Konstruktionstechnische Kompensation von Geschlechtereffekten (DIF)

Vier Items verzerren die Testfairness substanziell. Das ist ein massives Problem. Wenn Mädchen bei Aufgabe 8 durch ihre sequenzielle Verarbeitungsgenauigkeit bevorteilt sind, während Jungen bei Aufgabe 17 von intuitiven, visuo-räumlichen Heuristiken profitieren, misst der Test kognitive Stile und nicht mathematische Kernkompetenz.

Ein zukünftiges Instrument darf diese Asymmetrien nicht ignorieren. Die Lösung hierfür liegt in einer systematischen Parallelisierung der Aufgabenmerkmale. Für jedes Item, das sprachanalytische Kontrollmechanismen triggert, muss ein mathematisch äquivalentes Gegenstück implementiert werden, welches visuo-räumliche Repräsentationen anspricht. Das Ziel ist ein balancierter Gesamttest bei dem der kumulierte DIF auf Testebene gegen Null konvergiert (Moosbrugger & Kelava, 2020).

5.3.3 Transformation in ein Computergestütztes Adaptives Testverfahren (CAT)

Die empirischen Befunde zur dimensionalen Struktur weisen der Weiterentwicklung des BRT-2 einen klaren Weg. Die Überführung des Papier-und-Bleistift-Verfahrens in ein computergestütztes adaptives Testformat (Computerized Adaptive Testing, CAT). Dieser Schritt ist keine technische Spielerei, sondern eine psychometrische Notwendigkeit zur Sicherung der Messökonomie (Moosbrugger & Kelava, 2020).

Die Gesamtinformationskurve (TIC) des Generalfaktors zeigt eine deutliche Konzentration im negativen Fähigkeitsbereich ($\theta < 0$). Im analogen Papierformat bedeutet dies einen erheblichen Effizienzverlust. Leistungsstarke Lernende müssen den gesamten Test durchlaufen, obwohl die Items im oberen Spektrum keine diagnostisch verwertbare Differenzierungskraft besitzen. Ein digitaler CAT-Algorithmus steuert die Itempräsentation dynamisch auf Basis der berechneten Item- und Schwellenparameter (Moosbrugger & Kelava, 2020). Sobald das System eine überdurchschnittliche mathematische Fähigkeit schätzt, bricht das Verfahren ab. Das verkürzt die Testzeit drastisch (Chang, 2015).

Die wahre Stärke der Digitalisierung liegt jedoch in der Entschärfung der empirischen Bruchstellen, die das Fehlernetzwerk so drastisch offenbart. Die Netzwerkanalyse zeigt eine dichte, hochgradig asymmetrische Struktur. Die zwei gigantischen „Fehlersenken“ A2 und A17 beherrschen das Bild. Fast alle Fehlerpfade des Tests münden in diese beiden Items. Von der zentralen Schlüsselaufgabe A3 führen vier Pfeile mit maximaler bedingter Wahrscheinlichkeit ($P(B\ falsch|A\ falsch) > 80\ \%$) direkt zu den Aufgaben A2, A6, A8 und A9. Wer an Aufgabe 3 scheitert, bricht psychometrisch fast zwangsläufig bei diesen Aufgaben ein. Im analogen Setting erzeugt dieses starre Abarbeiten unlösbarer Folgeaufgaben bei rechenschwachen Kindern massive Demotivationseffekte.

Ein CAT-System nutzt diese topologische Gesetzmäßigkeiten für ein intelligentes „Smart Routing“ (van der Linden & Glas, 2010). Registriert die Software ein Scheitern an Aufgabe 3, blockiert der Algorithmus die Frustrations-Items A2 und A17 sofort. Das System bricht die Kaskade ab und leitet den Lernenden stattdessen zu den Aufgaben (A1, A4 und A5) um, die im Fehlernetzwerk als vollkommen isolierte Satelliten am unteren Rand schweben. Diese drei Items sind mathematisch und empirisch komplett vom restlichen Fehlersystem entkoppelt. Sie ermöglichen es, basale Kompetenzen völlig unabhängig vom zentralen Problemcluster stressfrei und präzise abzutesten (Moosbrugger & Kelava, 2020).

Zuletzt leistet die Digitalisierung einen Beitrag zur Aufklärung der identifizierten Fairness-Probleme (DIF). Die theoretischen Erklärungsansätze für den geschlechtsspezifischen DIF bei den Aufgaben A1, A8, A9 und A17 basieren maßgeblich auf unterschiedlichen kognitiven Verarbeitungsstilen unter Zeitdruck. Ein digitales Erhebungsinstrument ermöglicht die automatische Miterfassung von Log-Daten, insbesondere der exakten Antwortlatenzen pro Item (Kupiainen et al., 2014; Moosbrugger & Kelava, 2020). Die Integration dieser Prozessdaten erlaubt es, die postulierten Geschlechterunterschiede im Ar-

beitsverhalten empirisch direkt zu überprüfen. Die Validität der DIF-Interpretation steht damit auf einem stabilen Fundament (Kupiainen et al., 2014).

5.3.4 Schnittstelle zur qualitativen Vertiefung: Kopplung mit der Bayreuther Förderdiagnostik

Diagnostik darf kein Selbstzweck sein. Der standardisierte BRT-2 liefert das quantitative Fundament, stößt aber bei der Aufdeckung individueller Denkwege an systemische Grenzen. Hier wird die Brücke zur *Bayreuther Förderdiagnostik* (BFD) als qualitativem Interview-Verfahren geschlagen. Diese Synthese aus produkt- und prozessorientierter Diagnostik bildet das Herzstück des zweistufigen Bayreuther Konzepts (Steinecke & Richter, 2025).

Ein massives Problem im Schulalltag bleibt der Zeitmangel. Lehrkräfte können unmöglich bei jedem auffälligen Kind eine zeitintensive Volldiagnostik durchführen. Das quantitative Screening des BRT-2 schrumpft diesen diagnostischen Suchraum radikal ein, da die Auswahl der zu interviewenden Lernenden so an die personelle Kapazität der jeweiligen Schule angepasst werden kann (Steinecke & Richter, 2025). Die in dieser Arbeit identifizierten sechs Leistungstypen der Clusteranalyse bieten dafür das perfekte Gelenkstück. Zeigt das Testergebnis des BRT-2 etwa spezifische Einbrüche im Bereich des Stellenwertverständnisses, steuert das die Lehrkraft direkt zu den passenden qualitativ-prozessorientierten Interview-Leitfäden der BFD (Steinecke & Richter, 2025).

Im Einzelgespräch lassen sich die konkreten Fehlvorstellungen, wie beispielsweise beim Bündeln oder Entbündeln, sodann exakt freigelegen. Erst aus dieser mathematikdidaktisch fundierten Kombination von ökonomischer Strukturierung und qualitativer Mikroanalyse erwächst eine verlässliche Basis für die spätere Förderung (Steinecke & Richter, 2025).

6 Fazit

Der Rechentest BRT-2 hält der psychometrischen Belastungsprobe stand. Er tut es auf eine eigene, empirisch faszinierende Weise. Die theoretisch postulierte Dreiteilung der mathematischen Kompetenzen lässt sich in den empirischen Daten so jedoch nicht halten.

Zahlbegriff und Stellenwertverständnis bilden keine getrennten Schubladen, sondern verschmelzen zu einer untrennbaren, dezimalen Strukturkompetenz. Ergänzt um die Rechenoperationen und ein relationales Zahlverständnis zeigt die empirische Faktorenanalyse ein differenzierteres Bild kindlicher Kognition, als es die reine Theorie vermuten ließ. Dass die konfirmatorische Faktorenanalyse das Bifaktor-Modell als klaren Sieger kürt, liefert die gewünschte essenzielle Eindimensionalität. Lehrkräfte addieren am Ende eben doch keine Äpfel mit Birnen, sondern berechnen einen aussagekräftigen Gesamtwert.

Als Screening-Instrument zur Früherkennung von Rechenschwäche leistet der BRT-2 genau das, was er soll. Die Peaks der Informationskurven im Bereich von $-3 \leq \theta \leq 0$ auf der latenten Fähigkeitsskala zeigen die Messpräzision exakt am leistungsschwachen Rand, weshalb hier der Standardfehler am kleinsten ist. Dass der Test im oberen Leistungsbereich massive Deckeneffekte zeigt und dort kaum noch differenziert, ist kein Mangel. Es ist ein Ausweis ökonomischer Effizienz, da die Begabungsdiagnostik nie das Ziel des Tests war. Die Schwachstellen sind jedoch gravierend, so weisen die Items A11, A6 und A5 eine Schwelleninversion auf, wodurch die polytome Punktabstufung kollabiert und die Aufgabe faktisch dichotom codiert ist.

Noch schwerer wiegt der systematische Bias. Vier Items verzerren die Testfairness durch ein signifikantes Differential Item Functioning. Wenn Jungen bei der Subtraktion (A17) von intuitiven Abkürzungen und visuo-räumlichen Repräsentationen profitieren, während Mädchen beim strukturellen Zählen (A8) ihre sequenzielle Genauigkeit ausspielen, misst der Test unbewusst kognitive Verarbeitungsstile. Bei den Aufgaben A1 und A9 zeigt sich dieses asymmetrische Bild erneut, wobei hier jeweils im niedrigen Leistungsbereich ein Wechsel der bevorteilten Gruppe stattfindet. Hier ringen sprachlich-chronologische Fixierungen mit ganzheitlich-visueller Mengenerfassung – ein unfairer Nebeneffekt, der die diagnostische Neutralität gefährdet.

Diese empirischen Brüche zwingen zu einer grundlegenden didaktischen Reflexion. Sie offenbaren das ewige Spannungsfeld zwischen testtheoretischer Präzision und unterrichtlicher Realität. Aus rein mathematisch-statistischer Sicht müssten verzerrende oder informationsarme Aufgaben sofort eliminiert werden. Fachdidaktisch jedoch repräsentieren gerade diese Items unverzichtbare Schlüsselschritte des mathematischen Anfangsunterrichts. Ein Test, der diese didaktischen Kernstellen aus statistischen Gründen opfert, verliert seine inhaltliche Validität. Der BRT-2 ist in seiner jetzigen Form ein hervorragendes Beispiel für die Notwendigkeit einer Symbiose. Isoliert betrachtet erfüllt er seine

primäre Screening-Funktion zur Identifikation von Risikoprofilen dank der hohen Messgenauigkeit im negativen Fähigkeitsbereich zuverlässig. Die aufgedeckten Mängel schränken jedoch eine isolierte, rein produktorientierte Interpretation auf Item-Ebene ein. Erst die direkte Verknüpfung mit der Bayreuther Förderdiagnostik (BFD) heilt diese diagnostischen Unschärfen (Steinecke & Richter, 2025). Das quantitative Raster des Screenings liefert die verlässliche Alarmglocke und ordnet die Lernende präzise sechs empirischen Leistungstypen zu. Die Lehrkraft greift im Anschluss gezielt zum passenden, qualitativen Interviewleitfaden der BFD, um konkrete Fehlvorstellungen im Einzelgespräch freizulegen, wodurch Zeit gespart wird (Steinecke & Richter, 2025). Somit erhält man eine gelenkte Tiefenbohrung anstelle ungerichteter Suche.

Der Ausblick zeigt in zwei Richtungen. Die erste betrifft die testkonstruktive Weiterentwicklung des Instruments. Die Transformation des BRT-2 in ein computergestütztes adaptives Testverfahren (CAT) ist der einzig logische Entwicklungsschritt zur Steigerung der Messökonomie (Moosbrugger & Kelava, 2020). Ein smarter Algorithmus, gefüttert mit den empirischen Fehler-Abhängigkeiten des Netzwerks, bricht die Testung bei drohendem Frust gezielt ab. Sobald eine Überlastung an den Fehlersenken A2 und A17 droht, beispielsweise nach dem Scheitern an der zentralen Schlüsselaufgabe A3, sperrt das System diese unlösbaren Hürden. Es steuert stattdessen die im Fehlernetzwerk isolierten Items A1, A4 und A5 an. So lässt sich das basale Fundament eingrenzen, ohne psychische Überlastung der betroffenen Lernenden. Gleichzeitig liefert die Digitalisierung über Reaktionszeiten endlich die Daten, um den geschlechtsspezifischen DIF endgültig zu entschlüsseln (Kupiainen et al., 2014).

Die zweite Richtung betrifft die zukünftige empirische Forschung. Hier müssen die methodischen Limitationen dieser Arbeit als Ausgangspunkt dienen. Da EFA und CFA auf derselben Stichprobe gerechnet wurden, ist eine unabhängige Kreuzvalidierung der bifaktoriellen Struktur an einer frischen, deutlich größeren Stichprobe dringend erforderlich. Nur so lässt sich prüfen, ob die gefundene dimensionale Architektur stabil bleibt oder ein Artefakt der vorliegenden Daten darstellt. Ebenso verlangt der ausgeprägte Föderalismus im deutschen Bildungssystem eine überregionale Validierung. Es muss untersucht werden, wie sich curriculare Unterschiede anderer Bundesländer auf die Aufgabenschwierigkeit und die Struktur des Tests auswirken. Schließlich bleibt der Blick im Querschnitt eine Momentaufnahme. Zukünftige Längsschnittstudien sollten die Entwicklungsprozesse der Lernenden über das gesamte zweite Schuljahr hinweg begleiten. Sie könnten klären, ob die geschlechtsspezifischen DIF-Effekte temporäre Phänomene unterschiedlicher Reifungsgeschwindigkeiten sind oder sich als stabile Barrieren im mathematischen Spracherwerb verfestigen.

Literaturverzeichnis

Barel, E. & Tzischinsky, O. (2022). Sex differences in the sustained attention of elementary school children. *BMC psychology*, *10*(1), 307. https://doi.org/10.1186/s40359-022-01007-z

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Statistical Psychology*, *3*(2), 77–85. https://doi.org/10.1111/j.2044-8317.1950.tb00285.x

Benjamini, Y. & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bock, R. D. & Mislevy, R. J. (1982). Adaptive EAP Estimation of Ability in a Microcomputer Environment. *Applied Psychological Measurement*, *6*(4), 431–444. https://doi.org/10.1177/014662168200600405

Bortz, J. & Lienert, G. A. (2008). *Verteilungsfreie Methoden in der Biostatistik: Mit 247 Tabellen* (3., korrigierte Aufl.). *Springer-Lehrbuch Bachelor, Master*. Springer Medizin Verl.

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler: Mit 70 Abbildungen und 163 Tabellen* (7., vollständig überarbeitete und erweiterte Auflage, limitierte Sonderausgabe). Springer.

Burnham, K. P. & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). Springer.

Cai, L. (2010). High-dimensional Exploratory Item Factor Analysis by A Metropolis–Hastings Robbins–Monro Algorithm. *Psychometrika*, *75*(1), 33–57. https://doi.org/10.1007/s11336-009-9136-x

Carr, M. & Jessup, D. L. (1997). Gender differences in first-grade mathematics strategy use: Social and metacognitive influences. *Journal of Educational Psychology*, *89*(2), 318–328. https://doi.org/10.1037/0022-0663.89.2.318

Chalmers, R. P. (2012). mirt : A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6). https://doi.org/10.18637/jss.v048.i06

Chang, H.-H. (2015). Psychometrics behind Computerized Adaptive Testing. *Psychometrika*, *80*(1), 1–20. https://doi.org/10.1007/s11336-014-9401-5

Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. (2014). NbClust : An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, *61*(6). https://doi.org/10.18637/jss.v061.i06

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. ed.). Erlbaum. https://doi.org/10.4324/9780203771587

Cronbach, L. J. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Csárdi, G., Nepusz, T., Traag, V., Horvát, S., Zanini, F., Noom, D., Müller, K., Schoch, D. & Salmon, M. (2006). *igraph: Network Analysis and Visualization*. https://doi.org/10.32614/CRAN.package.igraph

Döring, N., Bortz, J., Pöschl-Günther, S., Werner, C. S., Schermelleh-Engel, K., Gerhard, C. & Gäde, J. C. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Aufl. 2016). *Springer-Lehrbuch*. Springer Berlin Heidelberg. http://nbn-resolving.org/urn:nbn:de:bsz:31-epflicht-1624548 https://doi.org/10.1007/978-3-642-41089-5

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K. & Japel, C. (2007). School readiness and later achievement. *Developmental psychology*, *43*(6), 1428–1446. https://doi.org/10.1037/0012-1649.43.6.1428

Dunn, T. J., Baguley, T. & Brunsden, V. (2014). From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. *British journal of psychology (London, England : 1953)*, *105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Epskamp, S., Borsboom, D. & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior research methods*, *50*(1), 195–212. https://doi.org/10.3758/s13428-017-0862-1

Field, A., Miles, J. & Field, Z. (2012). *Discovering statistics using R*. Sage.

Fligner, M. A. & Killeen, T. J. (1976). Distribution-Free Two-Sample Tests for Scale. *Journal of the American Statistical Association*, *71*(353), 210–213. https://doi.org/10.1080/01621459.1976.10481517

Friedman, M. (1937). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, *32*(200), 675–701. https://doi.org/10.1080/01621459.1937.10503522

Fruchterman, T. M. J. & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, *21*(11), 1129–1164. https://doi.org/10.1002/spe.4380211102

Gaidoschik, M., Moser Opitz, E., Nührenbörger, M., Rathgeb-Schnierer, E. & Götze, D. (2021). *Besondere Schwierigkeiten beim Mathematiklernen*. https://doi.org/10.13140/RG.2.2.15952.64004

Gallagher, A. M., De Lisi, R., Holst, P. C., McGillicuddy-De Lisi, A. V., Morley, M. & Cahalan, C. (2000). Gender differences in advanced mathematical problem solving. *Journal of experimental child psychology*, *75*(3), 165–190. https://doi.org/10.1006/jecp.1999.2532

Gallagher, A. M. & De Lisi, R. (1994). Gender differences in Scholastic Aptitude Test: Mathematics problem solving among high-ability students. *Journal of Educational Psychology*, *86*(2), 204–211. https://doi.org/10.1037/0022-0663.86.2.204

Geary, D. C. (1996). Sexual selection and sex differences in mathematical abilities. *Behavioral and Brain Sciences*, *19*(2), 229–247. https://doi.org/10.1017/S0140525X00042400

George, D. & Mallery, P. (2016). *IBM SPSS statistics 23 step by step: A simple guide and reference* (Fourteenth edition). Routledge Taylor & Francis Group.

Gerster, H.-D. & Schultz, R. (2007). *Schwierigkeiten beim Erwerb mathematischer Konzepte im Anfangsunterricht : Bericht zum Forschungsprojekt „Rechenschwäche - Erkennen, Beheben, Vorbeugen“*. Pädagogische Hochschule Freiburg.

Guttman, L. (1954). Some Necessary Conditions for Common-Factor Analysis. *Psychometrika*, *19*(2), 149–161. https://doi.org/10.1007/BF02289162

Harman, H. H. & Jones, W. H. (1966). Factor analysis by minimizing residuals (minres). *Psychometrika*, *31*(3), 351–368. https://doi.org/10.1007/BF02289468

Hendrickson, A. E. & White, P. O. (1964). Promax: A Quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology*, *17*(1), 65–70. https://doi.org/10.1111/j.2044-8317.1964.tb00244.x

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185. https://doi.org/10.1007/BF02289447

Hothorn, T., Winell, H., Hornik, K., van de Wiel, M. A. & Zeileis, A. (2005). *coin: Conditional Inference Procedures in a Permutation Test Framework*. https://doi.org/10.32614/CRAN.package.coin

Howard, M. C. (2016). A Review of Exploratory Factor Analysis Decisions and Overview of Current Practices: What We Are Doing and How Can We Improve? *International Journal of Human-Computer Interaction*, *32*(1), 51–62. https://doi.org/10.1080/10447318.2015.1087664

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Ingenkamp, K. & Lissmann, U. (2008). *Lehrbuch der pädagogischen Diagnostik* (6. Auflage). Beltz Verlag.

Kaiser, H. F. (1974). An Index of Factorial Simplicity. *Psychometrika*, *39*(1), 31–36. https://doi.org/10.1007/BF02291575

Kaiser, H. F. & Rice, J. (1974). Little Jiffy, Mark IV. *Educational and Psychological Measurement*, *34*(1), 111–117. https://doi.org/10.1177/001316447403400115

Kassambara, A. & Mundt, F. (2026). *factoextra:Extract and Visualize the Results of Multivariate Data Analyses*. https://doi.org/10.32614/CRAN.package.factoextra

Kaufman, L. & Rousseeuw, P. J. (2005). *Finding groups in data: An introduction to cluster analysis*. *Wiley-Interscience paperback series: Bd. 47*. Wiley. https://doi.org/10.2307/2532178

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1-2), 81–93. https://doi.org/10.1093/biomet/30.1-2.81

Krinzinger, H. (2010). *The role of multi-digit number processing in the development of numerical cognition* [Doktorarbeit, RWTH Aachen].

Kuhl, A. (2015). *Zwanzigeins vs. einundzwanzig: Profitieren rechenschwache Kinder in Bezug auf die Dauer von einer stellenwertgerechten Sprechweise? Eine empirische Untersuchung* [Masterarbeit, Pädagogische Hochschule Schwäbisch Gmünd]. Pädagogische Hochschule Schwäbisch Gmünd. https://77lw780y81stz7zy4g9m6lmx.justedit.com/downloads/Kuhl%20A_2025_Masterarbeit.pdf

Kupiainen, S., Vainikainen, M.-P., Marjanen, J. & Hautamäki, J. (2014). The role of time on task in computer-based low-stakes assessment of cross-curricular skills. *Journal of Educational Psychology*, *106*(3), 627–638. https://doi.org/10.1037/a0035507

MacQueen, J. B., Le Cam, L. M. & Neyman, J. (1967). *Some methods for classification and analysis of multivariate observations*. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. University of California Press.

Mair, P. (2018). *Modern Psychometrics with R* (1st ed. 2018). *Use R!* Springer International Publishing.

Mann, H. B. & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, *18*(1), 50–60. https://doi.org/10.1214/aoms/1177730491

Matsunaga, M. (2010). How to factor-analyze your data right: do's, don'ts, and how-to's. *International Journal of Psychological Research*, *3*(1), 97–110. https://doi.org/10.21500/20112084.854

Maydeu-Olivares, A. & Joe, H. (2006). Limited Information Goodness-of-fit Testing in Multidimensional Contingency Tables. *Psychometrika*, *71*(4), 713–732. https://doi.org/10.1007/s11336-005-1295-9

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological methods*, *23*(3), 412–433. https://doi.org/10.1037/met0000144

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*(2), 127–143. https://doi.org/10.1016/0883-0355(89)90002-5

Moosbrugger, H. & Kelava, A. (Hrsg.). (2020). *Testtheorie und Fragebogenkonstruktion* (3., vollständig neu bearbeitete, erweiterte und aktualisierte Auflage). Springer. https://doi.org/10.1007/978-3-642-20072-4

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. https://doi.org/10.1177/014662169201600206

Newman, M. E. J. (2010). *Networks: An introduction*. Oxford University Press.

Orlando, M. & Thissen, D. (2000). Likelihood-Based Item-Fit Indices for Dichotomous Item Response Theory Models. *Applied Psychological Measurement*, *24*(1), 50–64. https://doi.org/10.1177/01466216000241003

Padberg, F. (2015). *Einführung Mathematik Primarstufe - Arithmetik* (2. Aufl. 2015). Springer Spektrum. https://doi.org/10.1007/978-3-662-43449-9

Pedersen, T. L. (2019). *patchwork: The Composer of Plots*. https://doi.org/10.32614/CRAN.package.patchwork

Pedersen, T. L. (2026a). *ggraph: An Implementation of Grammar of Graphics for Graphs and Networks*. https://doi.org/10.32614/CRAN.package.ggraph

Pedersen, T. L. (2026b). *tidygraph: A Tidy API for Graph Manipulation*. https://doi.org/10.32614/CRAN.package.tidygraph

Pedersen, T. L., Ooms, J. & Govett, D. (2026). *systemfonts: System Native Font Finding*. https://doi.org/10.32614/CRAN.package.systemfonts

R Core Team. (2026). *R: A Language and Environment for Statistical Computing*. https://www.r-project.org/

Reckase, M. D. (2009). *Multidimensional Item Response Theory* (1st ed.). Springer New York. https://doi.org/10.1007/978-0-387-89976-3

Reise, S. P. (2012). Invited Paper: The Rediscovery of Bifactor Measurement Models. *Multivariate behavioral research*, *47*(5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Revelle, W. (2026). *psych: Procedures for Psychological, Psychometric, and Personality Research*. https://doi.org/10.32614/CRAN.package.psych

Riley, E., Okabe, H., Germine, L., Wilmer, J., Esterman, M. & DeGutis, J. (2016). Gender Differences in Sustained Attentional Control Relate to Gender Inequality across Countries. *PloS one*, *11*(11), e0165100. https://doi.org/10.1371/journal.pone.0165100

Rodriguez, A., Reise, S. P. & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological methods*, *21*(2), 137–150. https://doi.org/10.1037/met0000045

Rosenthal, R. (1991). *Meta-Analytic Procedures for Social Research*. SAGE Publications, Inc. https://doi.org/10.4135/9781412984997

Rosseel, Y. (2012). lavaan : An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2). https://doi.org/10.18637/jss.v048.i02

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Schmid, J. & Leiman, J. M. (1957). The Development of Hierarchical Factor Solutions. *Psychometrika*, *22*(1), 53–61. https://doi.org/10.1007/BF02289209

Seitz, K. & Schumann-Hengsteler, R. (2000). Mental multiplication and working memory. *European Journal of Cognitive Psychology*, *12*(4), 552–570. https://doi.org/10.1080/095414400750050231

Shapiro, S. S. & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, *52*(3/4), 591. https://doi.org/10.2307/2333709

Siegler, R. S., Thompson, C. A. & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive psychology*, *62*(4), 273–296. https://doi.org/10.1016/j.cogpsych.2011.03.001

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, *15*(1), 72. https://doi.org/10.2307/1412159

Steinecke, A. & Richter, K. (2025). *Bayreuther Testpaket zur Erfassung von Rechenschwäche im Mathematikunterricht der Jahrgangsstufe 2*. University of Bayreuth. https://doi.org/10.15495/EPub_UBT_00008536

Steinecke, A. & Ulm, V. (2025). *Rechenschwäche in der Sekundarstufe: Spezifische Schwierigkeiten verstehen, erkennen und überwinden : Sekundarstufe I* (1. Auflage). *Scriptor Praxis*. Cornelsen.

Tibshirani, R., Walther, G. & Hastie, T. (2001). Estimating the Number of Clusters in a Data Set Via the Gap Statistic. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *63*(2), 411–423. https://doi.org/10.1111/1467-9868.00293

Urhahne, D., Dresel, M. & Fischer, F. (Hrsg.). (2019). *Psychologie für den Lehrberuf*. Springer. https://doi.org/10.1007/978-3-662-55754-9_28

van der Linden, W. J. & Glas, C. A. W. (2010). *Elements of adaptive testing*. *Statistics for social and behavioral sciences*. Springer.

Weltgesundheitsorganisation. (2019). *International Classification of Diseases, Eleventh Revision (ICD-11)*. https://icd.who.int/browse11

Wickham, H. (2007). Reshaping Data with the reshape Package. *Journal of Statistical Software*, *21*(12). https://doi.org/10.18637/jss.v021.i12

Wickham, H. (2010). *reshape2: Flexibly Reshape Data: A Reboot of the Reshape Package*. https://doi.org/10.32614/CRAN.package.reshape2

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., . . . Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D. & van den Brand, T. (2026). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://doi.org/10.32614/CRAN.package.ggplot2

Wickham, H., Henry, L., Pedersen, T. L., Luciani, T. J., Decorde, M. & Lise, V. (2026). *svglite: An 'SVG' Graphics Device*. https://doi.org/10.32614/CRAN.package.svglite

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, *1*(6), 80. https://doi.org/10.2307/3001968

Yen, W. M. (1984). Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. *Applied Psychological Measurement*, *8*(2), 125–145. https://doi.org/10.1177/014662168400800201

Zinbarg, R. E., Revelle, W., Yovel, I. & Li, W. (2005). Cronbach's α Revelle's β and Mcdonald's ω H : their Relations with Each Other and Two Alternative Conceptualizations of Reliability. *Psychometrika*, *70*(1), 123–133. https://doi.org/10.1007/s11336-003-0974-7

Zuber, J., Pixner, S., Moeller, K. & Nuerk, H.-C. (2009). On the language specificity of basic number processing: transcoding in a language with inversion and its relation to working memory capacity. *Journal of experimental child psychology*, *102*(1), 60–77. https://doi.org/10.1016/j.jecp.2008.04.003

Anhang

Anhangsverzeichnis

Anhang A: Tabelle der Ergebnisse des Shapiro-Wilk-Tests für die Geschlechter

Anhang B: Spearman-Korrelationsmatrix

Anhang C: 6-Faktoren-Modell

Anhang D: Vollständige Ergebnismatrix der globalen Passung

Anhang E: Vollständige Q3-Matrizen

Anhang F: vollständige Parameterschätzung und Fit-Tabellen

Anhang G: vollständige Übersicht der CRC-Kurven

Anhang H: Vollständige Übersicht der CRC-Kurven der DIF-Analyse und DIF-Ergebnisse

Anhang I: vollständige CFA-Tabelle

Anhang J: Erklärung zur KI-Nutzung

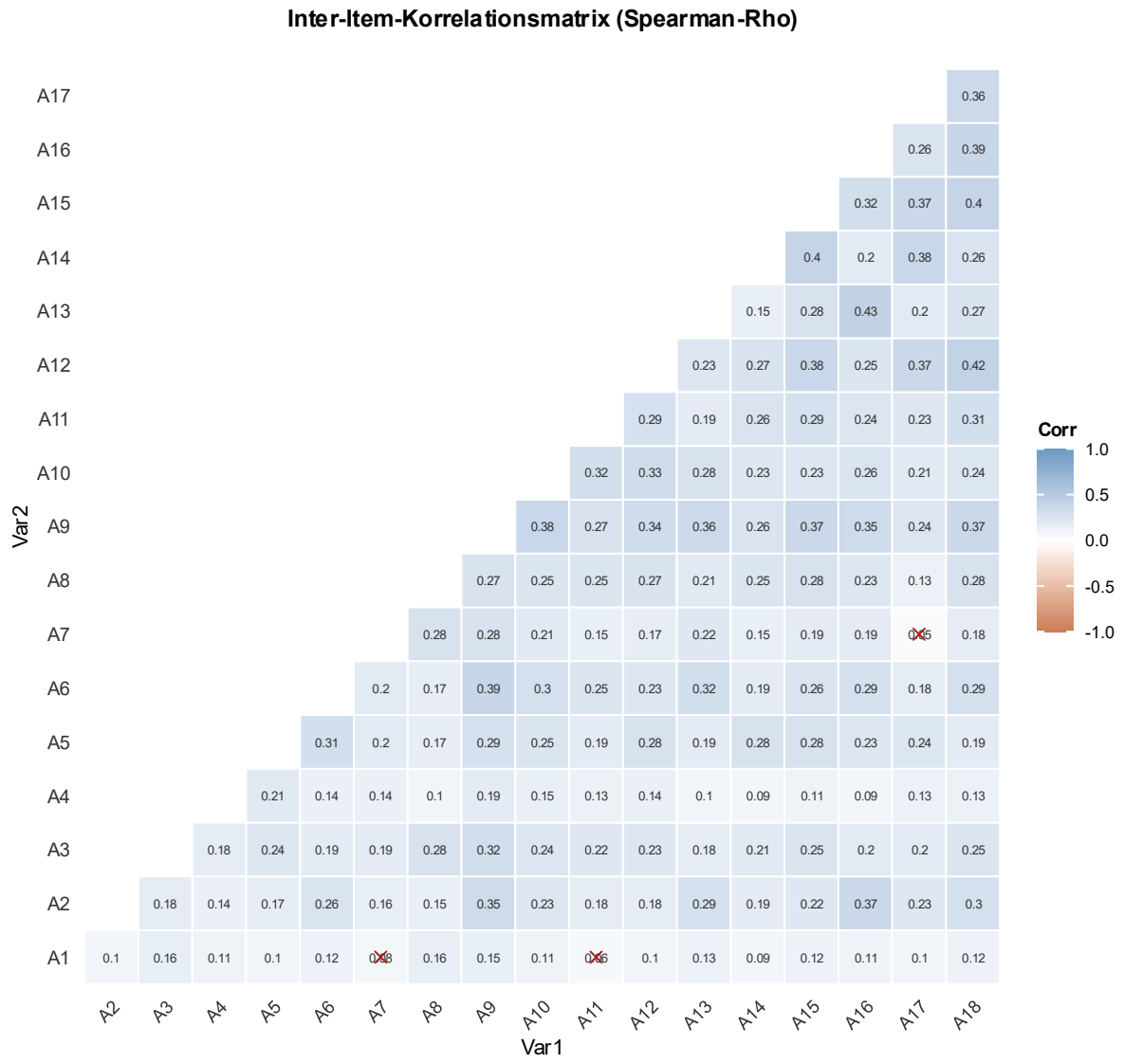
Anhang A

Tabelle A: Ergebnisse des Shapiro-Wilk-Tests für die Geschlechtergruppen

Kompetenzbereich	Geschlecht	N	Teststatistik W	p-Wert	Interpretation
Gesamt	Männlich	224	0,8646	$3{,}383 \cdot 10^{-13}$	Abweichung von Normalverteilung
	Weiblich	243	0,8949	$5{,}575 \cdot 10^{-12}$	Abweichung von Normalverteilung
Verständnis für natürliche Zahlen	Männlich	224	0,8371	$1{,}354 \cdot 10^{-14}$	Abweichung von Normalverteilung
	Weiblich	243	0,8554	$2{,}497 \cdot 10^{-14}$	Abweichung von Normalverteilung
Verständnis für das Stellenwertsystem	Männlich	224	0,8235	$3{,}150 \cdot 10^{-15}$	Abweichung von Normalverteilung
	Weiblich	243	0,8642	$7{,}594 \cdot 10^{-14}$	Abweichung von Normalverteilung
Verständnis für Rechenoperationen	Männlich	224	0,8975	$3{,}059 \cdot 10^{-11}$	Abweichung von Normalverteilung
	Weiblich	243	0,9288	$1{,}991 \cdot 10^{-9}$	Abweichung von Normalverteilung

Anhang B

Abbildung A: Spearman-Korrelationsmatrix der 18 Items. Nicht signifikante Korrelationen wurden ausgekreuzt.



Anhang C

Tabelle B: Faktorladungen des 6-Faktor-Modells

Item	Faktor 1	Faktor 2	Faktor 3	Faktor 4	Faktor 5	Faktor 6	h^2
A1					0,535		0,288
A2			0,667				0,462
A3					0,471		0,528
A4					0,346		0,258
A5				0,697			0,587
A6			0,397				0,473
A7						0,489	0,474
A8						0,773	0,677
A9	0,304		0,387				0,620
A10	0,762						0,610
A11	0,664						0,534
A12	0,416	0,409					0,502
A13			0,689				0,515
A14		0,664					0,557
A15		0,627					0,642
A16			0,805				0,603
A17		0,808				-0,326	0,646
A18		0,367	0,321				0,528
SS loadings	1,424	1,965	2,087	0,871	0,781	1,125	
Proportion Var	0,079	0,109	0,116	0,048	0,043	0,063	
Cumulative Var	0,079	0,188	0,304	0,352	0,395	0,458	
Faktorkorrelationen							
Faktor 1	-						
Faktor 2	0,674	-					
Faktor 3	0,682	0,552	-				
Faktor 4	0,610	0,329	0,514	-			
Faktor 5	0,605	0,417	0,519	0,491	-		
Faktor 6	0,603	0,464	0,564	0,487	0,447	-	

Anmerkung. N = 574. Faktorenextraktion mittels Minimum-Residual-Methode mit Promax-Rotation. h^2 = Kommunalität, SS loadings = Quadratsumme der Faktorladungen

Anhang D

Tabelle C: Globale Fit-Indizes der drei unifaktoriellen Faktoren des EFA-Modells

Faktor	M2	df	p	RMSEA	SRMSR	TLI	CFI
Faktor 1	17,198	20	0,64	0	0,035	1,005	1
Faktor 2	21,785	9	0,01	0,05	0,043	0,982	0,989
Faktor 3	3,327	2	0,19	0,034	0,026	0,99	0,997

Anhang E

Tabelle D: Summarische Kennwerte der Q_3-Statistik nach Dimensionen

Dimension	Minimum	1. Quartil	Median	Mittelwert	3. Quartil	Maximum
Faktor 1	-0,268	-0,141	-0,106	-0,095	-0,044	0,025
Faktor 2	-0,284	-0,235	-0,138	-0,156	-0,075	-0,042
Faktor 3	-0,342	-0,302	-0,263	-0,236	-0,174	-0,085

Tabelle E: Q_3-Residualkorrelationsmatrix für Faktor 1

Item	A1	A3	A4	A5	A7	A8	A9	A10
A1	1,000							
A3	0,016	1,000						
A4	0,002	-0,040	1,000					
A5	-0,051	-0,092	0,025	1,000				
A7	-0,082	-0,140	-0,045	-0,052	1,000			
A8	0,006	-0,025	-0,114	-0,121	0,024	1,000		
A9	-0,102	-0,196	-0,149	-0,171	-0,159	-0,268	1,000	
A10	-0,091	-0,205	-0,111	-0,116	-0,139	-0,116	-0,144	1,000

Tabelle F: Q_3-Residualkorrelationsmatrix für Faktor 2

Item	A11	A12	A14	A15	A17	A18
A11	1,000					
A12	-0,042	1,000				
A14	-0,119	-0,258	1,000			
A15	-0,115	-0,278	-0,058	1,000		
A17	-0,138	-0,181	-0,043	-0,284	1,000	
A18	-0,057	-0,092	-0,263	-0,211	-0,198	1,000

Tabelle G: Q_3-Residualkorrelationsmatrix für Faktor 3

Item	A2	A6	A13	A16
A2	1,000			
A6	-0,085	1,000		
A13	-0,280	-0,150	1,000	
A16	-0,246	-0,309	-0,342	1,000

Anhang F

Tabelle H: Tabelle mit den vollständigen $S-X^2$-Fit-Indizes

Item	$S-X^2$	$df(S-X^2)$	$RMSEA(S-X^2)$	$p(S-X^2)$
Faktor 1				
A1	14,651	11	0,024	0,199
A3	16,256	13	0,021	0,236
A4	21,719	19	0,016	0,298
A5	24,689	16	0,031	0,075
A7	11,689	10	0,017	0,306
A8	8,805	9	0,000	0,455
A9	7,672	14	0,000	0,906
A10	11,680	17	0,000	0,819
Faktor 2				
A11	36,278	17	0,044	0,004
A12	27,076	21	0,022	0,168
A14	31,126	19	0,033	0,039
A15	14,473	15	0,000	0,490
A17	15,425	15	0,007	0,421
A18	23,115	17	0,025	0,146
Faktor 3				
A2	4,479	5	0,000	0,483
A6	7,108	3	0,049	0,069
A13	13,482	9	0,029	0,142
A16	5,827	8	0,000	0,667

Tabelle I: Parameterschätzungen für das GPCM-EFA-Modell

Item	Faktor	Diskrimination (a)	b1	b2	b3	b4
A1	F1	0,785	-2,910	NA	NA	NA
A2	F3	1,481	-0,444	NA	NA	NA
A3	F1	1,403	-2,596	-1,355	NA	NA
A4	F1	0,649	-3,070	-1,132	NA	NA
A5	F1	0,353	3,948	-1,784	-1,532	-7,020
A6	F3	0,512	1,247	-4,873	2,627	-4,884
A7	F1	1,302	-1,623	NA	NA	NA
A8	F1	1,282	-1,012	NA	NA	NA
A9	F1	1,506	-0,720	-0,736	NA	NA
A10	F1	1,076	-0,822	-1,683	NA	NA
A11	F2	0,618	0,564	-3,562	-2,259	-3,248
A12	F2	0,901	-1,992	-1,913	-0,112	-0,586
A13	F3	1,676	-2,107	-0,433	NA	NA
A14	F2	1,244	-1,350	-1,052	NA	NA
A15	F2	1,720	-1,216	-0,275	NA	NA
A16	F3	1,795	-1,343	-1,343	NA	NA
A17	F2	1,314	-0,755	0,415	NA	NA
A18	F2	1,322	-1,179	0,346	NA	NA

Anhang G

Abbildung B: CRC-Kurven der IRT-Modellierung für Faktor 1

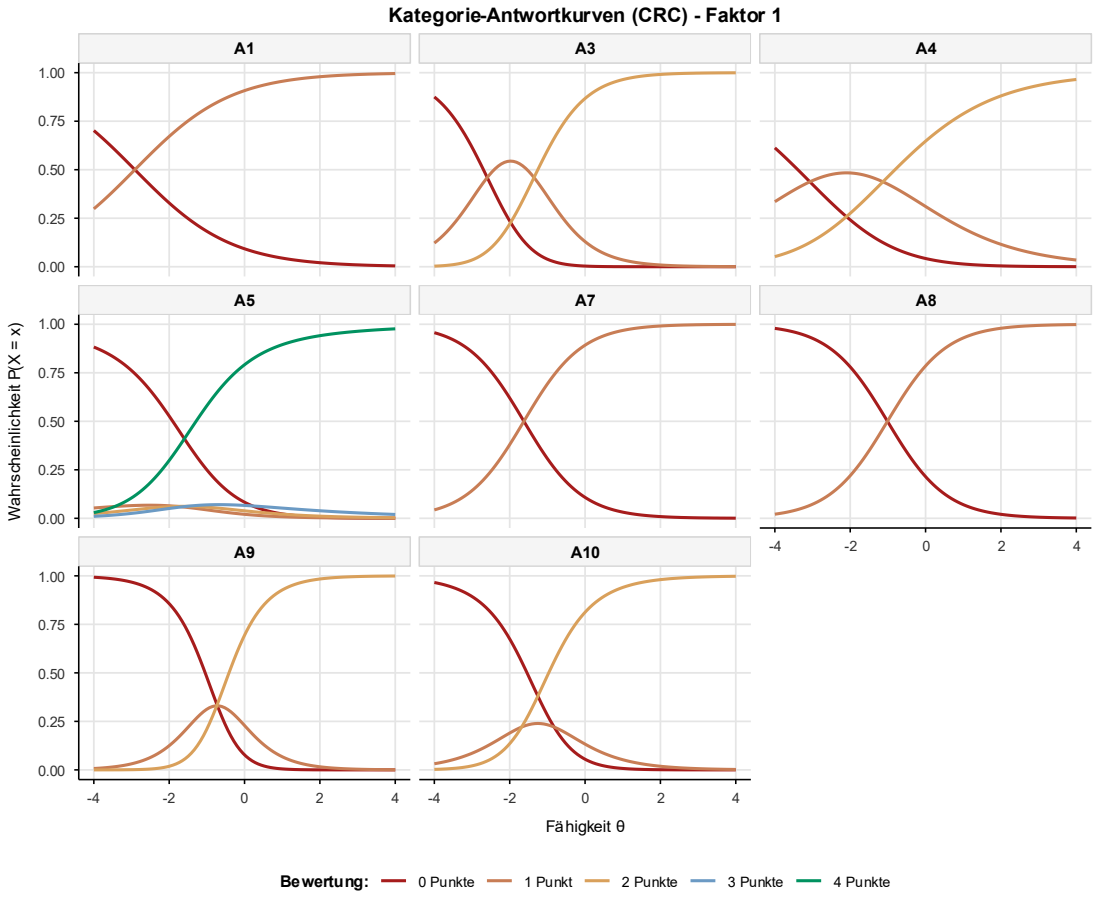


Abbildung C: CRC-Kurven der IRT-Modellierung für Faktor 2

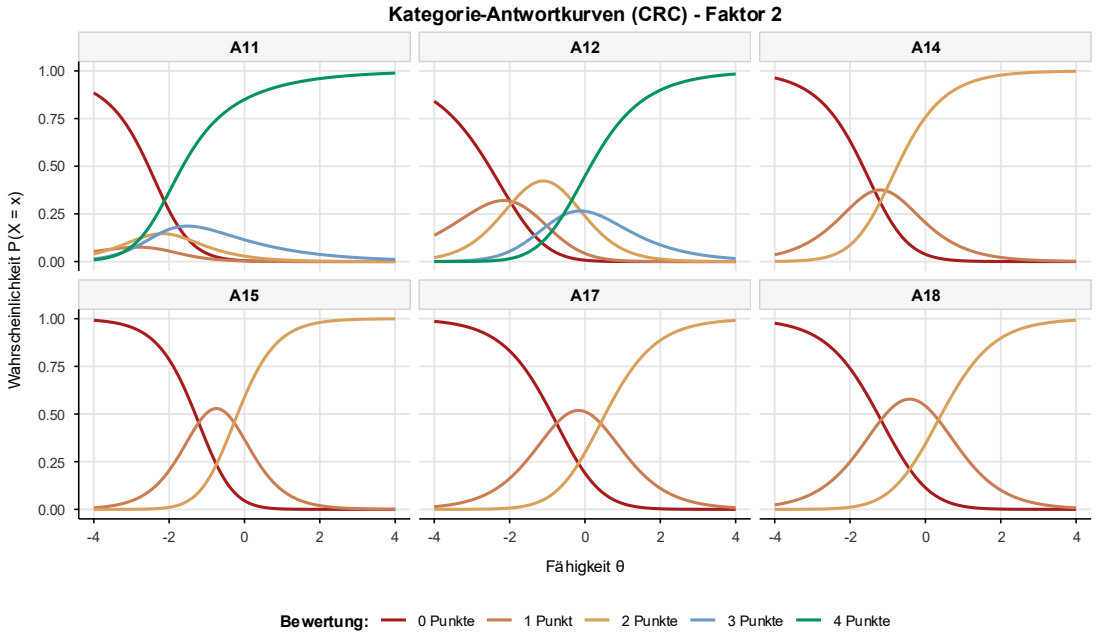
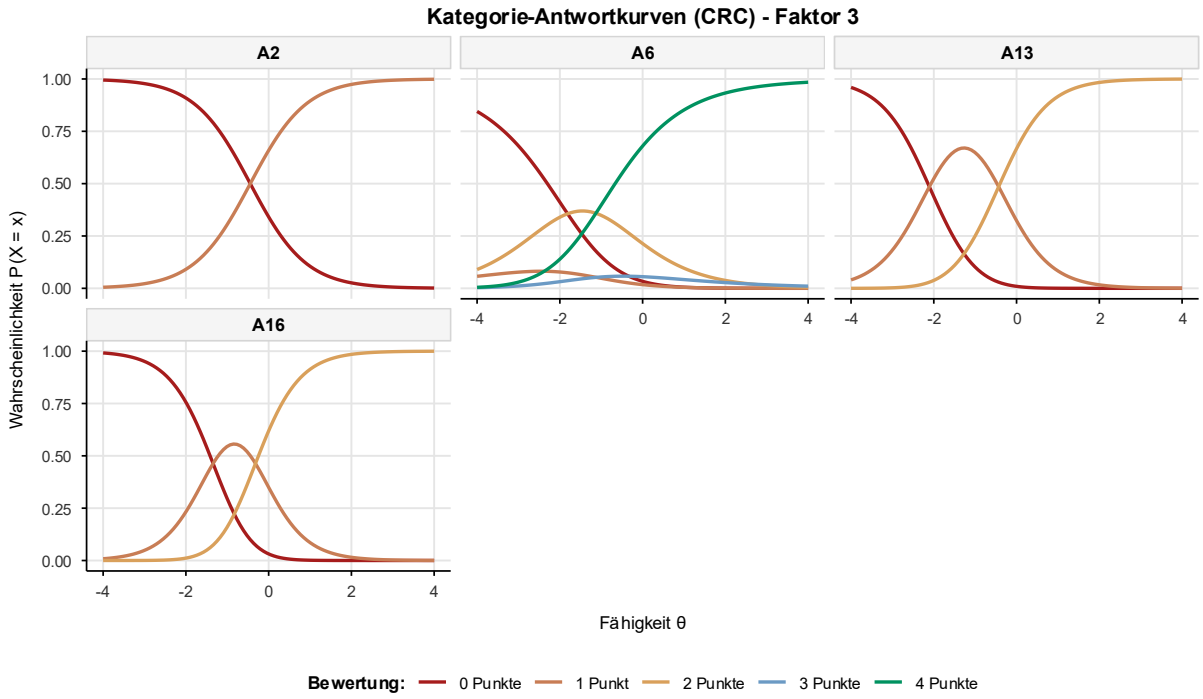


Abbildung D: CRC-Kurven der IRT-Modellierung für Faktor 3

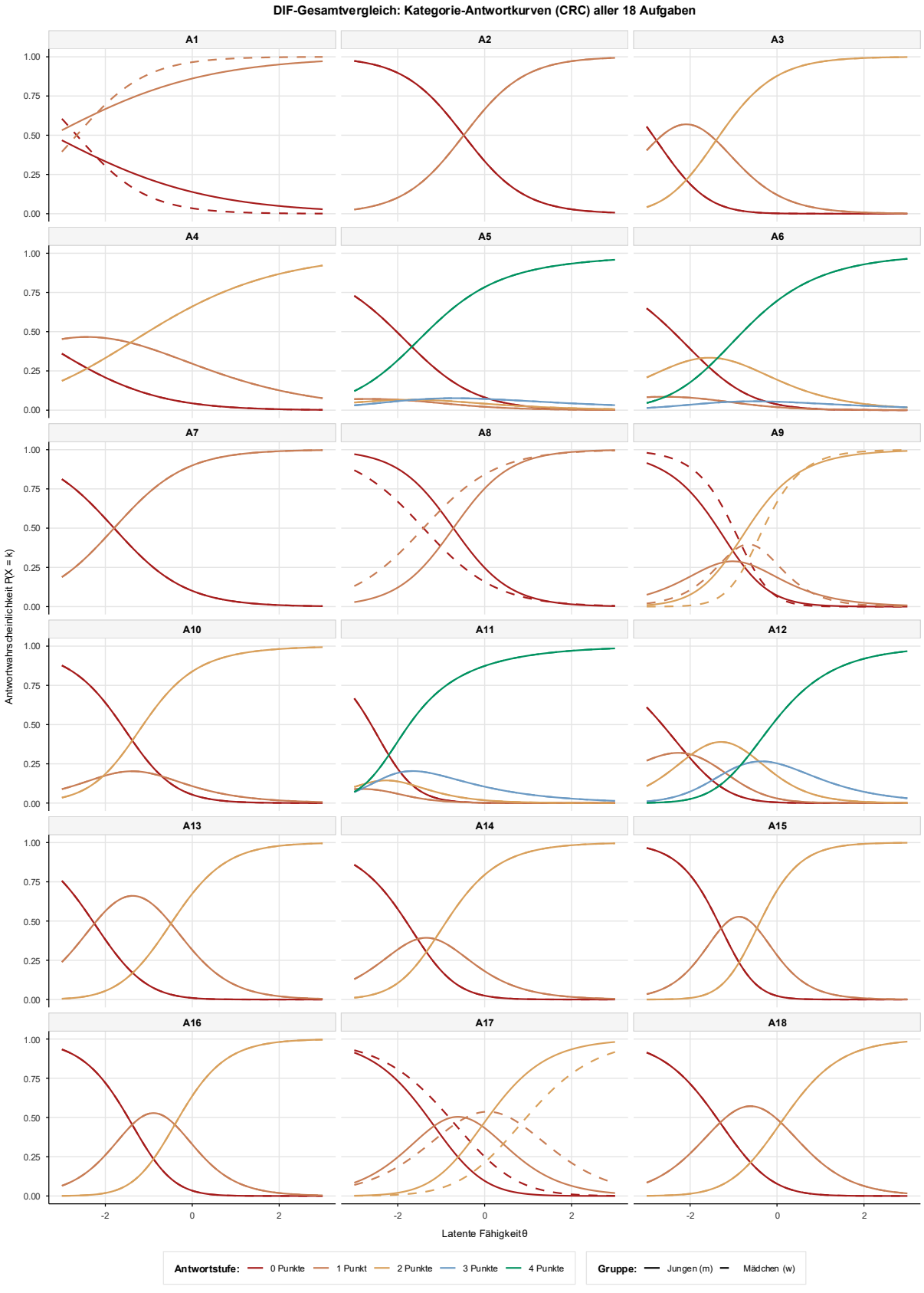


Anhang H

Tabelle J: Ergebnissen der DIF-Analyse

Item	Groups	Converged	ΔAIC	ΔBIC	p	Adj-p
A1	m,w	TRUE	-8,861	-0,568	0,002	0,015
A2	m,w	TRUE	0,496	8,788	0,173	0,284
A3	m,w	TRUE	2,760	15,199	0,356	0,458
A4	m,w	TRUE	2,597	15,036	0,334	0,458
A5	m,w	TRUE	7,924	28,656	0,839	0,893
A6	m,w	TRUE	5,283	26,015	0,451	0,542
A7	m,w	TRUE	3,691	11,984	0,857	0,893
A8	m,w	TRUE	-6,544	1,749	0,005	0,031
A9	m,w	TRUE	-5,619	6,820	0,009	0,040
A10	m,w	TRUE	-3,881	8,557	0,02	0,071
A11	m,w	TRUE	3,382	24,113	0,251	0,376
A12	m,w	TRUE	8,333	29,065	0,893	0,893
A13	m,w	TRUE	-2,572	9,867	0,036	0,107
A14	m,w	TRUE	-2,077	10,362	0,044	0,114
A15	m,w	TRUE	0,225	12,664	0,123	0,230
A16	m,w	TRUE	0,307	12,746	0,128	0,230
A17	m,w	TRUE	-17,024	-4,585	0	0,001
A18	m,w	TRUE	-1,535	10,904	0,057	0,127

Abbildung E: CRC-Kurven für alle 18 Aufgaben der DIF-Analyse



Anhang I

Tabelle K: Werten für den CFA-Modellvergleich für alle fünf Modelle

Modell	Robust CFI	Robust TLI	Robust RMSEA	SRMR
Einfaktormodell	0,879	0,862	0,079	0,064
Theoretisches Modell	0,898	0,882	0,073	0,060
EFA-3-Faktor-Modell	0,943	0,934	0,055	0,051
Theoretisches-Bifaktor-Modell		Konvergiert nicht		
EFA-Bifaktor-Modell	0,968	0,959	0,043	0,043

Anmerkung. N = 574. CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual.

Anhang J

In der Auswertungsphase wurde KI eingesetzt, um Probleme mit dem R-Code zu lösen oder in schwierigen Fällen statistische Methoden zur Beantwortung der forschungsleitenden Fragen vorgeschlagen zu kommen. Hierzu wurden die vorhandenen Probleme bzw. der R-Code geteilt und die gewünschte Lösung erläutert.

Im gesamten Verlauf der Arbeit diente KI als „Sparringspartner“ genutzt und mit ihm Ergebnisse diskutiert, um ein Feedback zu den eigenen Erkenntnissen zu erhalten und evtl. weitere Sichtweisen aufgezeigt zu bekommen, die dann eigenständig durchdacht und geprüft wurden.

Zum Abschluss wurde die generative KI auch als Lektor der Arbeit genutzt. Hierzu wurde der folgende Prompt genutzt:

Du bist der Lektor meiner Masterarbeit. Hier kommt Kapitel x. Prüfe Grammatik, Rechtschreibung und Zeichensetzung und optimiere hinsichtlich Struktur, Klarheit und Lesefluss. Die Aussage bleibt unverändert. Text: [Text]

Die verwendete KI hierfür war Gemini.

Impressum

Mathematikdidaktik im Kontext

ISSN 2568-0331

Heft 13

**Theoretische Struktur vs. empirische Realität –
Eine empirische Untersuchung des BRT 2 zur dimensionalen Architektur und seiner Eignung als Instrument zur Diagnostik von Rechenschwäche in der 2. Jahrgangsstufe**

Bayreuth, 2026

Elektronische Fassung unter:
https://epub.uni-bayreuth.de/view/series/Mathematikdidaktik_im_Kontext.html

Autor

Sebastian Rampp

Herausgeber

Carsten Miller und Volker Ulm
Universität Bayreuth
Lehrstuhl für Mathematik und ihre Didaktik
Universitätsstraße 30
95440 Bayreuth

https://www.dmi.uni-bayreuth.de