4. Validierung und Revision der ECON-2022-Assessmentumgebung

4.1 Einleitung

ECON-2022-Projektteam

Zur Überprüfung der Zuverlässigkeit und Aussagekraft der ECON-2022-Assessmentumgebung wurden verschiedene Validierungsprozesse durchlaufen. Abbildung 4.1.1 visualisiert die drei zu validierenden Instrumentarien des TBA-EL.

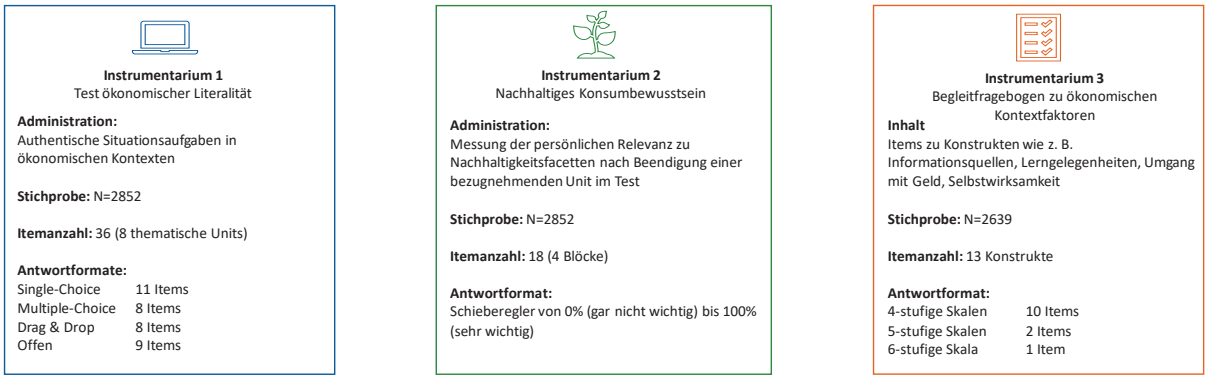


Abbildung 4.1.1: ECON-2022-Assessmentumgebung

Unterkapitel 4.2 stellt ex ante Konstruktionskriterien des Instrumentariums 1 vor. Eine besondere Rolle spielen hierbei schwierigkeitsgenerierende Merkmale und übergeordnete Testcharakteristika, die von Expert*innen bewertet wurden.

Im Anschluss daran liegt der Fokus im Unterkapitel 4.3 auf der quantitativen Validierung des Instrumentariums 1. Diese beinhaltet eine ausführliche Datendeskription, die Untersuchung von Itemfits und die Analyse empirischer Item- und Personenschwierigkeiten. Weiterhin wird auf die datengestützte Revision der Testitems eingegangen. Es wird eine vergleichende Perspektive der Feld- und Hauptstudiendaten eingenommen.

Abschließend beschäftigt sich Unterkapitel 4.4 mit Instrumentarium 2 und 3 und hier insbesondere mit der Validierung der Einstellungsfragen und des Fragebogens. Dies umfasst die Datendeskription, die Prüfung der Dimensionalität und Reliabilitäten der Skalen und die durchgeführten Revisionen der Items in diesen Instrumenten im Vergleich von Feld- zu Hauptstudie.

ECON-2022-Projektteam (2024). Einleitung. In E. Winther & H. J. Abs (Hrsg.), *ECON 2022. Ökonomische Bildung in Jahrgang 8: Konzeption, Kompetenzen und Einstellungen* (S. 109). Waxmann. CC BY 4.0. https://doi.org/10.31244/9783830998389.4_1

4.2 Konstruktionskriterien des Testinstruments in der Expertenvalidierung

Nina Johanna Welsandt, Fenna Henicz, Fabio Fortunati, Esther Winther & Hermann Josef Abs

4.2.1 Stichprobe

Expertengestützte Validitätsprüfung ECON 2022

Für die Überprüfung der Validität wurden Konstruktionskriterien des Instrumentariums in eine Expertenbefragung gegeben. Die Expert*innen haben die konstruierten Testitems einerseits entlang schwierigkeitsgenerierender Merkmale bewertet, um ex ante Vorstellungen davon zu entwickeln, welches Leistungsspektrum das Instrumentarium erfassen kann. Andererseits wurden Urteile über spezifische Testcharakteristika – hier: authentische Item- und Testgestaltung sowie Aspekte der Usability – eingeholt. Die Expert*innen (n = 25) repräsentieren Fachwissen aus drei Handlungsfeldern; sie bringen Expertisen aus den Forschungsbereichen Testentwicklung (n = 10), Wirtschaftswissenschaften/Wirtschaftspädagogik beziehungsweise Wirtschaftspsychologie (n = 11) sowie Schule und Unterricht (n = 12) ein.

Multidisziplinäre Expertisen der Expert*innen

4.2.2 Ratings der schwierigkeitsgenerierenden Merkmale

Ein Ex-ante-Rating der Itemschwierigkeiten

Ein Ex-ante-Rating der Itemschwierigkeiten erhöht die Wahrscheinlichkeit, das zu erfassende Konstrukt in angemessener Breite abbilden zu können. Die Expert*innen haben die Itemschwierigkeiten entlang dreier Merkmale bewertet: (1) inhaltliche Spezifität, (2) kognitive Beanspruchung sowie (3) funktionale Modellierung (Klotz et al., 2015; Winther, 2010; Beck, 2020):

- Die inhaltliche Spezifität bewertet Testaufgaben hinsichtlich des zu ihrer Lösung benötigten (Fach-)Wissens. Die Aufgabenkonstruktion wird hier auf *Inhaltsebene* beurteilt.
- Die Art der kognitiven Beanspruchung bewertet Testaufgaben mit Blick auf die zu ihrer Lösung eingeforderten Leistungsfähigkeiten der Schüler*innen. Die Aufgabenkonstruktion wird hier auf kognitiver *Prozessebene* beurteilt.
- Die funktionale Modellierung bewertet Testaufgaben dahingehend, wie anspruchsvoll es ist, die zur Lösung notwendigen Schritte aus der zugrundeliegenden Anforderungssituation zu extrahieren. Die Aufgabenkonstruktion wird hier auf *Kontextebene* beurteilt.

Konstruktionsprozess

Die systematische Konstruktion von Testaufgaben entlang der o. g. schwierigkeitsgenerierenden Merkmale stellt sicher, dass (1) Annahmen über kognitive Theorien in die Testaufgaben einfließen und (2) Testaufgaben formuliert werden, die gut zwischen den zu testenden Personen trennen können, da sie unterschiedliche Fähigkeitsstufen ansprechen.

Die nachfolgende Abbildung 4.2.1 zeigt das Ratingschema, das von den Expert*innen zur Schwierigkeitsprognose genutzt wurde. Die drei schwierigkeitsgenerierenden Merkmale differenzieren zwischen jeweils drei Schwierigkeitsstufen, wobei Stufe 1 für eine geringe Schwierigkeit und Stufe 3 für eine hohe Schwierigkeit der Testaufgabe steht.

Welsandt, N. J., Henicz, F., Fortunati, F., Winther, E. & Abs, H. J. (2024). Konstruktionskriterien des Testinstruments in der Expertenvalidierung. In E. Winther & H. J. Abs (Hrsg.), *ECON 2022. Ökonomische Bildung in Jahrgang 8: Konzeption, Kompetenzen und Einstellungen.* Waxmann (S. 110–116). CC BY 4.0. https://doi.org/10.31244/9783830998389.4_2

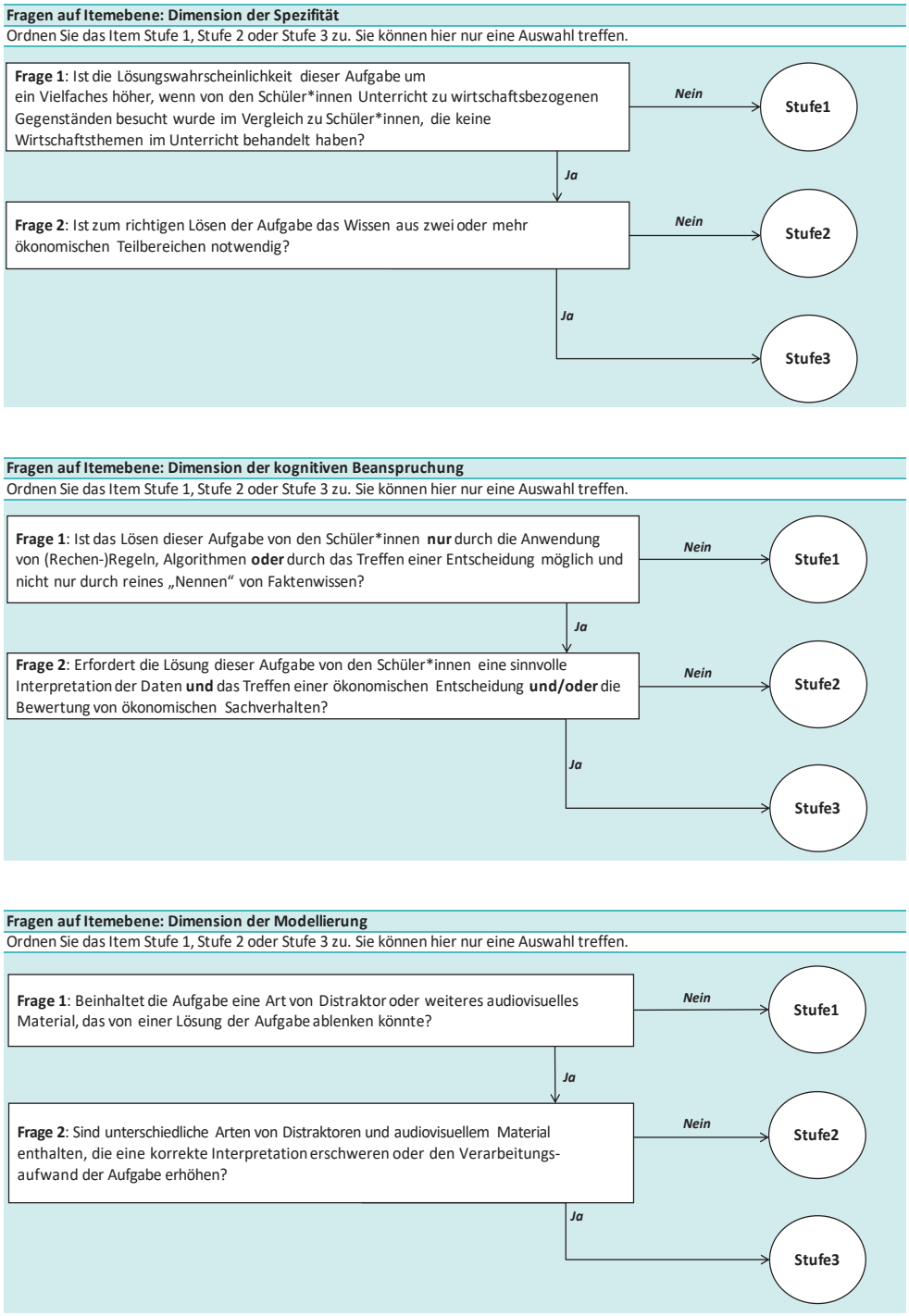


Abbildung 4.2.1: Ratingschema der schwierigkeitsgenerierenden Merkmale

Verteilung der Itemschwierigkeiten

Ein gutes, d.h. zwischen den Fähigkeiten der Schüler*innen hinreichend diskriminierendes, Testinstrument weist eine ausgewogene Verteilung der Itemschwierigkeiten auf. Für den konstruierten Test liegen auf Basis der Urteile der Expert*innen ex ante Schwierigkeitsprognostiken vor, die – wie in Abbildung 4.2.2 dargestellt – ein ausgewogenes Verhältnis von leichten, mittelschweren und schweren Testitems erwarten lassen.

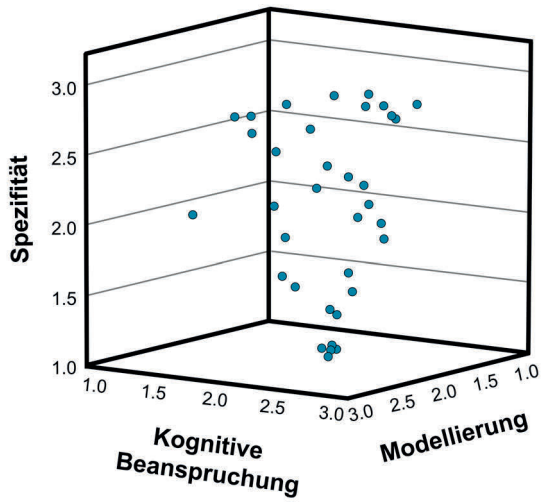
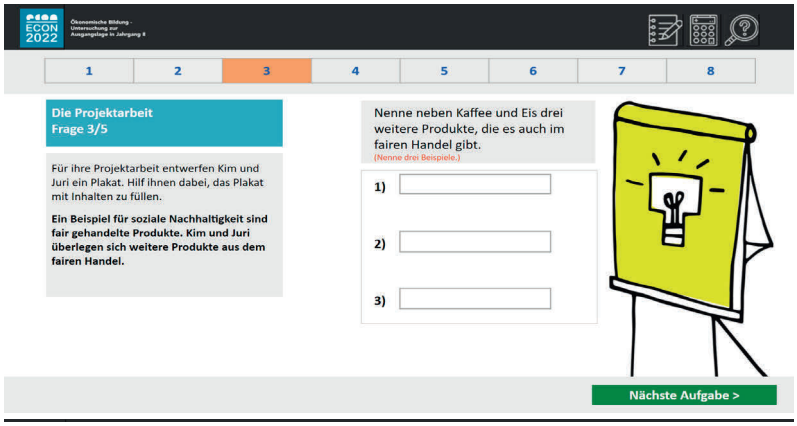


Abbildung 4.2.2: Ex-ante-Beurteilung der Itemschwierigkeiten

Abbildung 4.2.2 illustriert dreidimensional die Bewertungen der schwierigkeitsgenerierenden Merkmale Spezifität, kognitive Beanspruchung und Modellierung. Die durch Punkte dargestellten Werte der x-, y- und z-Achse stellen das durchschnittliche Rating der einzelnen Items dar. Zur Veranschaulichung werden in Abbildung 4.2.3 Beispiele verschieden schwieriger Testitems präsentiert.

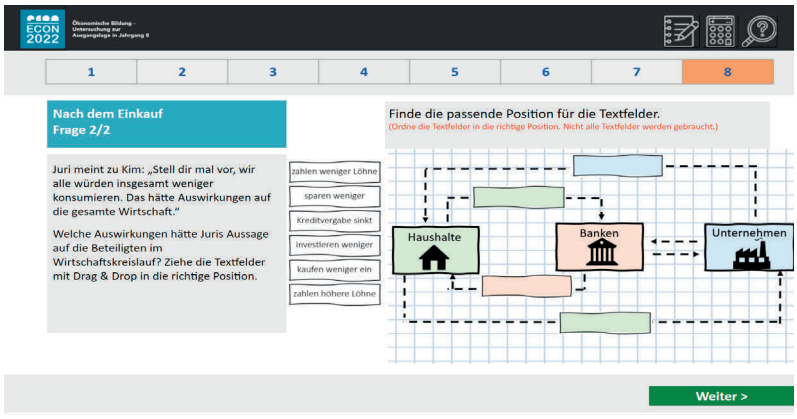
4.2.3 Ratings testcharakteristischer Merkmale

Neben einer schwierigkeitsbeschreibenden Beurteilung wurden die Expert*innen gebeten, testcharakteristische Merkmale zu bewerten. Hierzu war zunächst die authentische Gestaltung der Testinhalte und dann die Usability des Tests einzuschätzen. Zur Bewertung der Authentizität wurde ebenfalls ein dreistufiges Ratingschema genutzt (Abbildung 4.2.4): Es wird davon ausgegangen, dass lebensweltnahe Handlungssituationen eine Aufgabe besser zugänglich machen. Dafür müssen authentische Situationen modelliert werden. Stufe 1 wird erreicht, wenn das Item der Zielgruppe aus dem Alltag vertraut ist. Ist die Situation zumindest theoretisch zugänglich, kann sie Stufe 2 zugeordnet werden. Ein Item fällt unter Stufe 3, wenn nicht erwartet werden kann, dass die Zielgruppe einer solchen Situation im Alltag perspektivisch begegnen kann.



Prognostisch leichtes Testitem 3_3: Fair-Trade Produkte
Die Schüler*innen müssen in einem offenen Format Fair-Trade-Produkte aufzählen. Das Nennen der Produkte, sprich die Anwendung von Faktenwissen, ermöglicht eine korrekte Lösung der Aufgabe. Dabei beinhaltet die Aufgabe keine weiteren Distraktoren. Spezifität (M=1.37), kognitive Beanspruchung (M=1.12) und auch Modellierung (M=1.06) wurden in der durchschnittlichen Bewertung der ersten Stufe zugeordnet und somit als leicht eingestuft.

Prognostisch mittelschweres Testitem 1_2: Bedürfnisse und Bedarf
In dieser Hotspot-Aufgabe können durch Anklicken Produkte von der Einkaufsliste gestrichen werden, die nicht zum täglichen Bedarf zählen. Für das Lösen der Aufgabe ist es hilfreich, Unterricht zu wirtschaftsbezogenen Gegenständen besucht zu haben. Weiterhin reicht das reine Faktenwissen bei dieser Aufgabe nicht mehr aus. Das Testitem wurde in allen Merkmalen auf Stufe 2 geratet: Spezifität (M=1.83), kognitive Beanspruchung (1.73), Modellierung (1.90).



Prognostisch schweres Testitem 8_2: Wirtschaftskreislauf
Über Drag-&-Drop-Felder müssen die Verbindungen im Wirtschaftskreislauf passend hergestellt werden. Dafür muss das Wissen mehrerer Teilbereiche kombiniert angewendet werden. Die bereits vorgegebenen Begriffe sind zu interpretierten und Entscheidungen sind zu treffen. Unterschiedliche Arten von Distraktoren erhöhen die Schwierigkeit zusätzlich. Spezifität (M=2.96), kognitive Beanspruchung (M=2.82) und Modellierung (M=2.48) wurden im Mittel der Stufe 3 zugeordnet.

Abbildung 4.2.3: Beispiele unterschiedlich schwerer Testitems

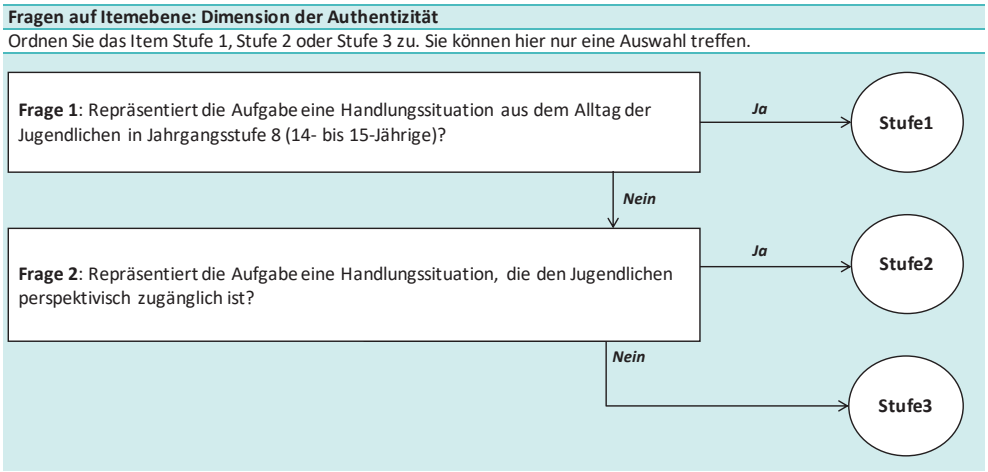


Abbildung 4.2.4: Beurteilung der Authentizität

Das Expertenrating zeigt, dass die Expert*innen die Aufgaben größtenteils als authentisch einschätzten. Eine Ausnahme bilden die Aufgaben 8_1 und 8_2. Die Grenzwerte wurden aus den Bewertungen der Expert*innen abgeleitet. So bilden nur die Aufgaben 8_1 und 8_2 laut den Expert*innen keine perspektivisch zugängliche Handlungssituation für Schüler*innen ab. Die Werte lagen hier über 2.30. 16 Aufgaben erhielten mit Werten zwischen 1.51 und 2.30 eine Bewertung mittlerer Authentizität. 17 Aufgaben wurden mit Werten unter 1.5 als sehr authentisch eingestuft.

Beurteilung der Usability

Zur Beurteilung der Usability auf Testebene wurden den Expert*innen am Ende der Befragung 15 Fragen gestellt (siehe Tabelle 4.2.1), bei denen zu jeder Aussage angegeben werden sollte, inwieweit diese in Bezug auf die Zielgruppe von Achtklässler*innen persönliche Zustimmung findet. Dabei wurde ein vierstufiges Antwortformat mit den Antwortmöglichkeiten 1 = stimme nicht zu, 2 = stimme eher nicht zu, 3 = stimme eher zu und 4 = stimme zu gewählt. Die Expert*innen bewerteten verschiedene Aspekte der Benutzerfreundlichkeit des Assessments.

Die Ergebnisse dieser Expertenbefragung lieferten wertvolle Einblicke in die Usability des TBA-EL aus der Perspektive von Fachexpert*innen. Die meisten Aspekte der Benutzerfreundlichkeit wurden positiv bewertet, während es einige Bereiche gibt, die weiterhin Aufmerksamkeit erfordern. Diese Ergebnisse bieten eine Grundlage für gezielte Verbesserungen des Assessments und eine Optimierung seiner Benutzerfreundlichkeit. Die Ergebnisse dieser Befragung werden im Folgenden präsentiert:

Befunde zu den Bewertungen der Usability

- Intuition des Userinterfaces: Die Auswertung von 21 Bewertungen ergab einen Durchschnittswert (M) von 3.71. Dies deutet darauf hin, dass die Expert*innen die Nutzung des Userinterfaces im Allgemeinen als intuitiv empfinden.
- Intuition der Funktionen einzelner Buttons: Auch hier wurde ein Durchschnittswert von 3.71 ermittelt. Dies legt nahe, dass die Expert*innen die Funktionen der Buttons überwiegend als intuitiv wahrnehmen.
- Verständnis eingesetzter Gestaltungselemente: Hier ergab sich basierend auf den Einschätzungen von 21 Expert*innen ein Durchschnittswert von 3.57, der darauf schließen lässt, dass die Gestaltungselemente für die Zielgruppe angemessen ausgewählt wurden.
- Verständnis der zu bearbeitenden Aufgaben: Basierend auf den Einschätzungen von 21 Expert*innen wurde hierbei ein Durchschnittswert von 3.62 er-

mittelt. Die zu bearbeitenden Aufgaben wurden demnach zielgruppengerecht eingeführt und erstellt.

- Häufigkeit von Überraschungen während der Aufgabenbearbeitung: Hier zeigte sich basierend auf den Einschätzungen von 20 Expert*innen ein Durchschnittswert von 2.05.
- Mehrwert des Erklärvideos als Unterstützung: Die Auswertung von 20 Bewertungen ergab einen Durchschnittswert von 3.15. So wurde der Aussage „Durch das Erklärvideo fällt mir der Umgang mit dem Assessment leichter" durchschnittlich nur eher zugestimmt. Infolgedessen wurde das Erklärvideo noch mal auf Funktionalität überprüft und angepasst.
- Offensichtlichkeit relevanter Informationen für die Aufgabenbearbeitung: Hierbei wurde basierend auf den Einschätzungen von 20 Expert*innen ein Durchschnittswert von 3.35 ermittelt. So konnte aus der Befragung abgeleitet werden, dass die relevanten Informationen teilweise nicht direkt ersichtlich gewesen sind. Dies führte zu einer Überprüfung und Anpassung des Aufgabendesigns.
- Ablenkung durch Gestaltungselemente. Die Auswertung von 21 Bewertungen ergab einen Durchschnittswert von 3.0. Gestaltungselemente wurden demnach teilweise als ablenkend empfunden und wurden daher überarbeitet. In Anbetracht der Modellierung einer Aufgabe dürfen Gestaltungselemente nur als Ablenkung empfunden werden, wenn diese gezielt als Distraktoren eingesetzt wurden. Ist dies nicht der Fall, müssen die Gestaltungselemente verändert werden.
- Lesbarkeit der Arbeitsaufträge des Assessments sowie der verwendeten Materialien: Die Auswertung ergab basierend auf den Einschätzungen von 21 Expert*innen Durchschnittswerte von 3.57 und 3.19. Aus dem Durchschnittswert lässt sich ableiten, dass Elemente nicht durchweg gut lesbar waren und dass eingesetzte Materialien unübersichtlich waren. Auch hier musste es Anpassungen geben.
- Lesbarkeit der Untertitel in den Videoelementen des Assessments: Hierbei wurde basierend auf den Einschätzungen von 20 Expert*innen ein Durchschnittswert von 3.55 ermittelt.
- Übersichtlichkeit der eingesetzten Materialien: Dies ergab basierend auf den Einschätzungen von 21 Expert*innen einen Durchschnittswert von 3.38.
- Erleichterung der Aufgaben durch gewähltes Design: Hierbei wurde basierend auf den Einschätzungen von 20 Expert*innen ein Durchschnittswert von 3.2 ermittelt. Das Design kann die Bearbeitung einer Aufgabe inhaltlich nicht vereinfachen. Trotzdem sollte es eine Aufgabe auch nicht erschweren.
- Verständlichkeit für Schüler*innen der 8. Klasse: Dies ergab basierend auf den Einschätzungen von 21 Expert*innen einen Durchschnittswert von 3.38.
- Angemessenheit der Distraktoren für Schüler*innen der 8. Klasse: Die Auswertung ergab basierend auf den Einschätzungen von 21 Expert*innen einen Durchschnittswert von 3.19. Die Einschätzung der Expert*innen wies darauf hin, dass Schüler*innen nicht vollkommen klar sein könnte, was bei den Aufgaben von ihnen gefordert wird. Dies geht mit der Aussage, dass die Distraktoren für Achtklässler*innen nur eher als angemessen empfunden wurden, einher. Auch hier wurde überprüft, welche der Distraktoren angepasst werden können.

Tabelle 4.2.1: Gesamtbewertung der Usability-Fragen auf Testebene

	Expertenrating (ER)				
	1 = stimme nicht zu, 2 = stimme eher nicht zu, 3 = stimme eher zu, 4 = stimme zu				
	N	Min	Max	M	Std. Abweichung
Die Nutzung des Userinterfaces ist intuitiv.	21	3	4	3.71	0.46
Die Funktionen der einzelnen Buttons sind intuitiv.	21	3	4	3.71	0.46
Die eingesetzten Gestaltungselemente habe ich schnell verstanden.	21	2	4	3.57	0.60
Wie ich eine Aufgabe bearbeiten soll, habe ich schnell verstanden.	21	2	4	3.62	0.59
Bei der Bearbeitung der Aufgaben war ich oft überrascht.	20	1	4	2.05	1.19
Durch das Erklärvideo fällt mir der Umgang mit dem Assessment leichter.	20	1	4	3.15	0.99
Für die Bearbeitung der Aufgaben sind alle relevanten Informationen ersichtlich.	20	2	4	3.35	0.75
Die eingesetzten Gestaltungselemente haben mich bei der Bearbeitung der Aufgaben nicht abgelenkt.	21	1	4	3.00	0.89
Die Arbeitsaufträge des Assessments sind gut lesbar.	21	2	4	3.57	0.60
Die Materialien sind gut lesbar.	21	1	4	3.19	0.87
Die Untertitel der Videoelemente sind gut lesbar.	20	1	4	3.55	0.83
Die eingesetzten Materialien im Assessment sind übersichtlich gestaltet.	21	2	4	3.38	0.59
Das Design der Aufgaben hat mir das Bearbeiten erleichtert.	20	2	4	3.20	0.62
SuS der 8. Klasse ist klar, was von ihnen in der Aufgabe gefordert wird.	21	2	4	3.38	0.67
Die Distraktoren sind für SuS der 8. Klasse angemessen.	21	2	4	3.19	0.75

4.2.4 Implikationen der Ergebnisse

Die Datenanalyse der Expertenbefragung zeigt, dass sie sich trotz unterschiedlicher Expertise in ihrer Bewertung der Spezifität, kognitiven Beanspruchung, Authentizität und Modellierung weitgehend einig sind. Basierend auf den Ergebnissen der Expertenbefragung und der Validierung der Feldtestdaten (vgl. Kapitel 4.3) wurde das Prüfinstrument überarbeitet. Besonders schwierige Items wurden zu leichteren Items abgeändert. Distraktoren wurden bspw. der Altersgruppe entsprechend angepasst. Weitergehend wurden überflüssige Elemente bei der Überarbeitung der Modellierung des Assessments entfernt.

In Bezug auf die negativen Authentizitätseinschätzungen des Expertenratings wurden die Aufgaben 8_1 und 8_2 nochmal genauer überprüft. Da es sich bei den beiden Aufgaben um die letzten des Assessments handelte, wurde hier entschieden, den Lebensweltbezug zu lockern, um Aufgaben zu generieren, die stärker auf die Reflexion über wirtschaftliche Systeme ausgerichtet waren.

Literatur

Beck, K. (2020). Ensuring content validity of psychological and educational tests – the role of experts. *FLR*, 1–37. https://doi.org/10.14786/flr.v8i6.517

Klotz, V. K., Winther, E. & Festner, D. (2015). Modeling the development of vocational competence: A psychometric model for economic domains. *Vocations and Learning*, *8*(3), 247–268. https://doi.org/10.1007/s12186-015-9139-y.

Winther, E. (2010). *Kompetenzmessung in der beruflichen Bildung.* Habilitation. Bertelsmann.

4.3 Validierung des Testinstruments anhand der Feldtestdaten

Fabio Fortunati, Nina Johanna Welsandt, Fenna Henicz, Hermann Josef Abs & Esther Winther

Dieses Unterkapitel betrachtet die psychometrischen Eigenschaften des Testinstruments zur Messung ökonomischer Kompetenz in Feld- und Hauptstudie. Zunächst wird kurz das methodische Vorgehen der Datenanalyse für beide Testzeitpunkte sowie der Umgang mit fehlenden Werten erläutert. Die psychometrischen Eigenschaften des Testinstruments werden für die Feld- und Hauptstudie vergleichend betrachtet und es wird untersucht, inwieweit vorgenommene Veränderungen zu einer Verbesserung des Testinstruments geführt haben.

4.3.1 Codebook, Scoring der Items und Interrater-Reliabilität

Codierung der Testitems

Zur Bewertung der Schülerantworten wurde ein Codebook entwickelt, das die korrekten Antworten für die Single- und Multiple-Choice-Items sowie den möglichen Lösungsraum bei Aufgaben mit offenem Antwortformat enthält. In der Feldtesterhebung wurden die Single- und Multiple-Choice-Items auf Grundlage des Datensatzes bewertet, der die Eingaben der Schüler*innen dokumentiert. Die Eingaben der Testteilnehmenden wurden in neue Variablen transformiert, die den jeweilig erreichten Score (Punktwert) gemäß dem Codebook enthalten.

Die offenen Antworten wurden sowohl im Feld- als auch im Haupttest manuell von drei Codierenden bewertet. Die Codierung der Items erfolgte zunächst unabhängig voneinander. Nach einer ersten Prüfung wurde der zuvor im Codebook entwickelte Lösungsraum um weitere korrekte Schülerantworten erweitert und die offenen Items nochmals codiert.

Datenaufbereitung: Interrater-Reliabilität und Ratereffekte

Die Prüfung der Interrater-Reliabilität (IRR) erfolgt mittels Krippendorffs Alpha. Krippendorffs Alpha berechnet die erwartete zufällige Übereinstimmung durch die durchschnittliche Übereinstimmung, wenn alle Codierungen aller Analyseeinheiten miteinander verglichen werden (Krippendorff, 2004; Hayes & Krippendorff, 2007). Ein Vorteil ist, dass eine Untersuchung der Übereinstimmung von zwei oder mehr Personen auf einem variablen Skalenniveau erfolgen kann. Darüber hinaus kann bei der Berechnung von Krippendorffs Alpha der 95 %-Konfidenzintervall angegeben werden, der Aufschluss über die Präzision der Reliabilitätsmessung gibt (Hayes & Krippendorff, 2007). Darüber hinaus ist bspw. Cohens Kappa, im Vergleich zu anderen Reliabilitätsmaßen, ein konservatives Maß, das bei ungleich verteilten Variablen zu tiefen Werten tendiert (Brennan & Prediger, 1981; Zhao et al., 2013). Die Berechnung von Krippendorffs Alpha erfolgt in SPSS (IBM Corp., 2021). Zur Präzision der Berechnung der 95 %-Konfidenzintervalle wurde das Bootstrapping-Verfahren (10.000 Bootstraps) angewendet. Tabelle 4.3.1 zeigt die IRR für die Items mit offenem Antwortformat der Hauptstudie. Die IRR zeigt für alle Items zufriedenstellende Werte an.

Fortunati, F., Welsandt, N. J., Henicz, F., Abs, H. J. & Winther, E. (2024). Validierung des Testinstruments anhand der Feldtestdaten. In E. Winther & H. J. Abs (Hrsg.), *ECON 2022. Ökonomische Bildung in Jahrgang 8: Konzeption, Kompetenzen und Einstellungen* (S. 117–133). Waxmann. CC BY 4.0. https://doi.org/10.31244/9783830998389.4_3

Tabelle 4.3.1: Interrater-Reliabilität für die offenen Items in der Hauptstudie

	ein Drittel der Fälle				
Reliabilität/Item	FT2_2	FT3_1	FT5_4	FT7_1	FT7_4
Krippendorffs Alpha					
Bei 2 Codierenden	0.713	0.856	0.836	0.815	0.803
	Zufallsstichprobe (5 % aller Fälle)				
Bei 3 Codierenden	0.723	0.752	0.656	0.753	0.867

Aufgrund der hohen Stichprobengröße in der Hauptstudie wurden die offenen Aufgaben von einem Codierenden vollständig codiert und von einem Weiteren zu jeweils einem Drittel (ca. 1.000 Fälle). Die Aufgabendrittel variierten dabei pro Item. Zusätzlich wurden die Items durch eine Zufallsziehung in 150 Fällen (ca. 5 %) von allen drei Codierenden bewertet, um etwaige Ermüdungseffekte ausschließen zu können (siehe Tabelle 4.3.1).

Die Ergebnisse zeigen sowohl für die Prüfung von einem Drittel der Fälle als auch der Zufallsstichprobe zufriedenstellende Werte an, die sich gegenüber dem Rating des Feldtests verbesserten ($0.63 \leq \alpha \leq 0.84$). Um Ratereffekte auf die Skalierung des Tests auszuschließen, wurde sowohl für den Feld- als auch für den Haupttest geprüft, inwieweit das Rating einzelner Codierender signifikante Unterschiede in der Itemschwierigkeit hervorruft. Hierzu wurden mittels einer Erweiterung des Partial-Credit-Models von Linacre (1994) Ratereffekte untersucht. Beim Feldtest konnte nur bei einem Item (FT22) ein signifikanter Unterschied zwischen zwei Ratern festgestellt werden; dieses weist einen DIF über 0.426 auf. Beim Haupttest betraf dies dasselbe Item (FT22), welches auch die niedrigste Inter-Rater-Reliabilität aufwies. Dieses Item wurde konsensual nachcodiert.

4.3.2 Umgang mit fehlenden Werten

Fehlende Werte pro Item und gruppenbezogene Unterschiede

Grundsätzlich gibt es für den Umgang mit fehlenden Werten kein pauschales Verfahren. In ECON 2022 betrachten wir für den Haupttest die fehlenden Werte pro Item und untersuchen zudem auf Fallebene, ob einzelne Fälle eine Häufung an fehlenden Werten aufweisen. Darüber hinaus wird geprüft, ob die Quote der fehlenden Werte eines Items von der Positionierung im Test abhängen.

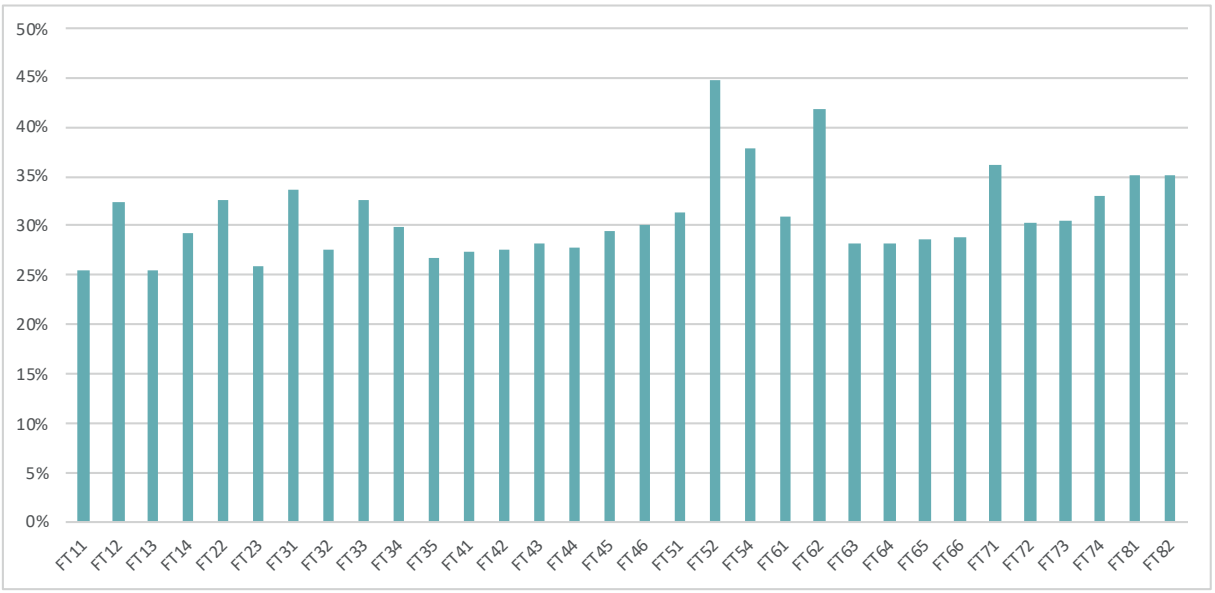


Abbildung 4.3.1: Fehlende Werte auf Itemebene der Hauptstudie

Abbildung 4.3.1 zeigt die fehlenden Werte für jedes Item der Hauptstudie. Die Spannweite reicht dabei von 25.40 % bei Item 1_1 bis 44.83 % bei Item 5_2. Im Mittel können pro Item fehlende Werte in Höhe von 31.03 % festgestellt werden. Zur Überprüfung, ob die Positionierung der Items einen Einfluss auf die Höhe der fehlenden Werte hat, wurde der Test gruppiert: (1) in zwei Hälften und (2) in drei Drittel. Für die Bestimmung signifikanter Gruppenunterschiede wurde für (1) ein t-Test für unabhängige Stichproben (siehe Tabelle 4.3.2) und für (2) eine einfaktorielle Varianzanalyse verwendet (siehe Tabelle 4.3.4). Während der t-Test einen signifikanten Unterschied zwischen den beiden Testhälften zeigt, kann bei einer Einteilung in Drittel kein signifikanter Unterschied mehr festgestellt werden.

Tabelle 4.3.2: T-Test für die fehlenden Werte der Testhälften der Hauptstudie

Variable	**Testhälften**	
Merkmal	Erste Hälfte	Zweite Hälfte
N	16	16
M	0.288	0.331
SD	0.027	0.050
t-Wert	-3.029	
df	30	
p	0.005	
Cohens d	0.040	

Zu beobachten ist, dass insbesondere die Items der Einheiten 5 und 6 von fehlenden Werten betroffen sind (siehe Tabelle 4.3.3). Dies könnte darauf zurückzuführen sein, dass hier eine Häufung von mathematisch bezogenen Aufgaben zu finden ist, die den Schüler*innen häufig schwerer erscheinen. Zur Prüfung, ob ein statistisch signifikanter Zusammenhang mit der Positionierung im Test festzustellen ist, wurde eine Korrelation nach Pearson durchgeführt. Der Befund ist signifikant ($r = 0.372$; $p < 0.036$). In Anbetracht der uneindeutigen Ergebnisse hin-

sichtlich der unterschiedlichen Positionierung und des lediglich mittleren Korrelationskoeffizienten kann daher nicht zweifelsfrei bestimmt werden, ob die Quote an fehlenden Werten mit der Positionierung im Test zusammenhängt oder ob Ermüdungseffekte einen Einfluss auf die Testleistung haben.

Tabelle 4.3.3: ANOVA für die fehlenden Werte der Testdrittel der Hauptstudie

Variable	Merkmal	N	M	SD	F	p
Testdrittel	Erstes Drittel	10	0.295	0.032	1.057	0.360
	Zweites Drittel	11	0.311	0.055		
	Drittes Drittel	11	0.323	0.043		

Ausschlusskriterien für einzelne Fälle

Für eine genauere Untersuchung wurde auf Fallebene geprüft, inwieweit hier einzelne Fälle eine hohe Quote an fehlenden Werten aufweisen. Hierfür wurden verschiedene Schwellenwerte entwickelt.

Tabelle 4.3.4: Fehlende Werte auf Fallebene (Hauptstudie)

	Hauptstudie	75 %-Missing	66 %-Missing	50 %-Missing	33 %-Missing	25 %-Missing
Stichprobe	3020	2852	2841	2807	2696	2540
Betroffene Fälle (kumuliert)		168	179	213	324	480

Tabelle 4.3.4 zeigt die fehlenden Werte auf Fallebene. Daraus kann geschlossen werden, dass die überwiegende Mehrheit der Teilnehmer*innen den Test ernsthaft bearbeitet hat. Bei Fällen, die mehr als 75 % der Aufgaben nicht bearbeitet hat, kann angezweifelt werden, inwieweit zuverlässig auf die Kompetenz des entsprechenden Teilnehmenden geschlossen werden kann. Aus diesem Grund wurden alle Fälle, die mehr als 75 % der Testaufgaben nicht beantwortet haben, von der Analyse ausgeschlossen. Dies betraf 168 bzw. 5.5 % der Fälle in der Hauptstudie und 63 bzw. 7.72 % der Fälle im Feldtest. Vorteilhaft ist hier, dass neben einer zuverlässigeren Schätzung der Personenfähigkeit auch eine linksschiefe Verteilung der Werte verringert wird und bei der Testwertinterpretation nicht vorschnell die Annahme eines zu schweren Tests getroffen wird. Die Befunde des Feldtests zu den fehlenden Werten auf Fallebene reihen sich prozentual betrachtet in die Ergebnisse der Hauptstudie ein.

In den Analysen zur Hauptstudie wird somit von einer Stichprobengröße von N = 2.852 Teilnehmer*innen ausgegangen und während des Feldtests von einer Stichprobengröße von 753 Personen.

4.3.3 Datenanalysemethoden

Modellauswahl

Die in diesem Kapitel vorgestellten Datenanalyseverfahren werden sowohl für die Analyse der Stichprobe des Feldtests sowie des Haupttests verwendet. Die Ergebnisse werden in Kapitel 4.2.4 vergleichend dargestellt, sodass etwaige Änderungen der psychometrischen Eigenschaften des Testinstruments transparenter dargestellt werden können.

Für die Analyse der Daten wurde ein polytomes 1PL-IRT-Modell, das Multidimensional-Random-Coefficients-Multinomial-Logit-Modell (MCMLM) (Adams et al., 1997), gewählt und mit dem Programm ACER ConQuest (Adams et al., 2018) skaliert. Bei der Datenerhebung im Feldtest konnten aufgrund eines Erfassungsfehlers der Testsoftware beim ersten Item 1_1 die Schülerantworten nicht reliabel zu den dargebotenen Antwortoptionen des Items zugeordnet werden, sodass Item 1_1 von der Analyse ausgeschlossen werden musste. In der Analyse der Feldtestdaten können somit nur 34 der 35 Items berücksichtigt werden. Tabelle 4.3.5 stellt übersichtlich dar, welche Analyseverfahren für das Bestimmen der psychometrischen Qualität des Testinstruments verwendet wurden. Vor der Analyse wurde geprüft, inwieweit das Rating der einzelnen Codierenden einen Einfluss auf die Itemschwierigkeit ausübt. Fehlende Schülerantworten wurden mit dem Wert 0 für falsche Antworten codiert und in die Modellberechnungen aufgenommen.

Analyseebene: Itemparameter

Für die Überprüfung der psychometrischen Eigenschaften des Testinstruments wurden zunächst die Personen- und Itemparameter sowie die Messgenauigkeit bestimmt (siehe Tabelle 4.3.5). Dazu wurden die Parameter mit der Marginal-Maximum-Likelihood-Methode (MML) geschätzt (Adams et al., 1997). Als Parameter wurden die Itemschwierigkeit sowie die Personenfähigkeitsschätzer (WLE) und deren Verteilung ermittelt. Darüber hinaus wurde mittels einfaktorieller Varianzanalyse (ANOVA) geprüft, ob sich die Itemschwierigkeit hinsichtlich der Aufgabentypen und Inhaltsbereiche unterscheidet.

Analyseebene: Personenparameter und Reliabilitätskoeffizienten

Die Messgenauigkeit des Tests kann anhand des Standardfehlers der einzelnen Personenparameter sowie der Reliabilitätskoeffizienten der probabilistischen und klassischen Testtheorie geprüft werden. Das zu messende Konstrukt gilt als zuverlässig schätzbar, wenn (1) die Standardfehler der Personenparameter gering und (2) die Reliabilitätskoeffizienten hoch sind (EAP/PV & WLE ≥ .70; Cronbachs $\alpha \geq .70$) (Frey, 2012). Die präzise Schätzung der Personenfähigkeiten ist von unmittelbarer Bedeutung für die valide Testwertinterpretation (American Educational Research Association [AERA] et al., 2014, S. 37ff.). Darüber hinaus wird mit der Person-Separation-Reliabilität (WLE) geprüft, ob die Reproduzierbarkeit der Personenparameter gewährleistet ist. Zudem soll mit der Messung der Item-Separation-Reliabilität untersucht werden, ob der Test tatsächlich zwischen leichten und schwierigen Items unterscheiden kann. Ebenfalls soll mit der Bestimmung der testcharakteristischen Kurve (TCC) untersucht werden, ob ein Zusammenhang zwischen den summierten Personen-Testwerten und der latenten Personenfähigkeit existiert (Rost, 2004). Die TCC wird konzeptionell als die Regression der summierten Antwortscores der Testteilnehmenden verstanden und kann grafisch als die Summe aller itemcharakteristischen Kurven (ICC) betrachtet werden. Der empirische Zusammenhang muss einen streng monotonen, hohen Zusammenhang aufweisen.

Analysebene: Itemhomogenität

Zur Beurteilung der Itemhomogenität wurden zunächst die Itemfitwerte des Weighted-Mean-Squares (wMNSQ) und die T-Werte als Indizes für die Qualität der Items herangezogen sowie grafisch nach Auffälligkeiten in den itemcharakteristischen Kurven (ICC) untersucht (Winther, 2010). Darüber hinaus wurden auch Maße der KTT, wie die Trennschärfe, berücksichtigt.

Tabelle 4.3.5: Analyseebenen und methodisches Vorgehen

Analyse-ebene	Itemparameter	Personen-parameter	Reliabilität	Itemhomogenität
Analyse-methoden	Personen-Item-Map		Kennwerte der IRT	Itemfit-Werte
	Verteilung der Item-schwierigkeiten	Verteilung der Personenparameter	Kennwerte der KTT	Grafische Analyse (ICC)
	Itemschwierigkeit nach Aufgabentyp	Testcharakteristi-sche Kurve (TCC)	Testinformations-kurve (TIF)	DIF-Analysen
	Itemschwierigkeit nach Inhaltsbereich			

Mithilfe einer DIF-Analyse wurde im Anschluss geprüft, ob bei gleicher Personenfähigkeit Unterschiede in der Lösungswahrscheinlichkeit von Items hinsichtlich eines personenbezogenen Merkmals bestehen. Dies dient zur Ermittlung der Testfairness über verschiedene Subgruppen hinweg (Paek, 2002; Teresi et al., 2008). Zunächst wurde auf Testebene untersucht, ob signifikante Gruppenunterschiede hinsichtlich der Hintergrundvariablen zu finden sind. Im zweiten Schritt wurde der Interaktionsterm auf Signifikanz geprüft, darüber hinaus wurde auf Itemebene untersucht, ob der DIF-Schätzer einzelner Items einer Subgruppe sich signifikant von der absoluten Itemschwierigkeit unterscheidet. Hierfür wurde mittels des Wald-Tests ein Chi-Quadrat-Wert für einen Freiheitsgrad ermittelt, um so auf Itemebene Signifikanz ermitteln zu können (Kirsch, 2021). Signifikante DIF-Unterschiede können als Indiz für das Verletzten der Testfairness gewertet werden. Für die DIF-Analysen wurden die Merkmale Geschlecht, Zuwanderungsgeschichte sowie die zu Hause gesprochene Sprache der Schüler*innen verwendet. Der Umgang mit Verletzung der Testfairness durch DIF-Effekte in einzelnen Subpopulationen einer Stichprobe wird in der Literatur kontrovers diskutiert. In der Forschungslandschaft gibt es keine allgemein anerkannten Grenzwerte für die Tolerierbarkeit von DIF-Effekten. Vielmehr existieren unterschiedliche Klassifizierungsschemata, die DIF-Effekte zu kategorisieren versuchen. In diesem Artikel wird sich auf das Klassifizierungsschema von Paek und Wilson (2011) bezogen, das DIFs nach der Stärke und ihrer statistischen Signifikanz bewertet. Dabei stellen DIF-Effekte der Kategorie A ($|DIF| < 0.426$ und $p > 0.05$) keine problematischen Items dar, während der Einsatz von Items der Kategorie B ($0.426 < |DIF| < 0.628$; $p < 0.05$) und Kategorie C ($|DIF| > 0.628$; $p < 0.05$) als begründenswert zu bewerten ist. Darüber hinaus ist zu bedenken, dass nicht jeder DIF-Effekt eine Verletzung der Testfairness darstellen muss, so kann bspw. ein DIF-Effekt bezogen auf das Vorwissen von Schüler*innen zu einer Thematik wünschenswert sein, da das Testinstrument somit instruktionssensitiv reagiert. Daher ist eine notwendige Elimination von Items aufgrund von DIF-Effekten nicht zwingend ratsam und sollte auch aus der Perspektive der Abwägung von Konstruktvalidität entschieden werden. Auf Grundlage der Personen- und Itemparameter sowie der Informationen zur Itemhomogenität wurden auffällige Items vom Text unter Berücksichtigung von Überlegungen zur Inhaltsvalidität und Reliabilität des Testinstruments exkludiert oder überarbeitet.

4.3.4 Vergleichende Ergebnisse von Feld- und Haupttest

Befunde zu den Itemschwierigkeiten im Feldtest

Ein Testinstrument gilt als ausreichend skaliert, wenn die Parameter des Modells und die dazugehörigen Items (1) ausreichend weit auf der Logit-Skala streuen und (2) eine möglichst hohe Varianz der Logit-Werte aufweisen. Die durch die MCMLM-Methode geschätzten Personen- und Itemparameter wurden auf eine gemeinsame Logit-Skala transformiert (Personen-Item-Map). Die Personen-Item-Map (Wright Map) zeigt die Anzahl der Fälle des Personenparameters und die Itemparameter (siehe Abbildung 4.3.2).

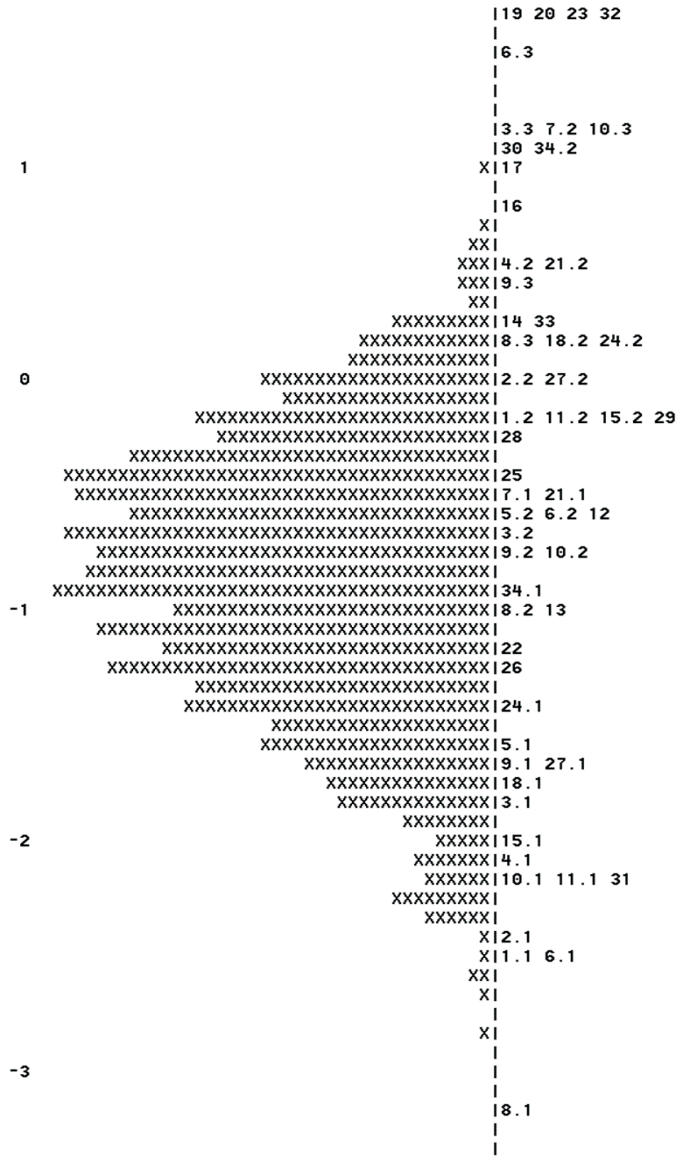


Abbildung 4.3.2: Personen-Item-Map für die Feldteststichprobe des Testinstruments TBA-EL

Die Itemschwierigkeitsparameter im Feldtest streuen im Bereich von -2.111 und 6.746 mit einer Spannweite von 8.857 (siehe Abbildung 4.3.2). Der Mittelwert der Itemschwierigkeitsparameter ist aufgrund der Modellspezifikation auf 0 fixiert ($M = 0$; $SD = 1.672$). Der Kolmogoroff-Sminorv-Test zeigt zunächst keine Normalverteilung der Itemparameter an ($K\text{-}S = 0.216$; $df = 34$, $p < 0.001$). Bei Exkludierung des Ausreißeritems 7_5 mit der Itemschwierigkeit von 6.746 Logits kann jedoch

die Normalverteilung angenommen werden (K-S = 0.178; df = 33; p = 0.10). Aus der Wright Map kann geschlossen werden, dass 22 von 35 Items eine negative Itemschwierigkeit aufweisen und somit als „eher einfach" zu werten sind. Die 12 Items im positiven Logit-Bereich sind hingegen als „eher schwierig" zu klassifizieren. Bei der Betrachtung der Itemschwierigkeiten nach Aufgabentyp zeigte sich mittels ANOVA, dass keine signifikanten Gruppenunterschiede hinsichtlich der Aufgabentypen Single-Choice, Multiple-Choice und eines offenen Antwortformats bestehen (F(3.164) = 31.64; p = 0.056). Ebenfalls wurden keine signifikanten Gruppenunterschiede in der Itemschwierigkeit hinsichtlich der Inhaltsbereiche des Domänenmodells gefunden (F(0.195) = 1.95; p = 0.824).

Befunde zu den Itemschwierigkeiten in der Hauptstudie

Die Itemschwierigkeitsparameter im Haupttest streuen im Bereich von -2.012 und 3.042 mit einer Spannweite von 5.054 (siehe Abbildung 4.3.3). Der Mittelwert der Itemschwierigkeitsparameter ist ebenfalls auf 0 fixiert (M = 0; SD = 1.00). Der Kolmogoroff-Sminorv-Test zeigt eine Normalverteilung der Itemparameter an (K-S = 0.126; df = 32, p = 0.200). Aus der Wright Map kann geschlossen werden, dass 17 von 32 Items eine negative Itemschwierigkeit aufweisen und somit als „eher einfach" zu werten sind. Die 15 Items im positiven Logit-Bereich

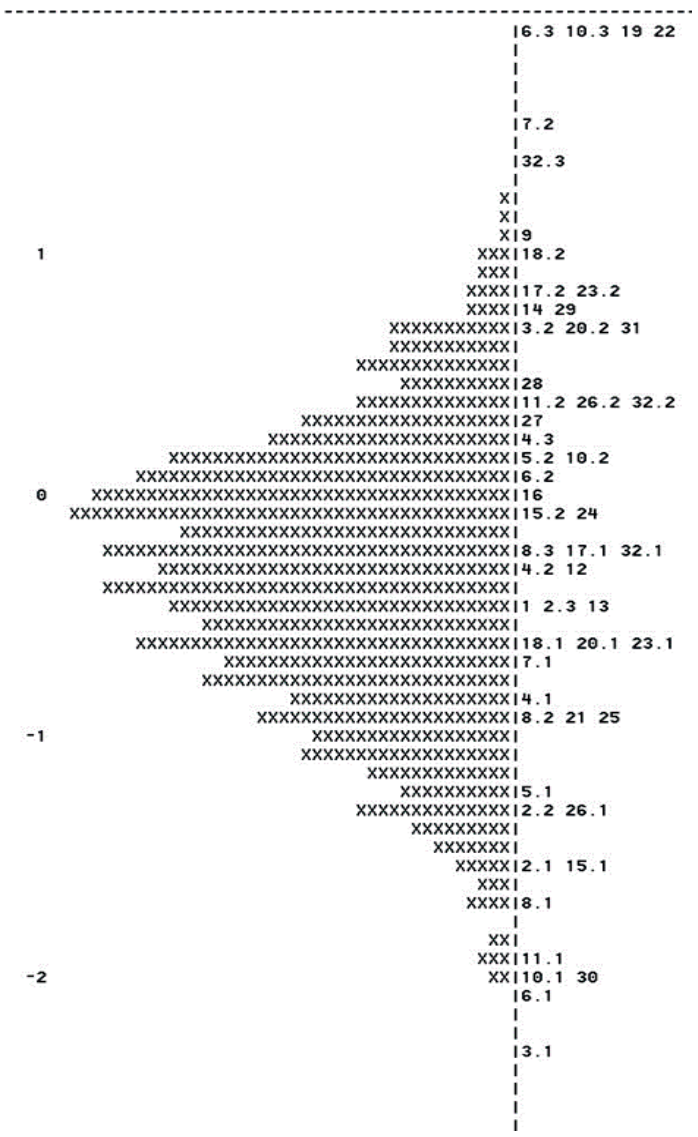


Abbildung 4.3.3: Personen-Item-Map für die Haupttestst ichprobe des Testinstruments TBA-EL

sind hingegen als „schwieriger“ zu klassifizieren. Bei der Betrachtung der Itemschwierigkeiten nach Aufgabentyp zeigte sich mittels ANOVA, dass keine signifikanten Gruppenunterschiede hinsichtlich der Aufgabentypen Single-Choice, Multiple-Choice und eines offenen Antwortformats bestehen ($F(0.102) = 1.478$; $p = 0.245$). Ebenfalls wurden keine signifikanten Gruppenunterschiede in der Itemschwierigkeit hinsichtlich der Inhaltsbereiche des Domänenmodells gefunden ($F(0.003) = 1.95$; $p = 0.949$). Es kann geschlossen werden, dass bei der Lösung der Items mögliche Effekte durch die Komplexität des Aufgabenformats zu vernachlässigen sind. Darüber hinaus kann gezeigt werden, dass für die Inhaltsbereiche des Domänenmodells eine ausgewogene Verteilung hinsichtlich der inhaltlichen Schwierigkeit der Aufgaben gegeben ist.

Befunde zu den Personenfähigkeitswerten im Feldtest

Die Personenparameter (WLE) in der Feldteststichprobe streuen zwischen den Werten -5.323 und 0.936 mit einer Spannweite von 4.167 Logits. Der Mittelwert des Personenfähigkeitsparameters beträgt -0.845 mit einer Standardabweichung von 0.700 Logits. Die Verteilung der Personenparameter ist somit linksschief und nicht normalverteilt (K-S = 0.072; df = 753; $p < 0.001$). Die Nichtnormalverteilung des Personenfähigkeitsparameters trotz Ausschluss von 63 Fällen, die mehr als 75 % der Aufgaben nicht beantwortet haben, könnte als Indiz dafür interpretiert werden, dass der Test für die Feldteststichprobe als etwas zu schwer konzipiert wurde. Schiefe (-0.360) und Kurtosis (-0.092) weisen keine exzessiven Werte auf, sodass nur eine leichte Verletzung der Normalverteilungsannahme angenommen werden kann. Für den Feldtest kann ein sehr hoher, monotoner, nicht linearer, s-förmiger Zusammenhang zwischen den Personen-Testwerten und den Personenparametern festgestellt werden ($r = 0.955$; $p < 0.001$).

Im Testinstrument nehmen die Standardfehler der Personenparameter geringe Werte an ($M = 0.309$; $SD = 0.0265$) und liegen im Logit-Bereich von 0.292 bis 0.520. Die Schätzung der Personenparameter ist für den Logit-Bereich zwischen -0.502 und 0.593 signifikant ($p < 0.05$). Die Standardfehler sind für den Bereich zwischen -2.00 Logits bis 0.935 Logits am geringsten (siehe Abbildung 4.3.4). Die Personenfähigkeiten < 2 Logits werden weniger zuverlässig geschätzt. Dies deckt sich auch mit der grafischen Beurteilung der Testinformationsfunktion. Die maximale Testinformation kann im Logit-Bereich zwischen -2 bis 1 beobachtet werden (siehe Abbildung 4.3.5).

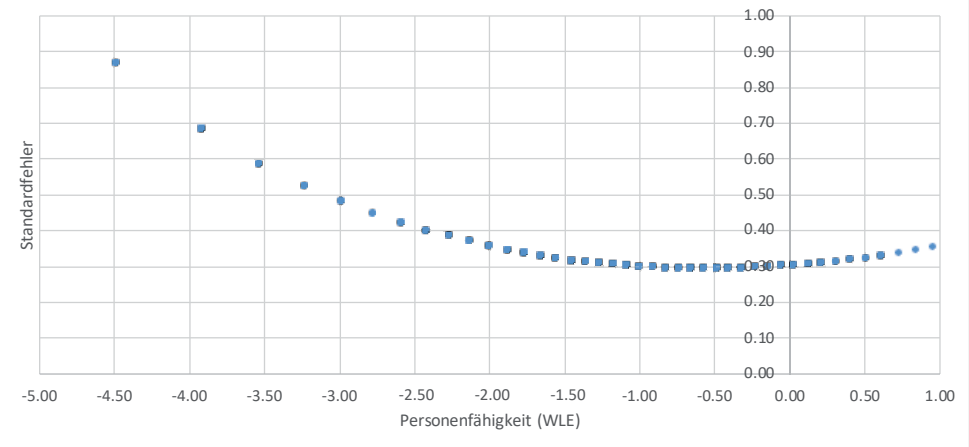


Abbildung 4.3.4: Personenfähigkeiten und Standardfehler für die Feldteststichprobe des Testinstruments TBA-EL

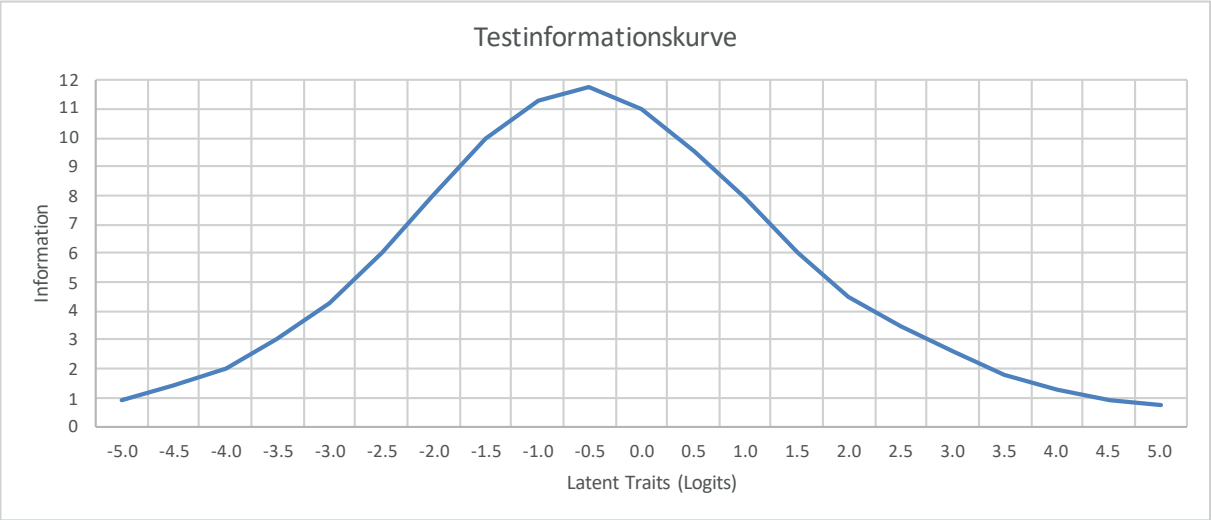


Abbildung 4.3.5: Testinformationsfunktion für die Feldteststichprobe des Testinstruments TBA-EL

Befunde zu den Personenfähigkeitswerten in der Hauptstudie

Die Personenparameter (WLE) im Haupttest streuen zwischen den Werten -3.900 und 1.599 mit einer Spannweite von 5.499 Logits. Der Mittelwert des Personenfähigkeitsparameters beträgt -0.392 mit einer Standardabweichung von 0.670 Logits. Die Verteilung der Personenparameter ist ebenfalls linksschief und nicht normalverteilt (K-S = 0.050; df = 2852; $p < 0.001$). Im Vergleich zum Feldtest hat sich der Mittelwert der Personenfähigkeit erhöht. Schiefe (-0.0416) und Kurtosis (-0.823) weisen keine exzessiven Werte auf, sodass nur eine leichte Verletzung der Normalverteilungsannahme angenommen werden kann. Für den Feldtest kann ein sehr hoher, monotoner, nicht linearer, s-förmiger Zusammenhang zwischen den Personen-Testwerten und den Personenparametern festgestellt werden (r = 0.998; $p < 0.001$).

Im Testinstrument der Hauptstudie nehmen die Standardfehler der Personenparameter ebenfalls geringe Werte an (M = 0.298; SD = 0.03) und liegen im Logit-Bereich von 0.282 bis 0.854. Die Schätzung der Personenparameter ist für den Logit-Bereich zwischen -0.582 und 0.564 signifikant ($p < 0.05$). Die Standardfehler sind für den Bereich zwischen -2.00 Logits bis 1.500 Logits am geringsten (siehe Abbildung 4.3.6). Die Personenfähigkeiten kleiner 2 Logits werden weniger

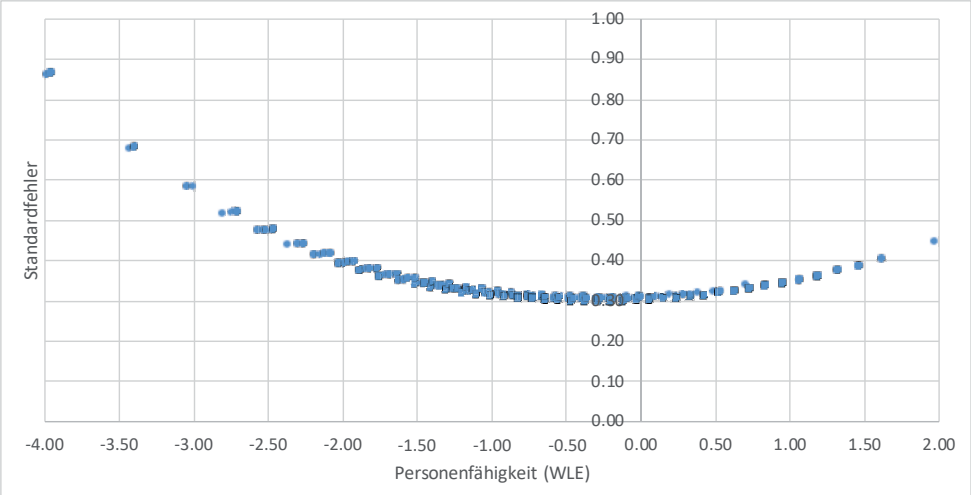


Abbildung 4.3.6: Personenfähigkeiten und Standardfehler für die Haupttestst ichprobe des Testinstruments TBA-EL

zuverlässig geschätzt. Diese deckt sich auch mit der grafischen Beurteilung der Testinformationsfunktion. Die maximale Testinformation kann im Logit-Bereich zwischen -2 bis 1.5 beobachtet werden, die auch in ihrem Maximum einen höheren Wert aufweist als die des Feldtests (siehe Abbildung 4.3.7).

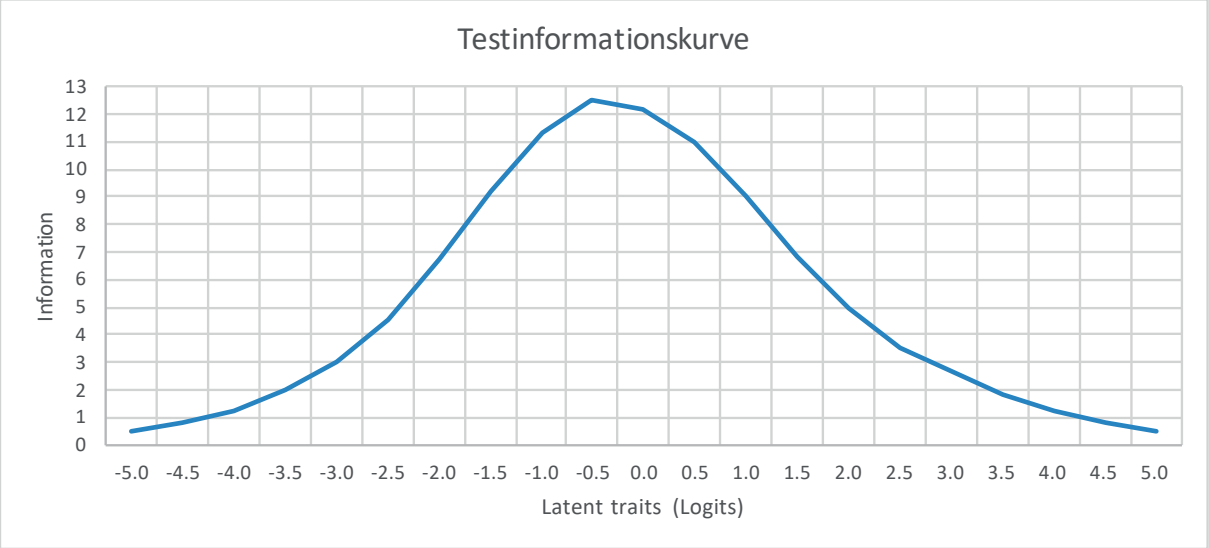


Abbildung 4.3.7: Testinformationsfunktion für die Haupttestst ichprobe des Testinstruments TBA-EL

Reliabilität der Testinstrumente: Vergleich zwischen Feld- und Hauptstudie

Die Messgenauigkeit des Testinstruments für die globalen Personenparameter kann über den Reliabilitätskoeffizient der IRT (EAP/PV) und der KTT (Cronbachs Alpha) beurteilt werden. Die Ergebnisse der globalen Reliabilitätsparameter konstatieren für alle Kennzahlen Werte deutlich über 0.70, was für eine hohe Reliabilität des Testinstruments spricht (siehe Tabelle 4.3.6). Die Höhe des WLE-Koeffizienten (0.879) lässt darauf schließen, dass das Testinstrument zuverlässig zwischen fähigen und weniger fähigen Personen unterscheiden kann.

Tabelle 4.3.6: Probabilistische und klassische Reliabilitätskennwerte

			Feldtest	Hauptstudie
Probabilistische Kennwerte	Personen-bezogen	EAP/PV	0.813	0.838
		MLE	0.804	0.809
		WLE	0.811	0.801
	Itembezogen	Item-Separation-Reliabilität	0.975	0.998
Klassische Testtheorie		Cronbachs Alpha	0.800	0.830

Zusammenfassend kann somit festgestellt werden, dass es eine ausreichende empirische Evidenz hinsichtlich der Messgenauigkeit des Instruments TBA-EL gibt. Die Personenfähigkeit der Teilnehmenden der Hauptstudie ist im Mittel um 0.500 Logits höher als die des Feldtests.

Die Annahme eines Partial-Credit-Rasch-Modells ist erfüllt, wenn (1) die Item-Infit-Werte (wMNSQ) idealerweise dem Erwartungswert 1 entsprechen sowie signifikant sind ($|t| < 1.96$ bzw. $p < 0.05$) und (2) die Schwellenparameter der Antwortkategorien aufsteigend angeordnet sind. Für den Cut-Off-Wert für den Item-Infit schlagen Adams und Khoo (1996) einen Wertebereich zwischen 0.75 und 1.33 vor, während in Large-scale Assessments wie PISA wMNSQ-Werte zwischen 0.85 und 1.15 als angemessen betrachtet werden (Kastberg et al., 2021).

Befunde zum Itemfit: Vergleich zwischen Feld- und Hauptstudie

Hinsichtlich der Feldteststichprobe erfüllen 33 von 34 Items des Testinstruments den strengen wMNSQ-Wertebereich, lediglich Item 7_5 weist einen Underfit auf (siehe Tabelle 4.3.7). Die T-Werte streuen in einem Bereich zwischen -4.40 und 5.50. Drei Items weisen einen T-Wert kleiner -1.96 auf, was für eine zu hohe Trennschärfe spricht und als eher nicht problematisch betrachtet wird. Bei fünf Items kann ein T-Wert größer 1.96 festgestellt werden, was auf eine signifikante Abweichung und eine niedrige Trennschärfe schließen lässt. Für eine präzisere Beurteilung des Item-Infits wird auch die Trennschärfe der klassischen Testtheorie berücksichtigt, die nicht unter dem Wert von 0.20 liegen sollte.

Tabelle 4.3.7: Vergleich der Itemparameter – WMNSQ der Testitems des Testinstruments TBA-EL im Feld- und Haupttest

Items	Nr.	Iteminhalt	WMNSQ-Feldtest	WMNSQ-Hauptstudie
1_1	1	Preisberechnung Einkaufszettel, Grundrechenarten	-	1.05
1_2	2	Bedürfnisse & Bedarf	0.98	1.04
1_3	3	Wirtsch. Unterschied Bio-Produkte/konv. Produkte	0.95	0.96
1_4	4	Knappheitskonzept	0.95	1.00
2_1	5	Influencer-Marketing	0.98	-
2_2	6	Wirkung von digitalen Marketingstrategien	0.84	0.95
2_3	7	Nutzen digitaler Marketingstrategien aus Unt.-Sicht	1.13	1.08
3_1	8	Definition Nachhaltigkeit	0.88	0.95
3_2	9	Facetten von Nachhaltigkeit	0.92	0.99
3_3	10	Fair-Trade-Produkte	1.07	1.00
3_4	11	Fair-Trade-Konzept	0.99	0.93
3_5	12	Informationsquellen zu Produktinformationen	0.94	0.93
4_1	13	Berechnung Jahreszinsen	1.03	1.01
4_2	14	Konzept Zinseszins	1.11	1.04
4_3	15	Berechnung unterjährige Zinsen	1.06	1.01
4_4	16	Zölle & Auswirkungen auf Unternehmen	0.98	0.98
4_5	17	Gewinnkonzept	1.06	0.93
4_6	18	Kaufkraft	1.07	1.24
5_1	19	Zentrale Begriffe Kaufvertrag	1.11	1.10
5_2	20	Prozentrechnen In-App-Kauf (verminderter GW)	0.99	1.00
5_3	21	Prozentrechnen In-App-Kauf (Preisdifferenz)	1.02	-
5_4	22	Gefahren des In-App-Kaufs	0.92	0.96
6_1	23	Subtraktion Preisunterschied	0.90	0.99
6_2	24	Prozentrechnen Angebotsvergleich (vermehrter GW)	0.98	0.99
6_3	25	Ursachen Kostenvorteile für Online-Shopping	1.04	1.00
6_4	26	Wirkung von AGBs bei Kaufverträgen	1.13	1.12
6_5	27	Währungsumrechnung	0.98	0.97
6_6	28	Institutionen des Verbraucherschutzes	0.99	1.00
7_1	29	Bezahlen mit Kleingeld	0.98	0.99
7_2	30	Funktionen des Geldes	1.08	1.09
7_3	31	Kaufvertrag rechtswirksam?	0.97	1.02
7_4	32	Bezahlen mit EC-Karte	0.95	0.97
7_5	33	Kaufvertrag Botengang	2.17	-
8_1	34	Preisbildung	1.08	1.07
8_2	35	Wirtschaftskreislauf	0.99	0.97

In Bezug auf den Haupttest kann festgestellt werden, dass sich die wMNSQ-Werte von 22 Items leicht verbessert haben, während 8 Items leicht niedrigere Werte aufweisen. Item 4_6 liegt mit einem Wert von 1.24 deutlich über dem Grenzwert von 1.15. Zwar konnte dies im Feldtest nicht beobachtet werden (wMNSQ-Wert von 1.07), jedoch traten bei diesem Item technische Probleme dergestalt auf, dass Schülerantworten nicht vollständig erfasst wurden, sodass eine vorbehaltlose Zuverlässigkeit der Schätzung des Itemfit-Werts für den Feldtest nur eingeschränkt möglich war. Drei von sechs Items mit einem T-Wert größer 1.96 haben Trennschärfen von größer 0.20. Drei Items weisen eine Trennschärfe von kleiner 0.20 auf. Für eine Revision der Testitems lässt sich daher schließen, dass eine Prüfung der Distraktoren in besagten Items notwendig sein kann, um eine sprachlich oder inhaltlich bessere Abgrenzung der Antwortoptionen zu gewährleisten. Hinsichtlich der Schwellenparameter weisen keine der 17 polytom codierten Items ungeordnete Schwellenparameter auf.

Für den Haupttest kann konstatiert werden, dass 8 Items einen T-Wert größer 1.96 aufweisen, fünf Items zeigten dies bereits im Feldtest. 1_1 und 4_6 unterlagen im Feldtest technischen Problemen, sodass eine inhaltliche Anpassung oder eine Designänderung vorab nicht möglich war. Item 8_1 zeigte zuvor keinerlei Auffälligkeiten. Hinsichtlich der Trennschärfe konnten bei 4 von 8 Items Trennschärfen kleiner 0.20 festgestellt werden.

Für eine bessere Veranschaulichung des Itemrevisionsprozesses wurde anhand eines exemplarischen Items aufgezeigt, inwiefern Veränderungen zwischen dem Feld- und Haupttest umgesetzt wurden (siehe Abbildung 4.3.8)

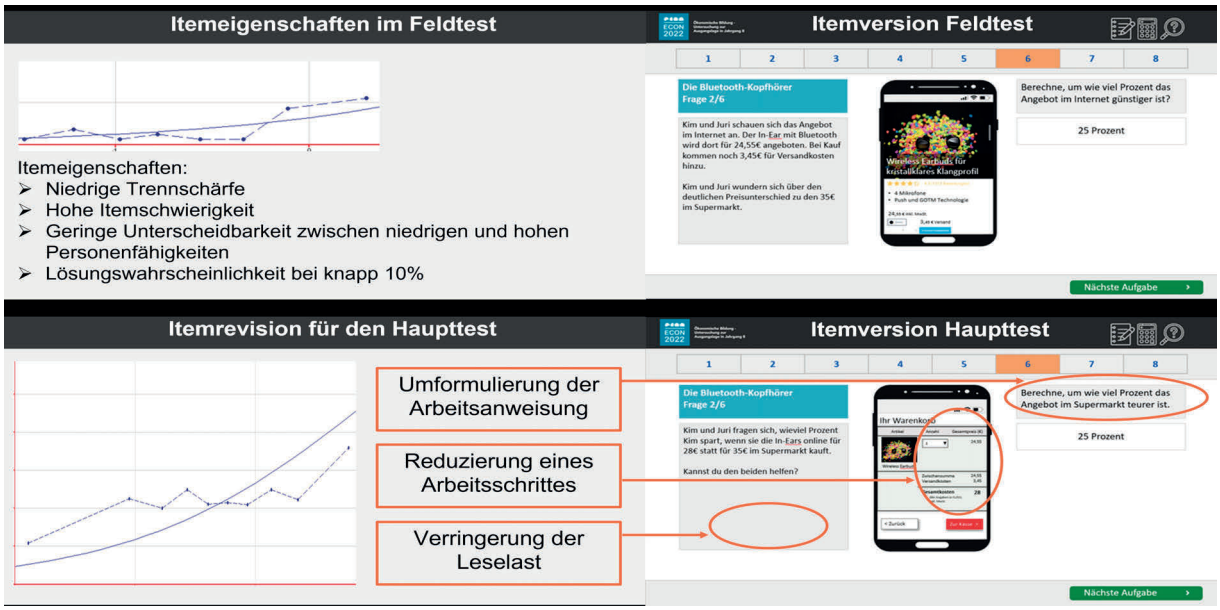


Abbildung 4.3.8: Itemrevision zwischen Feld- und Haupttest am Beispiel des Items 6_2 – Vermehrter Grundwert

Analyse des Informationsgehalts der Items

Mithilfe der Item-Characteristic-Curves (ICC) können die Items auch grafisch analysiert werden. Idealerweise sollten die Kurven s-förmig verlaufen, da so gewährleistet ist, dass ein Item zuverlässig zwischen fähigen und weniger fähigen Testpersonen unterscheiden kann und der Informationsgehalt eines Items in Bezug auf die unterstellte Kompetenz am höchsten ist (Moosbrugger & Kelava, 2012).

Bei den Items (4_6, 6_4, 7_2 und 8_1) wurden analog zur Prüfung des Feldtests die ICCs näher betrachtet. Item FT46 zeigt für ein Partial-Credit-Item einen zu flachen Anstieg der richtigen Antwortmöglichkeiten, dies spricht für ein tendenziell zu schwieriges Item für die Zielgruppe. Item 6_4 stellt sich auch nach Veränderung der Distraktoren im Feldtest ebenfalls als problematisch im Haupttest heraus. Die ICC erweist sich als nicht monoton ansteigend und die Trennschärfe ist mit 0.08 gering. Die Items 7_2 und 8_1 zeigen bei der grafischen Analyse zufriedenstellend ansteigende ICCs.

Testfairness des Testinstruments: Vergleich zwischen Feld- und Hauptstudie

Die Befunde bezogen auf die Analyse der Testfairness untersuchen, ob das Testinstrument gruppenspezifisch hinsichtlich personenbezogener Merkmale diskriminiert. Ziel sollte es sein, DIF-Effekte bei der Konstruktion eines Testinstruments möglichst gering zu halten. Leistungsunterschiede sollten möglichst durch die unterschiedlichen Personenfähigkeiten der Proband*innen erklärt werden können und nicht durch die Zugehörigkeit zu einer spezifischen Subpopulation.

Die Zuwanderungsgeschichte (ZWG) ist nominal skaliert und schließt Personen ein, bei denen mindestens ein Elternteil oder die Testperson selbst im Ausland geboren ist. Das Merkmal der familiären Verwendung der Landessprache definiert, ob Deutsch im häuslichen Umfeld die häufigste gesprochene Sprache darstellt oder nicht. Unter Geschlecht werden die Merkmale männlich, weiblich und divers verstanden, wobei das Merkmal divers von nur 42 Testteilnehmenden in der Hauptstudie gewählt wurde, sodass diese bei den DIF-Analysen aufgrund der geringen Gruppengröße nicht berücksichtigt werden kann.

Testfairness auf Testebene

Auf Testebene zeigt die DIF-Analyse bei der Feldteststichprobe keine signifikanten Geschlechts- und Sprachunterschiede (siehe Tabelle 4.3.8). Bei der Zuwanderungsgeschichte konnte ein signifikanter DIF-Effekt festgestellt werden. Dieser ist auf Testebene mit 0.282 sowie mit 0.429 gemäß der Klassifizierung nach Paek & Wilson (2011) als gering einzuschätzen. Im Vergleich zum Feldtest zeigt sich bei der Stichprobe der Hauptstudie ein ebenfalls signifikanter DIF-Effekt bei der häuslichen Verwendung der Landessprache. Dieser ist jedoch mit 0.444 ebenfalls als gering einzuschätzen.

Tabelle 4.3.8: DIF-Unterschiede der Subgruppen auf Testebene im Feld- und Haupttest

Testzeitpunkt	Subgruppen	Z	DIF-Wert	Standard-fehler	Chi-Quadrat	p-Wert
Feldtest	Keine ZWG	1	0.142	0.053	7.27(1)	0.007
	mit ZWG	2	-0.142	0.053		
Haupttest	kein ZWG	1	0.178	0.014	1748.48(1)	>0.001
	mit ZWG	2	-0.178	0.014		
Feldtest	männlich	1	0.009	0.043	0.04(1)	0.8415
	weiblich	2	-0.009	0.043		
Haupttest	männlich	1	0.007	0.013	0.28(1)	0.5967
	weiblich	2	-0.007	0.013		
Feldtest	Keine häusliche Verwendung der Landessprache	1	-0.088	0.051	2.95(1)	0.0859
	Häusliche Verwendung der Landessprache	2	-0.088	0.051		
Haupttest	Keine häusliche Verwendung der Landessprache	1	-0.222	0.016	202.15(1)	>0.001
	Häusliche Verwendung der Landessprache	2	0.222	0.016		

Testfairness auf Itemebene: Feldtest

Auf Itemebene wurde bei der Feldteststichprobe für den Interaktionsterm Item Geschlecht kein signifikanter DIF-Unterschied festgestellt (Chi-Square (df) = 32.18 (32), p = 0.458), sodass davon auszugehen ist, dass das Testinstrument in Hinblick auf das Geschlecht nicht diskriminiert. Für die Merkmale Zuwanderungsgeschichte (Chi-Square (df) = 55.20 (32), p = 0.007) und familiäre Verwendung der Landessprache (Chi-Square (df) = 79.18 (32), p < 0.001) konnten hingegen signifikante DIF-Unterschiede festgestellt werden. Zur Berechnung des DIF-Effekts wurde der DIF-Schätzer pro Item verdoppelt. In der Gesamtschau konnten 8 unterschiedliche Items identifiziert werden, die einen signifikanten DIF-Effekt über 0.426 aufweisen. Davon haben 3 Items bei mehr als einem personenbezogenen Merkmal einen DIF-Effekt (siehe Tabelle 4.3.9). Hinsichtlich des Merkmals „Zuwanderungsgeschichte" zeigen vier Items einen DIF-Effekt der Kategorie B und zwei Items der Kategorie C. Bezogen auf das Merkmal der familiären Verwendung der Landessprache kann bei drei Items ein DIF-Effekt der Kategorie B und bei zwei Items der Kategorie C festgestellt werden.

Tabelle 4.3.9: DIF-Effekte der Subgruppen auf Itemebene im Feldtest

Items	Nr.	Iteminhalt	Absolute Itemschwierigkeit der Stichprobe	Zuwanderungsgeschichte	Familiäre Verwendung der Landessprache
3_1	8	Definition Nachhaltigkeit	0.323		B+*
4_5	17	Gewinnkonzept	0.760	B+*	B-**
5_2	20	Prozentrechnen In-App-Kauf (verminderter GW)	2.561	B+*	
5_3	21	Prozentrechnen In-App-Kauf (Preisdifferenz)	3.145	B-*	
6_2	24	Prozentrechnen Angebotsvergleich (vermehrter GW)	2.675	C+**	
6_4	26	Wirkung von AGBs bei Kaufverträgen	-0.149	B+**	C-**
7_3	31	Kaufvertrag rechtswirksam?	1.017	C-*	B+**
8_1	33	Preisbildung	0.303		C-**

B = DIF-Effekt > 0.426; C = DIF-Effekt > 0.538; * p < 0.05; ** p < 0.001

Testfairness auf Itemebene: Haupttest

Bei der Hauptstudie wiesen auf Itemebene bei Betrachtung beider Merkmale nur drei Items signifikante DIF-Effekte auf (siehe Tabelle 4.3.10). FT64 erwies sich sowohl beim Merkmal der familiären Verwendung der Landessprache als auch bei der Zuwanderungsgeschichte insofern als problematisch, da dieses nicht deutschsprechende Personen sowie Menschen mit Zuwanderungsgeschichte trotz der Veränderungen im Feldtest diskriminiert. Bei sechs Items, die im Feldtest noch signifikante DIF-Effekte aufwiesen, konnte dies in der Hauptstudie nicht mehr festgestellt werden.

Tabelle 4.3.10: DIF-Effekte der Subgruppen auf Itemebene im Haupttest

Items	Nr.	Iteminhalt	Absolute Itemschwierigkeit der Stichprobe	Zuwanderungsgeschichte	Familiäre Verwendung der Landessprache
5_2	20	Prozentrechnen In-App-Kauf (verminderter GW)	3.230	B+*	
6_4	26	Wirkung von AGBs bei Kaufverträgen	-0.093	B-**	B+**
7_4	32	Bezahlen mit EC-Karte	-2.029		B+**

B = DIF-Effekt > 0.426; C = DIF-Effekt > 0.538; * p < 0.05; ** p < 0.001

Arten der Itemmodifikationen

Hinsichtlich itemspezifischer Veränderungen zwischen Feld- und Haupttest kann zwischen vier Änderungsverfahren differenziert werden:

1) *Sprachliche Präzisierung oder Vereinfachung*: Eine sprachliche Anpassung erfolgte bei 7 Items. Hier wurden zumeist überflüssige Fachtermini entfernt oder der Itembegleittext gekürzt, um die Leselast zu verringern.
2) *Veränderung bei Itemdistraktoren*: Eine Veränderung der Distraktoren betraf 3 Items. Hier wurden diese entweder umformuliert, durch fachlich eindeutigere ersetzt oder die Anzahl verringert.
3) *Formatänderung*: Eine Änderung des Itemformats wurde bei zwei Items vorgenommen. Item 7_3 wurde aufgrund der hohen Schwierigkeit von einem Item mit offenem Format zu einem Single-Choice-Item. Item 2_2 erfuhr eine Designänderung durch das Hinzufügen grafischer Elemente, die sich zuvor in Item 2_1 befanden.
4) *Itemexkludierung*: Eine Itemexkludierung wurde bei drei Items vorgenommen. Item 2_1 wurde mit Item 2_2 kombiniert. Item 5_3 wurde ersatzlos gestrichen, da dieses für die interne Validität des Konstrukts als nicht notwendig erachtet wurde und sich als zu schwierig erwiesen hat, dasselbe gilt für Item 7_5.

Überblick der Veränderungen am Testinstrument zwischen Feld- und Hauptstudie

In der Gesamtschau der Testrevision kann konstatiert werden, dass die Änderungen, die im Testinstrument des Feldtests vorgenommen wurden, sich überwiegend positiv auf die Befunde der Hauptstudie ausgewirkt haben. Die Itemfit-Werte haben sich überwiegend leicht verbessert, während negative Änderungen marginal sind. Sowohl Feld- als auch Haupttest decken in ihrer Itemverteilung das Personenfähigkeitsspektrum ausreichend ab. Die Linksschiefe konnte im Haupttest verringert werden, sodass von besserer Adäquanz des Schwierigkeitsniveaus des Testinstruments ausgegangen werden kann. Die Veränderung von Schwierigkeiten einzelner Items zwischen Feld- und Haupttest kann ursächlich auf die unterschiedliche Itemanzahl und die Exkludierung von zwei besonders schweren Items (5_3 und 7_5) zurückgeführt werden.

Literatur

Adams, R. J. & Khoo, S. T. (1996). *ACER Quest. interactive test analysis system. Version 2.1*. The Australian Council for Educational Research.

Adams, R. J., Wilson, M. & Wang, W.-C. (1997). The multidimensional random coefficient multinomial logit model. *Applied psychological measurement*, *21*(1), 1–23. https://doi.org/10.1177/0146621697211001.

Adams, R. J., Wu, M. L. & Wilson, M. (2018). ACER ConQuest. In W. J. van der Linden (Hrsg.), *Handbook of item response theory: Three volume set* (S. 495–505). CRC Press.

AERA, APA & NCME. (2014). *Standards for educational and psychological testing*. American educational research association.

Brennan, R. L. & Prediger, D. J. (1981). Coefficient Kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, *41*(3), 687–699. https://doi.org/10.1177/001316448104100307.

Frey, A. (2012). Adaptives Testen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 275–293). Springer.

Hayes, A. F. & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77–89. https://doi.org/10.1080/19312450709336664.

IBM Corp. (2021). *IBM SPSS Statistics for Windows (Version 28)* [Computer software]. IBM Corp.

Kastberg, D., Cummings, L., Ferraro, D. & Perkins, R. C. (2021). *Technical report and user guide for the 2018 Program for International Student Assessment (PISA)*. (NCES 2021-011). U.S.

Kirsch, A. (2021). *Professionalitätsentwicklung angehender Lehrkräfte*. Springer. https://doi.org/10.1007/978-3-658-36123-5.

Krippendorff, K. (2004). Reliability in content analysis. Some common misconceptions and recommendations. *Human Communication Research*, *30*(3), 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x.

Linacre, J. (1994). Sample size and item calibration stability. *Rasch Mes Trans*, *7*, 328. https://ci.nii.ac.jp/naid/10031091465/.

Moosbrugger, H. & Kelava, A. (2012). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 7–26). Springer.

Paek, I. (2002). *Investigations of differential item functioning: Comparisons among approaches, and extension to a* multidimensional context*. University of California. https://search.proquest.com/openview/a92992ec1fbc1ea0894b9e8e3842fabd/1?pq-origsite=gscholar&cbl=18750&diss=y.

Paek, I. & Wilson, M. (2011). Formulating the Rasch differential item functioning model under the marginal maximum likelihood estimation context and its comparison with Mantel–Haenszel Procedure in short test and small sample conditions. *Educational and Psychological Measurement*, *71*(6), 1023–1046. https://doi.org/10.1177/0013164411400734.

Rost, J. (2004). Psychometrische Modelle zur Überprüfung von Bildungsstandards anhand von Kompetenzmodellen. *Zeitschrift für Pädagogik*, *50*(5), 662–678. https://doi.org/10.25656/01:4834.

Teresi, J. A., Ramirez, M., Lai, J.-S. & Silver, S. (2008). Occurrences and sources of Differential Item Functioning (DIF) in patient-reported outcome measures: Description of DIF methods, and review of measures of depression, quality of life and general health. *Psychology science quarterly*, *50*(4), 538.

Winther, E. (2010). *Kompetenzmessung in der beruflichen Bildung*. W. Bertelsmann Verlag.

Ziesemer, F., Peyer, M., Klemm, A. & Balderjahn, I. (2016). Die Messung von nachhaltigem Konsumbewusstsein. *Ökologisches Wirtschaften – Fachzeitschrift*, (4), 24–26. https://doi.org/10.14512/OEW310424

Zhao, X., Liu, J. S. & Deng, K. (2013). Assumptions behind intercoder reliability indices. *Annales of the International Communication Association*, *36*(1), 419–480. https://doi.org/10.1080/23808985.2013.11679142.

4.4 Validierung des Fragebogeninstruments anhand der Feldtestdaten

Fenna Henicz, Nina Johanna Welsandt, Fabio Fortunati, Hermann Josef Abs & Esther Winther

Prüfung der Dimensionalität und der internen Konsistenz als Gütemaße für die Skalen

Das folgende Unterkapitel gibt einen Überblick zur Qualität der in Kap. 3.5 dargestellten Skalen mit Blick auf die ökonomischen Einstellungen und Bildungs- und Sozialisationskontexte. Zur Prüfung der Konstruktvalidität wurde die Dimensionalität anhand von Faktorenanalysen untersucht, sodass Aussagen darüber getroffen werden können, ob die Skalen das intendierte Konstrukt auch tatsächlich messen (Moosbrugger & Kelava, 2012). Die interne Konsistenz der jeweiligen Skalen beschreibt, inwieweit die enthaltenen Items zusammenhängen und für die Messung des Konstruktes geeignet sind. Sie stellt einen Aspekt der Reliabilität und somit eine Voraussetzung für die Validität dar (Moosbrugger & Kelava, 2012). Items, deren Werte in den Faktoren- und/oder Itemanalysen auf eine unzureichende Qualität hindeuteten, wurden modifiziert oder aus der Skala entfernt, um die Messeigenschaften des Instruments bestmöglich nach den Gütekriterien für quantitative Forschung auszurichten.

Cronbachs Alpha als Maß der internen Konsistenz

Zur Prüfung der internen Konsistenz der Skalen wurde die Korrelation unter den einzelnen Items geprüft. Je höher die Items dabei untereinander positiv korrelieren, desto stärker repräsentieren sie das jeweilige Konstrukt als dessen Indikatoren (Döring & Bortz, 2016; Moosbrugger & Kelava, 2012; Schermelleh-Engel & Werner, 2012). Ein hoher Cronbachs-Alpha-Koeffizient entspricht dabei einer hohen internen Konsistenz, die für die Eignung der Items zur Messung des jeweiligen Konstruktes spricht. Hier ist zu berücksichtigen, dass der α-Wert mit der Anzahl der Items steigt und so eine Skala mit einer hohen Anzahl an Items nicht unbedingt eine höhere interne Konsistenz aufweisen muss als Skalen mit weniger Items (Cortina, 1993).

Trennschärfe als Differenzierungsmaß zwischen Personen mit unterschiedlichen Merkmalsausprägungen

Schließlich sollten die Items außerdem Personen mit hoher und niedriger Merkmalsausprägung unterscheiden können, was sich über die Berechnung der Trennschärfe der Items prüfen lässt. Dabei wird die bivariate Korrelation des Items mit dem Gesamtscore der Skala betrachtet. Ein hoher Wert (zwischen -1 und 1) repräsentiert eine hohe Trennschärfe und differenziert entsprechend gut zwischen Schüler*innen mit hohen und Schüler*innen mit niedrigen Merkmalsausprägungen. Bei der Kürzung von Skalen sollten daher diejenigen Items modifiziert oder eliminiert werden, die eine geringe Trennschärfe und/oder einen niedrigen Cronbachs-Alpha-Koeffizienten aufweisen (Döring & Bortz, 2016).

Dimensionalität als Indikator für Konstruktvalidität

Um die Eignung der Skalen zu untersuchen, wurde nach dem Einsatz im Feldtest die Dimensionalität der Skalen anhand von explorativen Faktorenanalysen geprüft. Die Eindimensionalität ist dabei ein Indikator für die Validität einer Skala. Eindimensionalität ist gewährleistet, wenn jeweils nur ein Faktor mit einem Eigenwert größer als 1 identifiziert werden kann und die Ladungen auf diesem Faktor ausreichend hoch sind ($\lambda \geq .4$; Loewenthal & Lewis, 2021). Eine hohe positive Faktorladung deutet auf einen hohen Anteil der durch den Faktor erklärten Messwertevarianz hin (Gäde et al., 2020). Mehrere extrahierte Faktoren bei theoretisch angenommener Eindimensionalität des Konstruktes können darauf hindeuten, dass die Skala auch Aspekte anderer Konstrukte misst. Skalen für die Messung von Konstrukten mit theoretisch begründetem mehrdimensionalem Charakter

Henicz, F., Welsandt, N. J., Fortunati, F., Abs, H. J. & Winther, E. (2024). Validierung des Fragebogeninstruments anhand der Feldtestdaten. In E. Winther & H. J. Abs (Hrsg.), *ECON 2022. Ökonomische Bildung in Jahrgang 8: Konzeption, Kompetenzen und Einstellungen* (S. 134–144). Waxmann. CC BY 4.0. https://doi.org/10.31244/9783830998389.4_4

ergeben in der Analyse idealerweise die entsprechende Anzahl an Faktoren und jeweils hohe Faktorladungen auf dem angenommenen Faktor bzw. niedrige Faktorladungen auf den anderen Faktoren. Im Falle von Abweichungen der Faktorzahl zur theoretisch angenommenen Dimensionalität des Konstruktes oder hohen Querladungen sollten die jeweiligen Items überarbeitet oder eliminiert werden (Schwarz, 2023).

Im Folgenden werden die faktoriellen Strukturen, die Reliabilitätsmaße für die interne Konsistenz sowie die Trennschärfen der in Kapitel 3.5 aufgeführten Skalen des Fragebogens aus den ECON-2022-Feldtestdaten dargestellt und zu den Messeigenschaften des Instrumentes der Hauptstudie in Bezug gesetzt.

4.4.1 Validierung der Befragungsinstrumente zu ökonomischen Einstellungen

Ökonomische und nachhaltigkeitsbezogene Selbstwirksamkeit als zwei eigenständige Konstrukte

Die Faktorenanalyse der Skalen für die ökonomische und nachhaltigkeitsbezogene Selbstwirksamkeit ergab erwartungsgemäß zwei Faktoren: Die Befunde der Itemanalysen für die *ökonomisch orientierte Selbstwirksamkeit* ergeben eine Skalen-Reliabilität von $\alpha = 0.791$, welche einen guten Wert darstellt (Field, 2018). Für Item F5g zeigt sich ein höherer Cronbachs-Alpha-Wert, würde das Item aus der Skala ausgeschlossen (siehe Tabelle 4.4.1). Vor dem Hintergrund der Befunde der Faktorenanalyse sowie aufgrund der geringen Trennschärfe und des Alpha-Wertes wurde hier die Entfernung des Items aus der Skala diskutiert; theoretisch lässt sich der Erhalt des Items in der Skala allerdings mit der inhaltlichen Passung aus dem Test begründen und wurde deshalb beibehalten.

Tabelle 4.4.1: Trennschärfe und Cronbachs Alpha für die Skala der ökonomischen Selbstwirksamkeit

	Trennschärfe	α, wenn Item entfernt
Was denkst du: Wie gut bist du darin, die folgenden Dinge zu tun?		
F5a Die eigenen Ausgaben und Einnahmen immer im Blick behalten	0.433	0.779
F5b Über Handel und Konsum diskutieren	0.543	0.760
F5c Werbetricks erkennen und einschätzen	0.536	0.761
F5d Anderen schwierige wirtschaftliche Themen erklären (z. B. den Wirtschaftskreislauf)	0.597	0.748
F5e Preise vergleichen und verdeckte Preiserhöhungen erkennen	0.560	0.756
F5f Wirtschaftliche Kenntnisse bewusst einsetzen, um Entscheidungen zu treffen (z. B. Abschluss einer Handy-Versicherung)	0.621	0.743
F5g Rabatte in Euro und Prozent berechnen	0.376	0.797

Für die Skala der *nachhaltigkeitsbezogenen Selbstwirksamkeit* zeigen die Befunde der Itemanalysen eine hohe Reliabilität von $\alpha = 0.831$. Die Cronbachs-Alpha-Werte, wenn ein jeweiliges Item ausgeschlossen würde, zeigen, dass alle fünf Items die Selbstwirksamkeit mit Blick auf Nachhaltigkeit verlässlich zu messen scheinen und keines eliminiert werden muss (siehe Tabelle 4.4.2). Auch die Trennschärfe ist bei allen Items mit Werten von größer als 0.60 hoch. Somit wurden alle Items auch vor dem Hintergrund der hohen Faktorladungen für die Hauptstudie beibehalten.

Tabelle 4.4.2: Trennschärfe und Cronbachs Alpha für die Skala der nachhaltigkeitsbezogenen Selbstwirksamkeit

	Trennschärfe	α, wenn Item entfernt
Was denkst du: Wie gut bist du darin, die folgenden Dinge zu tun?		
F6a Nachhaltige Entwicklung erklären	0.604	0.806
F6b Nachhaltige Produkte von nicht nachhaltigen Produkten unterscheiden	0.615	0.802
F6c Etwas gegen Ressourcenverschwendung tun	0.642	0.794
F6d Auswirkungen meines Konsums auf die Umwelt beim Einkauf einbeziehen	0.659	0.790
F6e Auswirkungen meines Konsums auf andere Menschen beim Einkauf einbeziehen	0.632	0.797

Vier Dimensionen nachhaltiger Konsumentscheidungen

Für die Erhebung der Relevanz von nachhaltigen Konsumentscheidungen (Erhebungsinstrument 2) wurden vier Faktoren extrahiert, die sich den Dimensionen nachhaltiger Entwicklung (ökonomisch, ökologisch, sozial) zuordnen lassen bzw. die Unterteilung der ökonomischen Dimension in zwei Subdimensionen nahelegt. Dabei sind alle Faktorladungen hoch; überwiegend bewegen sich die Ladungen bei Werten über 0.80, was eine starke Zugehörigkeit zu den jeweiligen Dimensionen impliziert. In Anlehnung an die ökonomischen Subdimensionen der Originalskala lassen sich die zwei ökonomischen Faktoren als ‚Konsum innerhalb der eigenen Mittel' und ‚genügsamen Konsum' beschreiben. Die Subdimension des kollaborativen Bewusstseins fällt hier unter den Faktor des genügsamen Konsums, was sich inhaltlich erklären lässt, da alle fünf Items auf den Verzicht eines nicht unbedingt notwendigen Kaufes abstellen (Ziesemer et al., 2016). Zwar grenzen die Autor*innen den kollaborativen vom maßvollen Konsum ab, weil beim kollaborativen Konsum trotz Vermeidung des Produktkaufs weiterhin eine Produktnutzung stattfindet, die Items zum maßvollen Konsum erheben aber nicht explizit, ob statt eines Tauschs dennoch eine Produktnutzung über andere Wege stattfindet.

Der Cronbachs-Alpha-Wert für die vierdimensionale Gesamtskala beträgt 0.924, was für eine sehr hohe interne Konsistenz der Skalen spricht. Die Cronbachs-Alpha-Werte der jeweiligen Sub-Skalen liegen alle über 0.810 (siehe Tabelle 4.4.3). Dabei lassen sich keine Items identifizieren, ohne die eine der Skalen eine höhere Reliabilität aufweisen würde. Die Trennschärfe aller Items beträgt größer als 0.50 und ist somit als hoch zu bewerten (Döring & Bortz, 2016).

Tabelle 4.4.3: Trennschärfe und Cronbachs Alpha für die Sub-Skalen der Relevanz nachhaltiger Konsumentscheidungen

Ökologisch (α= 0.933)	Trennschärfe	α, wenn Item entfernt
Wie wichtig ist dir persönlich, dass ein Produkt aus recyclingfähigen Materialien besteht?	0.827	0.918
Wie wichtig ist dir persönlich, dass ein Produkt sich umweltschonend entsorgen lässt?	0.840	0.915
Wie wichtig ist dir persönlich, dass ein Produkt umweltverträglich verpackt ist?	0.840	0.915
Wie wichtig ist dir persönlich, dass ein Produkt rohstoffschonend hergestellt wird?	0.798	0.923
Wie wichtig ist dir persönlich, dass ein Produkt ohne Umweltverschmutzung hergestellt wird?	0.820	0.920
Sozial (α= 0.945)		
Wie wichtig ist dir persönlich, dass bei der Herstellung eines Produktes die Menschenrechte der Arbeitnehmer*innen eingehalten werden?	0.841	0.934
Wie wichtig ist dir persönlich, dass bei der Herstellung eines Produktes keine Kinderarbeit eingesetzt wird?	0.832	0.936
Wie wichtig ist dir persönlich, dass bei der Herstellung eines Produktes Arbeitnehmer*innen nicht diskriminiert werden?	0.857	0.931
Wie wichtig ist dir persönlich, dass bei der Herstellung eines Produktes keine Arbeitnehmer*innen zur Arbeit gezwungen werden?	0.860	0.931
Wie wichtig ist dir persönlich, dass bei der Herstellung eines Produktes Arbeitnehmer*innen fair bezahlt werden?	0.862	0.930
Ökonomisch (Konsum innerhalb der eigenen Mittel; α=0.890)		
Wie wichtig ist dir persönlich, Produkte zu kaufen, die dich finanziell nicht stark belasten?	0.760	0.864
Wie wichtig ist dir persönlich, Produkte zu kaufen, ohne dass du dich dafür in Zukunft einschränken musst?	0.803	0.827
Wie wichtig ist dir persönlich, Produkte zu kaufen, ohne dass du dich dadurch langfristig verschuldest?	0.792	0.837
Ökonomisch (genügsamer Konsum; α= 0.810)		
Wie wichtig ist dir persönlich, nur Produkte zu kaufen, die du wirklich brauchst?	0.661	0.753
Wie wichtig ist dir persönlich, nur Produkte zu kaufen, die du wirklich nutzt?	0.663	0.752
Wie wichtig ist es dir, nach Möglichkeit ein Produkt von Freunden oder Bekannten auszuleihen anstatt es zu kaufen?	0.549	0.788
Wie wichtig ist es dir, nach Möglichkeit ein Produkt mit anderen zu teilen anstatt es selbst zu besitzen?	0.552	0.787
Wie wichtig ist dir persönlich, nur Produkte zu kaufen, die keine Luxusprodukte sind?	0.563	0.783

Vor dem Hintergrund der hohen und eindeutigen Faktorladungen, der hohen Trennschärfen und Cronbachs-Alpha-Werte wurden in den o. g. Skalen keine Veränderungen für die Hauptstudie vorgenommen.

Kollaborativer und maßvoller Konsum

Die Prüfung der Skala zu ökonomisch nachhaltigem Konsum (Belief-Komponente) aus dem Erhebungsinstrument 3 ergab zwei Faktoren. Hier ist anzumerken, dass aufgrund der Einwilligung zur Teilnahme am Fragebogen ein geringerer Stichprobenumfang (N = 620) als für die Teilnahme am Test vorliegt. Bei der Faktorenstruktur zeigen sich Abweichungen zur Skala der Importance-Komponente in den Ladungen der Items auf die Faktoren, die sich zum Teil durch die veränderte Stichprobe erklären lassen: Die kollaborative Dimension wird hier als eigenständiger Faktor mit hohen Ladungen der Items (> 0.85) identifiziert. Der zweite Faktor lässt sich als das Bewusstsein für maßvollen Konsum beschreiben und umfasst den Konsum innerhalb der eigenen Mittel und den genügsamen Konsum,

wie die Struktur der Originalskalen es vorschlägt (Ziesemer et al., 2016). Es fällt auf, dass Item F8c („… wenn es kein Luxusprodukt ist.“) niedrigere Faktorladungen aufweist als die anderen Items und die Ladungen dadurch nicht eindeutig zu interpretieren sind. Die leicht höhere Ladung zeigt sich auf dem Faktor des kollaborativen Konsums, was bedeuten könnte, dass Jugendliche Luxusprodukte eher teilen oder ausleihen, als sie zu kaufen.

Adaption der Skala für ökonomisch nachhaltigen Konsum

Die Itemanalysen ergaben für die Skala des ökonomisch nachhaltigen Konsums eine Skalenreliabilität von $\alpha = 0.708$. Die beiden Items, die mit hohen Werten auf dem Faktor für kollaborativen Konsum laden (F8g und F8h), zeigen einen höheren oder gleich hohen Cronbachs-Alpha-Wert, wenn das jeweilige Item exkludiert würde (siehe Tabelle 4.4.4). Die Trennschärfe für diese beiden Items liegt zwar über 0.20, ist aber bei beiden Items mit 0.281 eher niedrig. Auch Item F8c zeigt eine eher geringe Trennschärfe (0.327), was vor dem Hintergrund der erschwerten Interpretation ein weiterer Hinweis für den möglichen Ausschluss des Items aus der Skala ist.

Tabelle 4.4.4: Trennschärfe und Cronbachs Alpha für die Skala für ökonomisch nachhaltigen Konsum

	Trennschärfe	α, wenn Item entfernt
Auch wenn ich mir ein Produkt finanziell leisten könnte, würde ich es nur dann kaufen, ...		
F8a … wenn ich dieses Produkt auch wirklich brauche.	0.505	0.660
F8b … wenn ich dieses Produkt auch wirklich nutze.	0.507	0.661
F8c … wenn es kein Luxusprodukt ist.	0.327	0.695
F8d … wenn mich die Ausgaben dafür finanziell nicht stark belasten.	0.449	0.669
F8e … wenn ich mich dadurch langfristig nicht verschulde.	0.466	0.665
F8f … wenn ich mich dadurch in der Zukunft nicht einschränken muss.	0.454	0.668
Auch bei Produkten, die ich mir finanziell leisten kann, überlege ich immer, …		
F8g … ob ich mir das Produkt von Freunden oder Bekannten ausleihen kann.	0.281	0.711
F8h … ob ich mir das Produkt mit anderen teilen kann anstatt es zu kaufen.	0.281	0.708

Bei der separaten Betrachtung der internen Konsistenz der Sub-Skalen fällt auf, dass die Items, die sich unter *maßvollem Konsum* zusammenfassen lassen, ein $\alpha = 0.787$ aufweisen, ohne dass der Ausschluss eines Items den Alpha-Wert der Skala erhöhen würde. Die Items, die sich unter *kollaborativem Konsum* (inklusive Item F8c) fassen lassen, weisen hingegen ein $\alpha = 0.599$ auf. Dabei wäre der Alpha-Wert deutlich höher, würde Item F8c aus der Skala exkludiert ($\alpha = 0.776$). Da eine Skala mit lediglich zwei Items ein Konstrukt nicht aussagekräftig repräsentiert und die Trennschärfen dazu eher niedrig waren, wurde der Ausschluss der Items F8g und F8h beschlossen.

So wurden letztlich drei Items (F8c, F8g und F8h) aus der Skala ausgeschlossen, um Eindimensionalität der Skala sowie eine höhere Reliabilität zu erreichen. Die anschließend erneute Prüfung der Faktorstruktur der verbleibenden fünf Items identifizierte eine eindimensionale Struktur der Skala mit insgesamt zufriedenstellenden Faktorladungen größer als 0.639. Der erhöhte Cronbachs-

Alpha-Wert der Skala von 0.787 spricht für eine höhere Reliabilität als die Ursprungsskala. Die Trennschärfen sind insgesamt ebenfalls zufriedenstellend (siehe Tabelle 4.4.5), sodass von einer verlässlichen Messung von maßvollem Konsum durch die Skala ausgegangen werden kann.

Tabelle 4.4.5: Trennschärfe und Cronbachs Alpha für die gekürzte Skala für ökonomisch nachhaltigen Konsum

	Trennschärfe	α, wenn Item entfernt
Auch wenn ich mir ein Produkt finanziell leisten könnte, würde ich es nur dann kaufen, ...		
F8a … wenn ich dieses Produkt auch wirklich brauche.	0.455	0.779
F8b ... wenn ich dieses Produkt auch wirklich nutze.	0.572	0.745
F8d … wenn mich die Ausgaben dafür finanziell nicht stark belasten.	0.588	0.738
F8e … wenn ich mich dadurch langfristig nicht verschulde.	0.599	0.735
F8f … wenn ich mich dadurch in der Zukunft nicht einschränken muss.	0.609	0.731

4.4.2 Validierung der Befragungsinstrumente zu ökonomischen Bildungs- und Sozialisationskontexten

Verschiedene Kategorien von Quellen zur Informationsbeschaffung

Für die faktoranalytische Prüfung der Skalen der Quellen zur Informationsbeschaffung von Schüler*innen wurden zunächst beide Skalen gemeinsam untersucht: Die Analyse ergab vier Faktoren. Für beide Skalen laden dieselben Items auf einem der vier extrahierten Faktoren. Die Mehrdimensionalität deutet darauf hin, dass hier die Faktoren nicht auf Basis der inhaltlichen Struktur identifiziert wurden, sondern auf Basis der Art der Quelle: Für beide Themengebiete lassen sich das private Umfeld wie Freund*innen oder Verwandte, schulische Kontexte (Lehrkräfte), Medien, die eher passiv konsumiert werden, wie Fernsehen oder Zeitschriften, sowie das Internet als Faktoren für die Quellen differenzieren.

Inhaltliche Dimensionen der Quellen zur Informationsbeschaffung

Bei einzelner faktoranalytischer Untersuchung lassen sich erwartungsgemäß jeweils einzelne Faktoren extrahieren, die sich als Quellen der Informationsbeschaffung für ökonomische und nachhaltigkeitsbezogene Fragen interpretieren lassen.

Die Befunde der Itemanalysen für die ökonomisch orientierten Quellen ergeben eine Skalen-Reliabilität von $\alpha = 0.684$, für die Informationsquellen zu Nachhaltigkeit beträgt $\alpha = 0.762$. Tabellen 4.4.6 und 4.4.7 zeigen jeweils die Alpha-Werte der Skalen, wenn Items ausgeschlossen würden sowie die Trennschärfe der jeweiligen Items.

Tabelle 4.4.6: Trennschärfe und Cronbachs Alpha für die Skala für Quellen der Informationsbeschaffung zu ökonomischen Themen

	Trennschärfe	α, wenn Item entfernt
Bitte gib an, woher du wie viele Informationen zum Thema Geld und Wirtschaft bekommst.		
F1a Eltern, Erziehungsberechtigte oder erwachsene Verwandte	0.324	0.670
F1b Freundinnen und Freunde	0.406	0.645
F1c Lehrer oder Lehrerinnen	0.414	0.643
F1d Fernsehen oder Radio	0.504	0.612
F1e Zeitschriften	0.456	0.628
F1f Das Internet	0.383	0.657

Tabelle 4.4.7: Trennschärfe und Cronbachs Alpha für die Skala für Quellen der Informationsbeschaffung zu nachhaltigkeitsbezogenen Themen

	Trennschärfe	α, wenn Item entfernt
Bitte gib an, woher du wie viele Informationen zu Fragen der Nachhaltigkeit wie z. B. Nachhaltigem Konsum bekommst.		
F1aa Eltern, Erziehungsberechtigte oder erwachsene Verwandte	0.425	0.747
F1bb Freundinnen und Freunde	0.515	0.724
F1cc Lehrkräfte	0.447	0.742
F1dd Fernsehen oder Radio	0.572	0.709
F1ee Zeitschriften	0.567	0.710
F1ff Das Internet	0.498	0.729

Besonderheit der Skala

Für kein Item der beiden Skalen zeigt sich ein höherer Cronbachs-Alpha-Wert, wenn es exkludiert würde. Da die Items für diese Skala allerdings keine klassischen Indikatoren darstellen, die verschiedene Facetten eines Konstruktes messen, sondern die Indikatoren möglicherweise unabhängig voneinander sind, ist der Cronbachs-Alpha-Wert hier als Richtmaß nur bedingt aussagekräftig bzw. die interne Konsistenz der Skala nicht von so hoher Relevanz wie bei Skalen, die verschiedene Aspekte eines Personenmerkmals messen.

Verlässliche Messung des familiären Diskurses

Für die Skala des familiären Diskurses zu ökonomischen Themen lässt sich ein Faktor identifizieren. Die Items laden dabei im Wertebereich von 0.648 bis 0.741 auf dem Faktor, sodass angenommen werden kann, dass die Items durch den Faktor repräsentiert werden.

Der Cronbachs-Alpha-Wert für die Skala beträgt $\alpha = 0.786$ und ist somit akzeptabel. Bei keinem der Items ist der Cronbachs-Alpha-Wert höher, wenn das Item weggelassen würde (siehe Tabelle 4.4.8), was dafürspricht, dass die Items in der Summe eine reliable Skala für das Konstrukt bilden. Auch die Trennschärfe liegt für alle Items über dem Grenzwert von 0.20, sodass alle Items zwischen hoher und niedriger Zustimmung auf der Antwortskala differenzieren können. Vor dem Hintergrund der Ergebnisse der Faktorenanalyse und der Itemanalyse wurden alle sechs Items der Skala beibehalten.

Tabelle 4.4.8: Trennschärfe und Cronbachs Alpha für die Skala zum familiären Diskurs

	Trennschärfe	α, wenn Item entfernt
Wie oft sprichst du über die folgenden Geldfragen mit deinen Eltern oder Erziehungsberechtigten?		
F4a Wofür du dein Geld ausgibst	0.530	0.755
F4b Wofür du dein Geld sparst	0.568	0.745
F4c Das Einkommen der Familie	0.535	0.753
F4d Wie du Geld für Sachen bekommst, die du kaufen möchtest	0.584	0.741
F4e Nachrichten über Finanz- und Wirtschaftsfragen	0.485	0.765
F4f Online-Einkäufe	0.507	0.760

Einkommensquellen von Jugendlichen innerhalb und außerhalb des eigenen Haushaltes

Die Analyse der Skala für die Erhebung von Quellen für Einkommen der Schüler*innen ergab zwei Faktoren. Die faktorielle Struktur legt eine Differenzierung in Geldquellen innerhalb und außerhalb des eigenen Haushalts nahe. Da sich die Faktoren als zwei Dimensionen innerhalb desselben Konstruktes interpretieren lassen, ist die zweifaktorielle Struktur der Skala nicht als problematisch einzuschätzen.

Der Cronbachs-Alpha-Wert für die Skala beträgt 0.614, was eher als unzureichend zu bewerten ist (Field, 2018). Da die Items allerdings nicht klassischerweise verschiedene Facetten desselben Personenmerkmals erheben, sondern Aspekte erfassen, die nicht notwendigerweise untereinander korrelieren müssen (verschiedene Geldquellen), sollte die interne Konsistenz für diese Skala nicht als leitendes Gütemaß überschätzt werden. Bei der Bewertung der Reliabilität sollte stets der Kontext des Messinstrumentes einbezogen werden (Cortina, 1993). Die Trennschärfe ist bei den meisten Items eher gering, liegt aber bei allen Items über dem Grenzwert von 0.20. Für diese Skala wurden alle Items beibehalten.

Tabelle 4.4.9: Trennschärfe und Cronbachs Alpha für die Skala zu Einkommensquellen

	Trennschärfe	α, wenn Item entfernt
Wie oft hast du auf die folgende Art in den letzten zwölf Monaten Geld bekommen?		
F3a Geld als Belohnung für Mithilfe in der Familie	0.397	0.545
F3b Taschengeld, ohne dafür etwas getan zu haben	0.288	0.602
F3c Jobs außerhalb der Schule (z. B. Werbung austragen, Babysitten)	0.331	0.579
F3d Geschenke von Verwandten oder Freund*innen	0.512	0.512
F3e Verkauf von Sachen (z. B. Flohmarkt oder Internet)	0.366	0.561

Hinweise auf unterschiedliche Dimensionen im Umgang mit Geld und Konsum

Für die Skala zur Erhebung der Erfahrung als Konsument*innen wurden zwei Faktoren identifiziert: Die Zuordnung der Items zu den Faktoren lässt sich dabei nicht eindeutig interpretieren. So lädt das Item F7b („Etwas online eingekauft (allein oder mit jemandem aus der Familie)“) auf beiden Faktoren, wenn auch auf dem ersten Faktor weniger stark. Der zweite Faktor könnte darauf hindeuten, dass das eigene Geld zu verwalten und zu managen ein von den Konsumerfahrungen zu differenzierendes Konstrukt darstellt.

Das Reliabilitätsmaß der Gesamtskala liegt bei $\alpha = 0.765$, was einen akzeptablen Wert darstellt. Bei der Entfernung von Item F7a („Das Wechselgeld beim Einkaufen überprüft") ließe sich ein Cronbachs Alpha von 0.786 erreichen; dazu weist dieses Item eine sehr geringe Trennschärfe auf, die unter dem Grenzwert von 0.2 liegt (siehe Tabelle 4.4.10).

Tabelle 4.4.10: Trennschärfe und Cronbachs Alpha für die Skala für Erfahrungen als Konsument*innen

	Trennschärfe	α, wenn Item entfernt
Wie oft hast du die folgenden Dinge in den letzten zwölf Monaten gemacht?		
F7a Das Wechselgeld beim Einkaufen überprüft	0.173	0.786
F7b Etwas online eingekauft (allein oder mit jemandem aus der Familie)	0.431	0.746
F7c Etwas mit dem Smartphone bezahlt	0.576	0.726
F7d Etwas mit einer Bankkarte bezahlt (z. B. EC-Karte/Girokarte)	0.555	0.727
F7e Etwas mit einer anderen Guthabenkarte bezahlt (z. B. Prepaid-Card für Essen, Transport oder Eintritte)	0.543	0.730
F7f Überprüft, wie viel Geld du zur Verfügung hast	0.357	0.755
F7g Anderen Menschen mit dem Smartphone Geld geschickt	0.555	0.730
F7h Geld auf ein Sparkonto eingezahlt, z. B. bei einer Bank	0.588	0.725
F7i Bargeld zu Hause gespart	0.134	0.786
F7j Bezahlt um eine App herunterzuladen	0.501	0.737

Präzisierung der Skala für Konsumerfahrungen

Bei separater Überprüfung der internen Konsistenz der Sub-Skalen zeigte sich, dass die Sub-Skala des ersten Faktors einen deutlich höheren Alpha-Wert aufweist ($\alpha = 0.851$) als die Gesamtskala. Die Sub-Skala des zweiten Faktors hingegen zeigt einen deutlich niedrigeren Wert von $\alpha = 0.598$. Bei beiden Skalen ließen sich die Alpha-Werte nicht durch Ausschluss eines oder mehrerer Items erhöhen. Die Ergebnisse der Analysen sprechen dafür, Items F7a und F7b aus der Skala zu entfernen. Um Eindimensionalität der Skala zu erreichen, wurden auch die Items F7f und F7i aus der Skala exkludiert, die den zweiten Faktor bilden, sodass sechs Items zur Messung von Konsumerfahrungen bestehen blieben.

Der deutlich höhere Cronbachs-Alpha-Wert bei der gekürzten Skala spricht dafür, dass die Skala in reduzierter Form reliabler misst. Daher wurden die sechs Items für die Haupterhebung beibehalten (siehe Tabelle 4.4.11).

Tabelle 4.4.11: Trennschärfe und Cronbachs Alpha für die gekürzte Skala für Erfahrungen als Konsument*innen

	Trennschärfe	α, wenn Item entfernt
Wie oft hast du die folgenden Dinge in den letzten zwölf Monaten gemacht?		
F7c Etwas mit dem Smartphone bezahlt	0.601	0.833
F7d Etwas mit einer Bankkarte bezahlt (z. B. EC-Karte/Girokarte)	0.629	0.828
F7e Etwas mit einer anderen Guthabenkarte bezahlt (z. B. Prepaid-Card für Essen, Transport oder Eintritte)	0.627	0.828
F7g Anderen Menschen mit dem Smartphone Geld geschickt	0.717	0.811
F7h Geld auf ein Sparkonto eingezahlt, z. B. bei einer Bank	0.652	0.823
F7j Bezahlt, um eine App herunterzuladen	0.595	0.833

Lerngelegenheiten zu wirtschaftlichen Themen bilden eine Dimension

Für die Skala zur Messung des Umfangs an schulischen Lerngelegenheiten ließ sich ein Faktor identifizieren, sodass angenommen werden kann, dass die Items eine Dimension abbilden. Lediglich Item F2c („Wie man Zinsen und Zinseszinsen berechnet“) zeigt eine deutlich niedrigere Faktorladung von 0.443.

Der Cronbachs-Alpha-Wert für die Skala beträgt 0.898, was als hoch zu bewerten ist. Die interne Konsistenz der Skala würde sich erhöhen, wenn o. g. Item F2c exkludiert würde (siehe Tabelle 4.4.12). Ebenso weist dieses Item im Vergleich zu den anderen Items dieser Skala eine deutlich geringere Trennschärfe auf (0.373; die restlichen Items > 0.60). Auf Basis der auffälligen Werte in beiden Analysen wurde die Entfernung des Items F2c diskutiert. Aufgrund der inhaltlichen Passung zum Test wurde es jedoch in der Skala beibehalten.

Tabelle 4.4.12: Trennschärfe und Cronbachs Alpha für die Skala für den Umfang an schulischen Lerngelegenheiten zu wirtschaftlichen Themen

	Trennschärfe	α, wenn Item entfernt
In welchem Umfang hast du in der Schule etwas über die folgenden Dinge gelernt?		
F2a Worauf man beim Online-Handel achten muss	0.607	0.892
F2b Wie man Bedürfnisse und Bedarf unterscheiden kann	0.614	0.891
F2c Wie man Zinsen und Zinseszinsen berechnet	0.373	0.905
F2d Wie man wirtschaftliche Ziele erreichen kann	0.725	0.885
F2e Welchen Einfluss man als Verbraucher*in auf die Preisbildung hat	0.669	0.888
F2f Wie eigene Kaufentscheidungen das Leben von Menschen in anderen Ländern beeinflussen können	0.693	0.886
F2g Welche Folgen das eigene wirtschaftliche Handeln für andere hat	0.722	0.885
F2h Wie sich die Arbeitswelt durch die Digitalisierung ändert	0.618	0.891
F2i Welche Rechte und Pflichten man als Verbraucher*in hat	0.648	0.889
F2j Wie man die eigenen Einnahmen und Ausgaben im Blick behalten kann	0.681	0.887
F2k Wie Unternehmen funktionieren	0.635	0.890

Vergleich der internen Konsistenz der Skalen aus dem Feldtest und aus dem Haupttest

In Tabelle 4.4.13 sind nachfolgend die Cronbachs-Alpha-Werte der Skalen aus der Hauptstudie im Vergleich zu den Skalen aus dem Feldtest zu entnehmen. Im Vergleich zeigt sich, dass sich die Reliabilitäten überwiegend verbessert oder kaum verändert haben. Die interne Konsistenz der Skalen der ökonomischen Dimension von Nachhaltigkeit, der ökonomischen Informationsquellen, der Einkommensquellen sowie der schulischen Lerngelegenheiten zeigen in der Hauptstudie geringere Werte als im Feldtest. Die Veränderungen sind allerdings entweder gering und die Alpha-Werte sind weiterhin hoch (z. B. die ökonomische Dimension von Nachhaltigkeit und schulische Lerngelegenheiten) oder es handelt sich um die Skalen, für die das Cronbachs Alpha weniger aussagekräftig ist (Informations- und Einkommensquellen). Daher sind die Werte der internen Konsistenz insgesamt als zufriedenstellend zu beurteilen.

Tabelle 4.4.13: Vergleich der internen Konsistenz der Skalen aus dem Feldtest und aus dem Haupttest

Skala	Cronbachs Alpha Feldtest	Cronbachs Alpha Haupttest
Selbstwirksamkeit (ökonomisch)	0.791	0.800
Selbstwirksamkeit (Nachhaltigkeit)	0.832	0.836
Relevanz nachhaltiger Konsumentscheidungen (gesamt)	0.924	0.937
Ökologisch	0.933	0.919
Sozial	0.945	0.943
Konsum innerhalb der eigenen Mitte (3 Items ökonomisch)	0.890	0.893
Genügsamer Konsum (5 Items ökonomisch)	0.810	0.828
Ökonomische Dimension Nachhaltigkeit (separat)	0.787	0.768
Quellen zur Informationsbeschaffung (ökonomisch)	0.684	0.627
Quellen zur Informationsbeschaffung (Nachhaltigkeit)	0.762	0.767
Familiärer Diskus	0.786	0.801
Einkommensquellen	0.614	0.603
Erfahrungen als Konsument*in	0.851	0.799
Schulische Lernerfahrungen	0.898	0.898

Literatur

Cortina, J. (1993). What is Coefficient Alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98.

Döring, N. & Bortz, J. (2016). Operationalisierung. In N. Döring & J. Bortz (Hrsg.), *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (S. 221–283). Springer. http://doi.org/10.1007/978-3-642-41089-5.

Field, A. (2018). Discovering Statistics Using IBM SPSS Statistics. SAGE.

Gäde, J. C., Schermelleh-Engel, K. & Brandt, H. (2020). Konfirmatorische Faktorenanalyse (CFA). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 615–659). Springer. https://doi.org/10.1007/978-3-662-61532-4_24.

Loewenthal, K. M. & Lewis, C. A. (2021). *An Introduction to Psychological Tests and Scales*. Routledge.

Moosbrugger, H. & Kelava, A. (2012). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 7–26). Springer.

Schermelleh-Engel, K. & Werner, C. (2012). Methoden der Reliabilitätsbestimmung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 119–142). Springer.

Schwarz, J. (2023). *Faktoranalyse*. Universität Zürich. https://www.methodenberatung.uzh.ch/de/datenanalyse_spss/interdependenz/reduktion/faktor.html

Ziesemer, F., Peyer, M., Klemm, A. & Balderjahn, I. (2016). Die Messung von nachhaltigem Konsumbewusstsein. *Ökologisches Wirtschaften*, *4*(31), 24–26. https://doi.org/10.14512/OEW310424.

Bildnachweis S. 107:
Edward Burtynsky, Lithium Mines #1, Salt Flats, Atacama Desert, Chile, 2017