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.

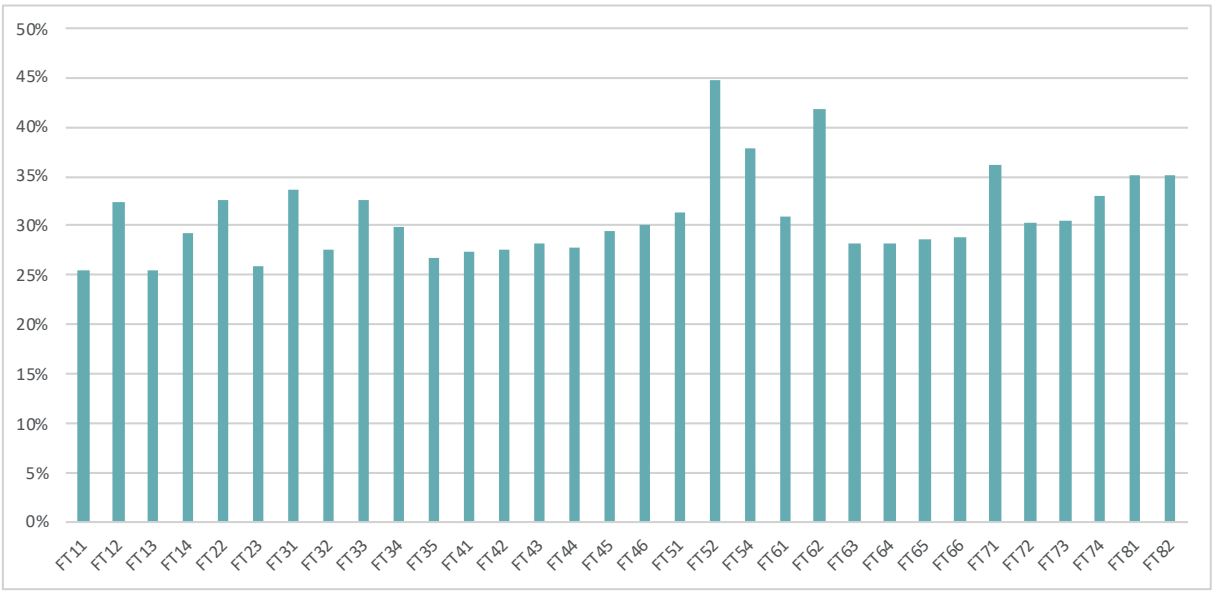

Abbildung 4.3.1: Fehlende Werte auf Itemebene der Hauptstudie

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).

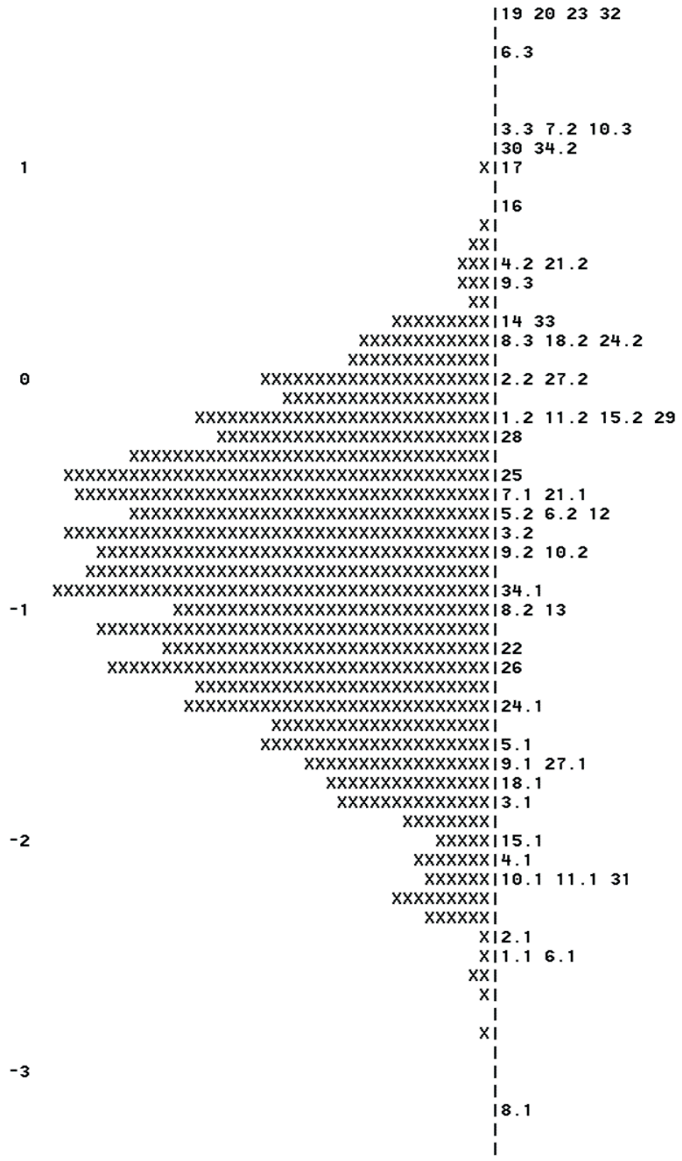

Abbildung 4.3.2: Personen-Item-Map für die Feldteststichprobe des Testinstruments TBA-EL

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

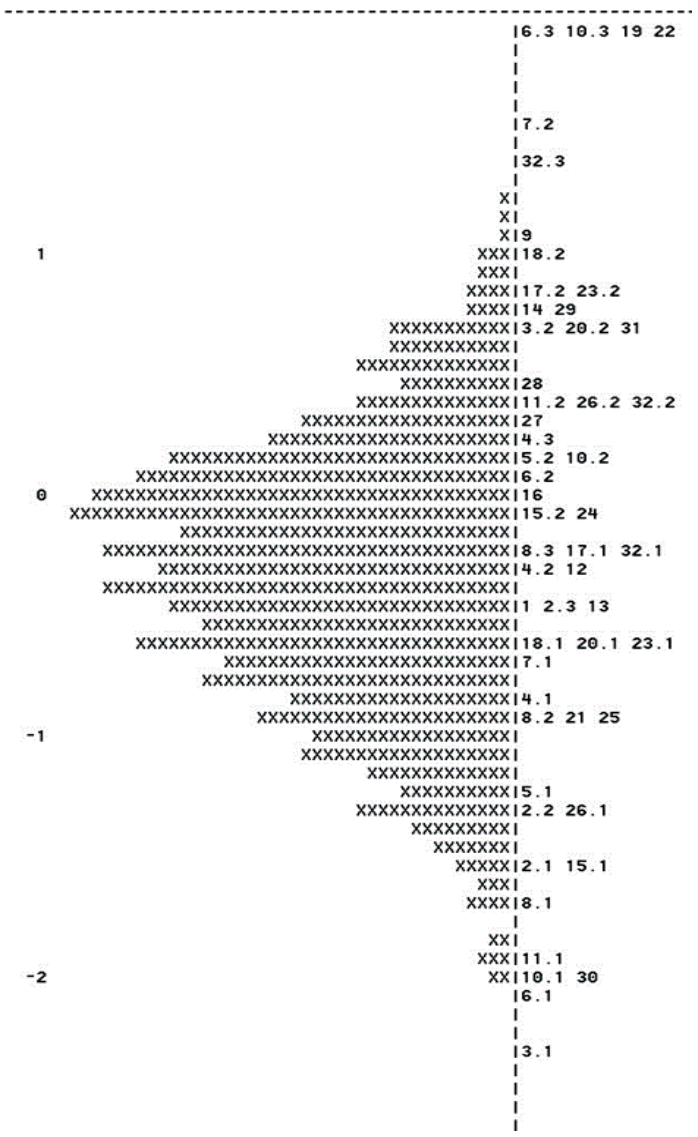

Abbildung 4.3.3: Personen-Item-Map für die Haupttestst ichprobe des Testinstruments TBA-EL

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).

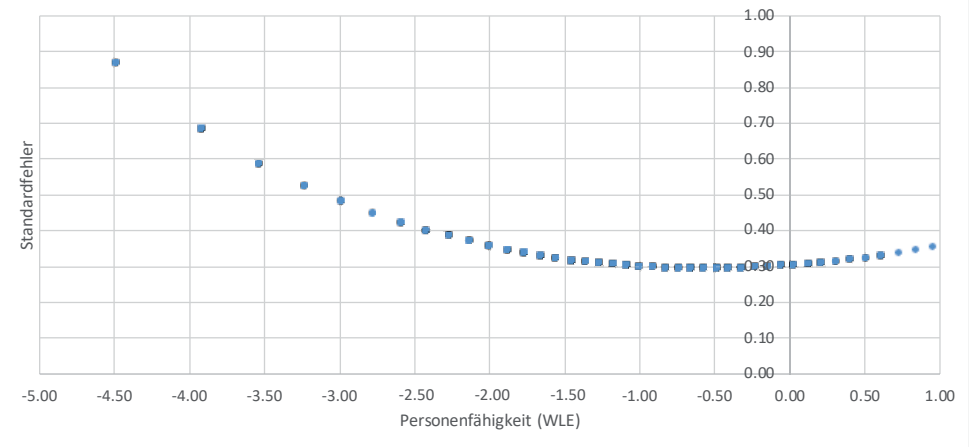

Abbildung 4.3.4: Personenfähigkeiten und Standardfehler für die Feldteststichprobe des Testinstruments TBA-EL

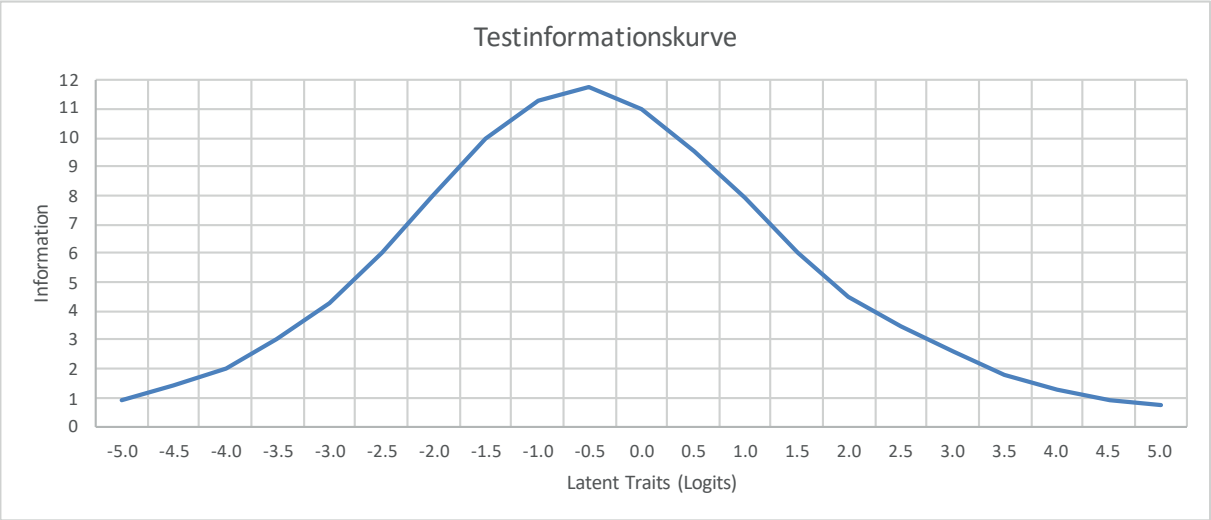

Abbildung 4.3.5: Testinformationsfunktion für die Feldteststichprobe des Testinstruments TBA-EL

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

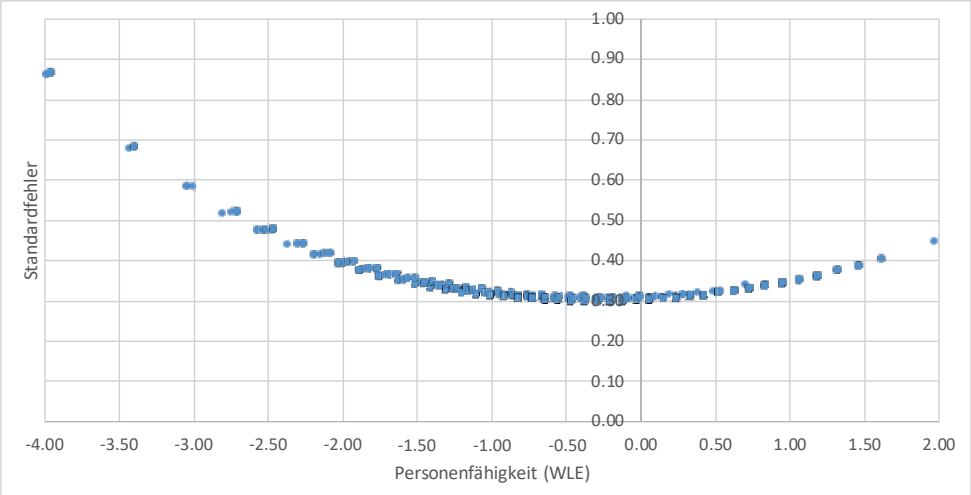

Abbildung 4.3.6: Personenfähigkeiten und Standardfehler für die Haupttestst ichprobe des Testinstruments TBA-EL

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).

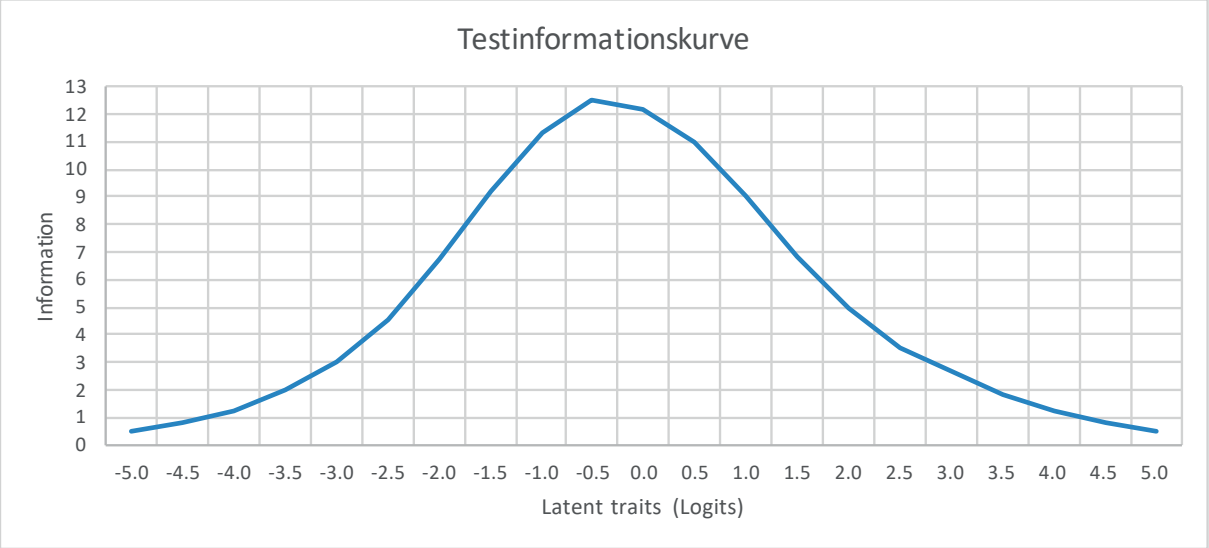

Abbildung 4.3.7: Testinformationsfunktion für die Haupttestst ichprobe des Testinstruments TBA-EL

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)

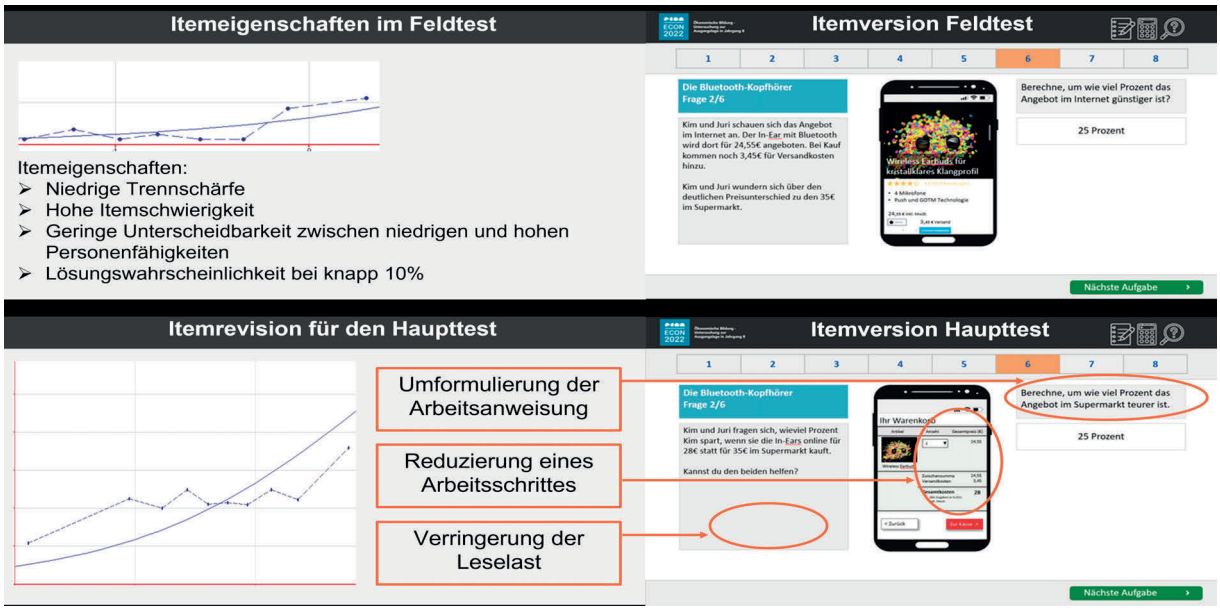

Abbildung 4.3.8: Itemrevision zwischen Feld- und Haupttest am Beispiel des Items 6_2 – Vermehrter Grundwert

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.