Zeitschrift für ökonomische Bildung, Heft 8, Jahrgang 2019, S. 60-98

Ökonomische Kompetenzen von Heranwachsenden: Entwicklung und Validierung eines Testinstruments

Luis Oberrauch

Universität Koblenz-Landau

Zusammenfassung

Die WIKO-BW-Studie untersucht die Entwicklung ökonomischer Kompetenz von Lernenden der Sekundarstufe I in Baden-Württemberg über vier Schuljahre hinweg. In der Startkohorte werden im Rahmen dieser Teilstudie kognitive Facetten ökonomischer Kompetenz am Ende der Klassenstufe 7 mit einem eigens für die WIKO-BW-Untersuchung entwickelten Instrument erfasst. Der Beitrag gibt einen Überblick über die Entwicklung des Testinstruments sowie dessen Skalierung. Die Inhaltsvalidität wurde im Vorfeld durch Expertenratings und curriculare Analysen überprüft. Mithilfe von Methoden der Item Response Theory werden anhand einer repräsentativen Stichprobe mit 1.689 Lernenden Itemparameter, Dimensionalität und Differential Item Functioning (DIF) untersucht. Der umfangreiche Datensatz ermöglicht zudem eine Analyse der Variation ökonomischer Kompetenz entlang von Schulformen und individuellen Schülermerkmalen. Das Instrument kann als eindimensional identifiziert werden und weist eine befriedigende psychometrische Qualität auf. Die Regressionsanalyse dokumentiert die existierende Heterogenität hinsichtlich ökonomischer Kompetenzen in Abwesenheit von Fachunterricht in der ökonomischen Domäne. Der Beitrag setzt den Startpunkt für eine Reihe von Kompetenzmessungen, die zusammen in einen klassen- und länderübergreifend anwendbaren Test of Economic Competences (TEC) münden.

Abstract

The WIKO-BW study examines the longitudinal development of economic competences in lower secondary schools in the German federal state of Baden-Württemberg. This substudy assesses cognitive facets of economic competences by means of a newly developed instrument in the first cohort at the end of grade 7. It provides a review of development processes and scaling. Content validity was established in advance by expert ratings and an analysis of school curricula. By using methods from Item Response Theory, we investigate item parameter, dimensionality and Differential Item Functioning (DIF) in a representative sample of 1,689 students. Moreover, the rich data set allows for examination of economic competences across school types and individual student characteristics. The instrument is identified as unidimensional and shows satisfactory psychometric characteristics. Regression analysis documents existing heterogeneity of economic competences in absence of a particular school subject within the economic domain. This article sets the stage for a series of analyses, that culminate in a Test of Economic Competences (TEC) which will be applicable across grades and countries.

1 Einleitung

Während der letzten Jahre hat die Implementierung ökonomischer Bildung in deutsche Schulcurricula neue Impulse erfahren. Neben Baden-Württemberg, welches zum Schuljahr 2017/18 das eigenständige Fach „Wirtschaft, Berufs- und Studienorientierung“ im Rahmen des neuen Bildungsplanes einführte, wird „Wirtschaft“ auch in Nordrhein-Westfalen ab dem Schuljahr 2019/20 flächendeckend einen höheren Stellenwert erhalten. Parallel entstanden über die letzten Jahre im deutschsprachigen Raum mehrere quantitative Querschnittserhebungen ökonomischer Fähigkeiten (Seeber et al. 2018; Macha 2015; Rumpold und Greimel-Fuhrmann 2016; Loerwald und Schnell 2014; Schumann und Eberle 2014), die größtenteils defizitäre Kenntnisse unter Lernenden feststellten.

Im Bundesland Baden-Württemberg legte die Testung eines Wirtschaftswissens unter Lernenden der 8. bis 12. Jahrgangsstufe (Würth und Klein 2001) und eine Kompetenzerhebung unter Lernenden der 9. bis 11. Jahrgangsstufe (Seeber et al. 2018) erhebliche Förderbedarfe offen. Mit dem Start des neuen Schulfaches untersucht die angelaufene WIKO-BW–Panelstudie Entwicklungen von Wirtschaftskompetenzen im Längsschnitt von der 7. bis zur 10. Klassenstufe. Langfristiges Erkenntnisinteresse ist die quasi-experimentelle Identifikation von Effekten des Schulfaches auf Kompetenzen, Einstellungen und das ökonomische Verhalten von Schülerinnen und Schülern. Dies erforderte die Entwicklung bzw. Adaption eines einfach zu implementierenden Testinstruments für die 7. Klassenstufe.

Bestehende Instrumente zeigen sich hinsichtlich des zugrunde liegenden theoretischen Modells und der zu erfassenden Taxonomiestufe uneinheitlich – viele Arbeiten zielen unter Vernachlässigung der Handlungsperspektive eher auf ein Lehrbuchwissen ab. Andere Beiträge beziehen zwar die ökonomisch geprägten Lebenssituationen in ihr Modell ein, machen sie aber nicht zum Ausgangspunkt der Testentwicklung, wie es das Integrationsmodell ökonomischer Kompetenz (IÖK) (Seeber et al. 2012) vorsieht. Unter Berücksichtigung, dass dieses Modell eine maßgebliche Rolle bei der Lehrplanentwicklung des neuen Faches in Baden-Württemberg spielte, wird in diesem Beitrag die Entwicklung und Validierung eines auf dem IÖK basierenden Testinstruments für die siebte Klassenstufe skizziert. Im Mittelpunkt stehen die inhaltliche Herleitung, psychometrische Eigenschaften der Skala sowie bedeutende Korrelate ökonomischer Kompetenz. Der Beitrag untersucht die Startkohorte aus einer Reihe von Wirtschaftskompetenzmessungen, die zusammen in einen

international und klassenübergreifend anwendbaren *Test of Economic Competence (TEC)* münden.

Im Ergebnis kann die Inhaltsvalidität des Instruments durch curriculare Analysen und Expertenratings bestätigt werden. Die empirische Analyse identifiziert faktoranalytisch sowie auf Basis der probabilistischen Testtheorie ein eindimensionales Konstrukt. Hinsichtlich Modellpassung, Trennschärfe und Item Bias, welcher durch eine Analyse des *Differential Item Functioning (DIF)* untersucht wird, zeigt das Instrument befriedigende psychometrische Eigenschaften. Die Regressionsanalyse bestätigt die bestehende Heterogenität ökonomischer Kompetenz in Abwesenheit von Fachunterricht. Demnach korrelieren ökonomische Kompetenzen positiv mit dem (approximierten) Bildungshintergrund der Eltern, dem Geschlecht (männlich) sowie mit Einstellungs- und Interessensdimensionen. Negative Effekte auf die Kompetenz können für den Migrationsstatus gezeigt werden. Den stärksten Gesamteffekt zeigt die Zugehörigkeit zur Schulform Gymnasium.

Der vorliegende Beitrag gliedert sich wie folgt: Das nachfolgende Kapitel beschreibt das Integrationsmodell ökonomischer Bildung als theoretische Grundlage sowie dessen Abgrenzung zu anderen elaborierten Ansätzen und gibt im Anschluss einen kursorischen Überblick über quantitative Wirtschaftskompetenzstudien sowie über festgestellte Korrelate ökonomischer Fähigkeiten. Der Prozess der Itementwicklung mit Blick auf die Inhaltsvalidität des Konstrukts wird in Kapitel 3 skizziert. Ein eigenes Kapitel widmet sich den Gütekriterien Validität und Reliabilität und beschreibt insbesondere, wie die verwendete Methodik eine Annäherung an den Validitätsbegriff leisten kann (Kapitel 4). Kapitel 5 beschreibt die Methode der Stichprobenziehung und die dazugehörige Design-Gewichtung. Die psychometrischen Eigenschaften des Tests sind Gegenstand von Kapitel 6. Dabei wird zuerst die Dimensionalität der Daten mithilfe einer Hauptkomponentenanalyse untersucht. Testtheoretische Analysen werden mit Methoden der *Item Response Theory* (*IRT)* durchgeführt, mit deren Hilfe die Passung von erhobenen Daten zu theoretischen Modellannahmen verifiziert werden kann. Ferner wird der Einsatz mehrparametrischer IRT-Modelle diskutiert. Die Testfairness für verschiedene Subgruppen wird anhand einer DIF-Analyse (*Differential Item Functioning*) auf Basis der verbreiteten ETS-Klassifikation überprüft. Zuletzt werden mittels eines hierarchischen Regressionsmodells Prädiktoren ökonomischer Kompetenz ausgewiesen.

2 Theoretischer Hintergrund und Forschungsstand

Theoretische Grundlage für die Testentwicklung bildet das Integrationsmodell ökonomischer Kompetenz (IÖK) von Seeber et al. (2012), das Anforderungen an Heranwachsende in ökonomisch geprägten Lebenssituationen systematisiert und in Baden-Württemberg in die Lehrplanentwicklung für das neue Schulfach Wirtschaft-, Berufs- und Studienorientierung Eingang gefunden hat. Die folgenden Abschnitte beschreiben die Kernaspekte des IÖK und wie sich dieses zu anderen (elaborierten) Konzepten verhält. Zuletzt wird ein kursorischer Überblick über bisherige Kompetenzmessungen sowie über bedeutende Korrelate gegeben.

2.1 Das Integrationsmodell ökonomischer Kompetenz (IÖK)

Der Kompetenzbegriff im IÖK orientiert sich grundsätzlich an der Definition von Weinert (2001), die motivationale, volitionale und soziale Bereitschaften einschließt. Der vorgestellte Test betrachtet jedoch ausschließlich kognitive Facetten ökonomischer Kompetenz. Volition und Motivation werden als Einstellungsdimension nur explorativ erfasst. Inhaltlich richtet sich das IÖK an der Kompetenzdefinition der DeGÖB und deren Rollenbeschreibungen aus. Kernfrage für die Konzeption war demnach, welche ökonomisch geprägten Lebenssituationen von Relevanz sind und welche Kompetenzen für die Bewältigung dieser benötigt werden (Seeber et al. 2018, 34). Im ersten Schritt werden die ökonomisch geprägten Lebenssituationen über verschiedene Rollen, die Akteure einnehmen können, definiert. Dazu gehören die Rollen der Verbraucher, der Erwerbstätigen und der Bürger (Retzmann et al. 2010, 78). Zur Bewältigung dieser Situationen benötigen die Heranwachsenden Kompetenzen, die mithilfe dreier Kompetenzbereiche klassifiziert werden. Diese gründen – neben den ökonomisch geprägten Lebenssituationen – auf dem Bildungsauftrag der Schulen, der neben Persönlichkeitsentwicklung und praktischer Lebensgestaltung die politische Partizipationsfähigkeit stärken soll. Ebenso steht gemäß allgemeinem Erziehungsauftrag die Bildung von Verantwortlichkeit im Mittelpunkt – Individuen sollen verantwortlich gegenüber sich selbst, gegenüber anderen Agenten und gegenüber der Sache handeln. Diese Maßstäbe wurden in die drei Kompetenzbereiche „Entscheidung und Rationalität“ (E & R), „Beziehung und Interaktion“ (B & I) und „System und Ordnung“ (S & O) übersetzt (Seeber et al. 2018, 36). Sie bilden folglich keine eigenen Dimensionen ab, sondern inhaltlich-theoretische Abgrenzungen eines globalen Kompetenzkonstrukts. Die Kompetenzbereiche werden mit den ökonomisch geprägten Lebenssituationen gekreuzt und in eine Matrix überführt (Abbildung 1).

		Lebenssituationen		
Kompetenzbereiche		**Verbraucher**	**Erwerbstätige**	**Bürger**
Entscheidung + Rationalität	➡	Ökonomische Situationen analysieren Handlungsfolgen analysieren und bewerten Veränderbarkeit erkennen und analysieren		
Beziehung + Interaktion	➡	Interessenkonstellationen verstehen und analysieren Kooperationen analysieren und bewerten Beziehungsgefüge analysieren		
System + Ordnung	➡	Märkte analysieren Wirtschaftssysteme und Ordnungen analysieren Politik ökonomisch beurteilen		

Abbildung 1: Integratives Modell ökonomischer Kompetenz (IÖK) in Seeber et al. 2012

Auf Basis der Kompetenzbereiche wurden Kompetenzitems entwickelt (Abschnitt 3), die Heranwachsende in eine fiktive Entscheidungssituation bringen (*Entscheidung & Rationalität*), sie das Handeln des ökonomischen Gegenübers beurteilen (*Beziehung & Interaktion*) oder systemische Zusammenhänge analysieren (*System & Ordnung*) lassen.

2.2 Weitere Modelle ökonomischer Kompetenz

Im Kompetenzraster der Deutschen Gesellschaft für Ökonomische Bildung (DeGÖB 2004) sind ebenfalls die Lebenssituationen (und somit auch die ökonomischen Rollen) maßgeblich für die Kompetenzanforderungen – das IÖK stellt somit eine Weiterentwicklung dieses Ansatzes dar. Innerhalb des DeGÖB-Ansatzes wurden Kompetenzbereiche definiert, die nicht eindeutig voneinander abgrenzbar sind und mit der Ethik einen domänenfremden Bereich einschließen (für eine kritische Auseinandersetzung vgl. Seeber et al. 2012, 19).

Die Bildungsstandards der DeGÖB haben ebenso in das sehr umfassende und sich an Weinerts Kompetenzdefinition ausrichtende Siegener Modell ökonomischer Kompetenz (Macha und Schuhen 2011) Eingang gefunden. Ökonomische (Kern-)Kompetenz wird dort beschrieben als „Fähigkeit, in verbal und mathematisch geprägten Situationen, Rollen und Kontexten (1) ökonomische Fragestellungen zu erkennen, (2) ökonomische Phänomene zu beschreiben, (3) ökonomisches Wissen in unterschiedlichen Handlungssituationen anzuwenden und (4) … sich mit ökonomischen Themen…reflektierend… auseinander zu

setzen…“ (Macha und Schuhen 2011, 21). Die ersten beiden Aspekte greifen dabei in Anlehnung an Achtenhagen und Winther (2006) verstehensbasierte Kompetenzfacetten, die Punkte (3) und (4) eher handlungsorientierte Kompetenzfacetten auf. Spätere Modifikationen übersetzten die vier Aspekte in einen Nukleus ökonomischer Kompetenz und bildeten entsprechende Schnittmengen für eine Itementwicklung (Macha 2015, 39). Die ökonomisch geprägten Lebenssituationen fließen somit in das Modell ein, sie sind jedoch – anders als im IÖK – nicht der Ausgangspunkt für die Entwicklung eines Testinstruments. Zudem haben mathematische Kompetenzen im Modell einen relativ höheren Stellenwert.

Einen alternativen Ansatz verfolgen die Schweizer Wissenschaftler Schumann und Eberle (2014), deren Kompetenzerhebung ebenso motivationale und volitionale Aspekte gemäß Weinert einschließt. Die Autoren definieren eine ökonomische Grundkompetenz („Economic Literacy“), die den Menschen als Referenzfigur in den Mittelpunkt stellt. Dieser soll in der Lage sein, „wirtschaftsbezogene Problemstellungen zu verstehen, zu analysieren und begründete Schlüsse für (potentielle) Lösungen daraus zu ziehen, also über wirtschaftsbürgerliche Kompetenz [verfügen]“ (Schumann und Eberle 2014, 107). Testitems wurden hier jedoch nicht aus Lehrplananalysen abgeleitet, sondern sollen „authentische und alltagsbezogene Darstellungen“ (ebd.) einbeziehen. Dazu wurden mittels Inhaltsanalysen aus 1.400 Zeitungen 80.000 ökonomiebezogene Begriffe extrahiert und den klassischen Teildomänen Volkswirtschaftslehre, Betriebswirtschaftslehre und Rechnungswesen zugeteilt, woraus wiederum modifizierte Zeitungsartikel als Inhaltsanker für die Itementwicklung dienten. Es handelt sich somit um einen sehr inhaltsbezogenen Ansatz, der die im IÖK verankerte Handlungsperspektive vernachlässigt.

2.3 Financial Literacy

Grundsätzlich werden im IÖK finanzielle Kompetenzen als Teilmenge von ökonomischen Kompetenzen betrachtet – das Konzept umfasst folglich auch finanzielle Grundkompetenzen („Financial Literacy“), wie sie im Rahmen zahlreicher englischsprachiger Studien (z. B. in Lusardi und Mitchell 2011 oder Bucher-Koenen et al. 2016) unter Erwachsenen erhoben wurden. Höchst umstritten ist dabei, inwieweit Financial Literacy (FL) eine eigene Kompetenzdimension darstellt (Schürkmann und Schuhen 2013; Retzmann und Frühauf 2014; Schuhen und Schürkmann 2014). Um internationale Vergleiche anstellen zu können und mit der FL verbundene Korrelate zu erforschen, verwenden OECD-Untersuchungen häufig nur drei Aufgaben aus einem Pool von 16. Die Gütekriterien für Leistungstests für dieses Instrument erscheinen zumindest fragwürdig. Zudem erfordert die Lösung dieser Items

auch mathematische Kompetenzen. Eine groß angelegte Messung ökonomischer Kompetenzen gemäß IÖK in Seeber et al. (2018) nahm diese drei Items mit „kanonischem Status“ (Kaiser und Lutter 2015) in den Itempool auf und identifizierte hohe Ladungen auf den gleichen Faktor wie die restlichen Items – ein weiteres Indiz dafür, dass finanzielle (Grund-)Bildung als Teilmenge von ökonomischer Bildung verstanden werden kann (siehe auch Seeber und Retzmann 2017). Bei Schülerinnen und Schülern wurde FL – ebenfalls unter Annahme eines eigenständigen Konstrukts – im Rahmen von PISA in den Jahren 2012 und 2015 miterhoben (OECD 2014a). Das Konzept nimmt die finanzielle Entscheidungsfähigkeit von Lernenden in den Fokus und lässt Ansätze eines elaborierteren Konstrukts im Sinne einer Verbesserung des finanziellen „Well-being“ für Individuum und Gesellschaft erkennen (Schürkmann und Schuhen 2013, 76). Bemängelt wird dabei, dass zumindest die wenigen veröffentlichten Aufgaben diesem Anspruch nicht genügen (vgl. Loerwald und Schnell 2016, 62). Das dort verwendete Konzept richtet sich nicht an deutschsprachigen Kompetenzdefinitionen aus und bildet zudem kein theoretisch fundiertes Globalkonstrukt ab.

2.4 Empirische Erhebungen im Bereich der ökonomischen Bildung

Die Erforschung des ökonomischen Verständnisses von Lernenden hat im angelsächsischen Raum sowohl in der Ökonomik als auch in der ökonomischen Bildung eine lange Tradition (Reviews finden sich u. a. in Allgood et al. 2015; Becker et al. 1990; Siegfried und Fels 1979).

Erste systematische Messungen ökonomischer Kenntnisse erfolgten bereits in den 1960er-Jahren durch den Test of Economic Understanding, der von Soper und Walstad (1987) in einen international einsetzbaren Test of Economic Literacy (TEL) transformiert und in der Folge weiter modifiziert wurde (Walstad und Rebeck 2001). Während sich viele Untersuchungen weltweit auf Messungen einer limitierten Finanical Literacy konzentrierten, bleibt die Anzahl von Untersuchungen außerhalb der Vereinigten Staaten vergleichsweise spärlich (Kaiser und Menkhoff 2017). Im deutschsprachigen Raum existiert mittlerweile eine Reihe von Testinstrumenten, die sowohl ökonomisches Wissen als auch Kompetenzen erheben:

Der in der Vergangenheit häufig eingesetzte wirtschaftskundliche Bildungstest (WBT; Beck und Krumm 1998), der den weltweit eingesetzten *Test of Economic Literacy* von Soper und Walstad (1987) adaptierte, ermöglichte durch seine Vielzahl von Anwendungen bundeslandübergreifende Vergleiche. Anwendung fand der WBT beispielsweise in der Studie von Würth und Klein (2001), die das Wirtschaftswissen anhand einer großen Stichprobe bei

Jugendlichen in Baden-Württemberg erhob. Weitere Anwendungen finden sich in Müller et al. (2007), die Lernende der Sekundarstufe I in Sachsen befragten oder in Bank und Retzmann (2012), die Lehrkräfte für eine Identifikation von Weiterbildungsbedarfen untersuchten. Das Instrument beinhaltet weitgehend lehrbuchbasierte Fragen (Seeber et al. 2018, 40) und legt dabei seinen Fokus auf volkswirtschaftliche Bereiche. Viele Fragen im WBT bilden gemäß IÖK lediglich den Inhaltsbereich „System und Ordnung" ab, das Instrument wird folglich dem in Seeber et al. (2012) formulierten Globalkonstrukt ökonomischer Kompetenz nicht gerecht. Weitere Studien erheben Wirtschaftswissen als zentrale Facette ökonomischer Kompetenzen von Lernenden der Sekundarstufe I in Österreich (Rumpold und Greimel-Fuhrmann 2016) und im deutschen Bundesland Niedersachsen (Loerwald und Schnell 2014) oder testen Teilaspekte ökonomischen Wissens (Verbraucherzentrale Bundesverband 2006). Alle Studien stellen dabei defizitäre Wissensbestände fest.

Ein Instrument, welches auf dem in 2.1 beschriebenen Ansatz der Schweizer Forschenden um Stephan Schumann fußt, wurde im Rahmen des OEKOMA-Projekts (Schumann et al. 2011) entwickelt. Anstatt die Aufgabeninhalte ex ante festzulegen, wurden diese aus gesichteten Zeitungsartikeln zweier führender Schweizer Tageszeitungen extrahiert. Die daraus gewonnenen 30.000 Begriffe wurden in das zuvor erarbeitete Kategoriensystem eingeordnet, woraus die Forschenden ein dreidimensionales Messmodell gemäß den drei Hauptkategorien (BWL, VWL und Rechnungswesen) erstellten. Die Mehrdimensionalität bestätigte sich empirisch. Die Autoren stellen ihr Konzept als ein dem wirtschaftskundlichen Bildungstest (WBT; Beck und Krumm 1998) ähnliches dar, wodurch der Bezug zum persönlichen Handeln fehlt (Seeber et al. 2018, 33).

Wertvolle Beiträge leisteten in diesem Rahmen auch die Siegener *ECOS*-Studien (Macha 2015; Macha und Schuhen 2013), die ihrem Instrument das Siegener Modell ökonomischer Bildung zugrunde legten. Für das eigens ausgearbeitete Messmodell wurden anhand einschlägiger Studien relevante Messdimensionen identifiziert und in einem Kompetenzoktagon zusammengefasst (Macha 2015, 49). Die komplexe Struktur konnte anhand einer Stichprobe von 580 Lernenden empirisch nachgewiesen werden.

Den bisher einzigen auf dem Integrationsmodell ökonomischer Kompetenz basierenden Test lieferten Seeber et al. (2018) im Rahmen einer Status-quo-Feststellung ökonomischer Kompetenz von Lernenden der 9. bis 11. Klassenstufe vor der Einführung des neuen Schulfaches

„Wirtschaft, Berufs- und Studienorientierung“ in Baden-Württemberg. Das 48 Items umfassende Testlet erlaubte durch sein adaptives Design eine hohe Messpräzision und konnte letztlich als eindimensional identifiziert werden. Es schien daher plausibel, auch in diesem Beitrag ein eindimensionales Konstrukt zu unterstellen.

Die beschriebenen Erhebungen in Deutschland beruhen bisher – mit Ausnahme von Seeber et al. (2018) – auf Convenience Samples, woraus sich neben der Entwicklung eines Instruments für die siebte Klassenstufe auf Basis des IÖK ein weiteres Forschungsdesiderat ableiten lässt.

2.5 Prädiktoren ökonomischer Kompetenz

Neben den Eigenschaften des Messinstruments interessiert insbesondere, welche Heterogenität hinsichtlich ökonomischer Kompetenz in Abwesenheit von Fachunterricht existiert. In Bezug auf Korrelate ökonomischer Kompetenz zeigen die meisten Untersuchungen Geschlechterunterschiede zugunsten der Männer (Kaiser und Lutter 2015; Davies et al. 2005; Heath 1989) – die Ursachen für diesen „Gender-Gap“ sind noch nicht eindeutig geklärt. Zudem zeigen sich negative Effekte für den Migrationsstatus (z. B. Schnell 2017; Erner et al. 2016; Würth und Klein 2001) oder für zuhause nicht deutsch sprechende Lernende (z. B. Seeber et al. 2018; OECD 2014a). Der sozioökonomische Status (OECD 2014b) sowie das Alter (Kotte und Lietz 1998) hängen dagegen häufig positiv mit ökonomischen Kenntnissen zusammen. Wesentliche Auskunft über die Heterogenität ökonomischer Kenntnisse gibt auch die Schulform: Sowohl in Studien auf Basis von Wissenstests (Müller et al. 2007; Beck und Krumm 1998) als auch in Kompetenzerhebungen (Seeber et al. 2018; Macha 2015) schnitten Lernende des Gymnasiums am besten ab. In beiden Kompetenzmessungen hatte in einer multiplen Regression die Zugehörigkeit zu einem Gymnasium den stärksten Gesamteffekt. Hinsichtlich Einstellungsdimensionen werden positive Zusammenhänge zwischen ökonomischen Kenntnissen und dem Interesse für Wirtschaftsthemen belegt (Schumann und Eberle 2014; Würth und Klein 2001), die empirische Evidenz insgesamt ist jedoch uneinheitlich (Seeber et al. 2018). Weitere Korrelate existieren hinsichtlich finanzieller Erfahrungen von Jugendlichen (Leiser und Ganin 1996), der innerfamiliären Kommunikation über Finanzen (OECD 2014a), der kognitiven Grundfähigkeit (Macha 2015; Nickolaus et al. 2008) und Leistungsdispositionen in den Fächern Deutsch und Mathematik (Beck und Krumm 1998).

3 Itementwicklung

Die Itementwicklung orientierte sich am Vorgehen einer vorangegangenen Querschnittserhebung in Baden-Württemberg (Seeber et al. 2018). Da die Items dort für Lernende der Klassenstufe 9 bis 11 ausgerichtet waren, wurde für die hier relevante Testung in Klassenstufe 7 lediglich eine Teilmenge aus dem unteren Teil der Schwierigkeitsskala übernommen. Ferner war im Itempool der erste Kompetenzbereich („Entscheidung und Rationalität“) leicht unterrepräsentiert, was eine Entwicklung zusätzlicher Items sinnvoll erscheinen ließ. Das Vorgehen gleicht sich in beiden Studien: Ökonomisches Wissen und Denken wurde anhand der Matrix des IÖK (Abbildung 1) operationalisiert. Die Aufgaben sollen demnach „Handlungsfolgen einer ökonomischen Entscheidung analysieren und/oder bewerten (Entscheidung und Rationalität), Kooperationen analysieren und/oder bewerten (Beziehung und Interaktion) sowie Politik ökonomisch beurteilen (System und Ordnung)“ (Seeber et al. 2018, 65). Da der Itempool inhaltlich relevante Kontexte abbilden soll, wurden – in Abwesenheit eines deutschen Kerncurriculums – Lehrpläne aus drei Bundesländern für die Sekundarstufe I sowie Vorschläge für Bildungsstandards auf ihre Inhalte und Anforderungen untersucht. Die Inhalte wurden anschließend den Kompetenzfeldern in Abbildung 1 zugeordnet und dienten letztlich der Generierung von Items (ebd.).

Die inhaltliche Validierung gliederte sich in zwei qualitative Auswertungsschritte. Für die Überprüfung der Altersadäquanz wurde im ersten Schritt exemplarisch eine qualitative *Think-Aloud*-Studie (Schnell 2016) mit drei Lernenden der siebten Klasse an einer Gesamtschule durchgeführt. Dabei konnte die Verständlichkeit der Fragen für diese Zielgruppe weitgehend bestätigt werden, jedoch mussten vereinzelt Begriffe an die niedrigere Altersgruppe angepasst werden. Nach dieser ersten Überarbeitung wurden die Items einer Expertenvalidierung (Bernd Remmele und Franziska Birke, PH Freiburg) unterzogen. Der Schwierigkeitsgrad der Items wurde leicht nach oben korrigiert und einzelne Distraktoren wurden durch plausiblere Alternativen ersetzt. Um weiteren Ausschlüssen durch psychometrische Tests in der Pilotierung vorzubeugen, wurde für den ersten Inhaltsbereich ein leichtes Übergewicht erzeugt. Die Pilotierungsstudie mit insgesamt 40 Items wurde an 20 Schulen mit 355 Lernenden (*Convenience Sample*) durchgeführt und führte zu weiteren acht Itemausschlüssen. Das finale Itemset zeigte eine weitgehend gleichmäßige Verteilung über die Kompetenzbereiche und umfasste 27 offene sowie fünf Multiple-Choice-Fragen. Tabelle 1 zeigt die Itemverteilung nach Kompetenzbereichen und Rollen.

	Entscheidung & Rationalität	Beziehung & Interaktion	System & Ordnung	Summe
Konsument	3	3	4	**10**
Erwerbstätiger	6	6	6	**18**
Bürger	1	2	2	**5**
Summe	**9**	**11**	**12**	

Tabelle 1: Anzahl der Items gemäß Rollen und Kompetenzbereichen

Aufgrund guter psychometrischer Eigenschaften wurden zwei Items aus externen Quellen aufgenommen (Tabelle 2: Lusardi und Mitchell 2014; OECD INFE 2012). Abbildung 2 zeigt beispielhaft ein Item aus dem Kompetenzbereich „Beziehung und Interaktion".

Familie Marone betreibt die Eisdiele Fantasia in der Innenstadt. Dieses Jahr erhöht sie die Preise pro Kugel von 1,00 € auf 1,20 €.

Welche Auswirkungen hat die Erhöhung auf den Umsatz der Eisdiele?

☐ *Der Umsatz wird durch die Preiserhöhung zurückgehen.*

☐ *Der Umsatz wird sich durch die höheren Preise ebenfalls erhöhen.*

☐ *Der Umsatz wird trotz Preiserhöhung gleichbleiben.*

☐ *Der Umsatz hängt davon ab, wie die Kundschaft reagiert.*

Abbildung 2: Item Nr. 9

4 Validität und Reliabilität

Die Validität eines Tests beschreibt grundsätzlich das „Ausmaß, in dem der Test das misst, was er behauptet zu messen" (Brown 1996) und kann in ihre drei Hauptzugänge untergliedert werden: Inhalts-, Konstrukt- und Kriteriumsvalidität. Sie ist jedoch nicht unmittelbar aus einem Testlet ableitbar, sondern verlangt eine Verknüpfung von inhaltlicher Logik und empirischer Analyse (Cronbach 1976). Aufgrund der bereits seit Jahrzehnten andauernden Diskussion über eine angemessene Evaluation von Validität (Wainer und Braun 2013; Shepard 1993; Messick 1980; Cronbach 1946), kann im Rahmen dieses Beitrags im besten Fall eine Annäherung an dieses Gütekriterium geleistet werden.

Die inhaltliche Logik der Validität kommt im Zugang der Inhaltsvalidität zum Tragen und beschreibt, inwieweit der behandelte Test eine Stichprobe aus einem hypothetischen Universum aller Items darstellt, die dasselbe Konstrukt repräsentieren – sie verkörpert somit

eine erste Approximation an eine Validitätsprüfung. Trotz mehrerer Versuche, eine Prüfung der Inhaltsvalidität zu quantifizieren (Lawshe 1975; Rubio et al. 2003), stützt sich diese in der Regel auf fachliche Überlegungen (Haynes et al. 1995). In diesem Beitrag erfolgt eine Annäherung über eine Expertenvalidierung (Kapitel 3).

Die Konstruktvalidität drückt hingegen aus, inwieweit die Antwortcharakteristika des Tests durch das zugrunde liegende Konstrukt zusammengefasst werden (Anastasi und Urbina 1997, 197). Loerwald und Schnell (2016, 61) sehen in diesem Zugang „das bedeutendste und zugleich komplexeste Gütekriterium“. Empirisch wird dieser Zugang u. a. mit faktoranalytischen Methoden (Thompson und Daniel 1996; Guilford 1946), aber auch mit Methoden der Item Response Theory (Henning 1992; Hambleton und Rovinelli 1986) untersucht. Letztere erlauben zudem die (stichprobenunabhängige) Extraktion von Schwierigkeits- und Diskriminationsparametern für eine Itemanalyse. Ist die Einschränkung gegeben, dass dieser Beitrag keine umfassende Konstruktvalidität gemäß dem in Cronbach und Meehl (1955) postulierten nomologischen Netz nachweisen kann, erfolgt hier eine Annäherung über den Nachweis der Dimensionalität mithilfe von Faktoranalysen sowie im Rahmen der Item Response Theory über die Prüfung der Modellpassung und des Differential Item Functioning (DIF) unter Subgruppen. Der Nachweis der Unidimensionalität der Antwortcharakteristika stellt dabei durch die Extraktion möglichst weniger Linearkombinationen sicher, dass die Testitems kein weiteres zugrunde liegendes Konstrukt messen (Green 2013). Da IRT-Modelle ein (oder mehrere) zugrundeliegende Konstrukt(e) unterstellen, kann die Dimensionalität ebenso mithilfe einer Prüfung der Modellpassung nachgewiesen werden (z. B. Stewart-Brown et al. 2009). Die Analyse von Antwortmustern (anstatt Korrelationsmatrizen) oder auch die Annahme eines logistischen Zusammenhangs zwischen Testleistung und latenter Fähigkeit werden als Vorteile gegenüber faktoranalytischen Methoden angesehen (Li et al. 2012). Komplementär wird mithilfe eines multidimensionalen IRT-Modells geprüft, inwieweit die im Integrationsmodell ökonomischer Kompetenz ausgewiesenen Kompetenzbereiche eigene Dimensionen darstellen. Mitentscheidend für die Konstruktvalidität ist auch die Test- und Itemfairness unter Subgruppen, die in älteren Beiträgen unter dem Rubrum *item bias* (Mellenbergh 1989) diskutiert wurde. Grundannahme ist dabei, dass ein Item von Lernenden mit gleichem Fähigkeitsniveau aufgrund ihrer Subgruppenzugehörigkeit (z. B. Geschlecht oder Migrationshintergrund etc.) unterschiedlich beantwortet wird, somit zusätzlich ein außenstehendes Konstrukt gemessen wird und folglich die aus dem Test gezogenen

Schlussfolgerungen verzerrt sind (Salehi und Tayebi 2012, 86). Statistisch kann ein solcher Bias im Rahmen der Item Response Theory durch das Differential Item Functioning (DIF) festgestellt werden (Holland und Thayer 2013), welches auch in Messungen von ökonomischen Fähigkeiten Anwendung findet (Seeber et al. 2018; Walstad und Robson 1997).

Die Kriteriumsvalidität bezieht sich auf korrelative Zusammenhänge zwischen dem gemessenen Konstrukt und unabhängigen Außenkriterien (Cronbach und Meehl 1955, 282). Da im Rahmen des Beitrags keine longitudinalen Außenkriterien erhoben wurden, kann lediglich die konkurrente Validität betrachtet werden. Dabei wird mithilfe einer multiplen Regression geprüft, inwieweit sich erwartbare Subgruppenunterschiede in den Testergebnissen hinsichtlich demografischer Aspekte oder der Schulformzugehörigkeit ergeben. Die Reliabilität der Kompetenzskala als nicht hinreichende, aber notwendige Bedingung für Skalenvalidität (Clark und Watson 1995, 314) wird durch den Nachweis der inneren Konsistenz mithilfe von Maßen aus der Klassischen Testtheorie (KTT) geprüft.

5 Daten

Die folgenden Unterkapitel beschreiben das Stichprobenverfahren und geben anschließend einen Überblick über deskriptive Statistiken der erhobenen Variablen. Zuletzt wird der Umgang mit fehlenden Daten skizziert.

5.1 Stichprobe

Die Grundgesamtheit umfasst alle Lernenden der öffentlichen Schulen in Baden-Württemberg der siebten Klassenstufe im Schulahrjahr 2016/2017. Im Zuge des Stichprobenverfahrens wurde die Grundgesamtheit anhand der Variablen Schulform und Urbanisierungsgrad in Teilpopulationen aufgeteilt (explizite Stratifizierung). Der Urbanisierungsgrad umfasst drei Ausprägungen (hoch, mittel und niedrig) und wurde mithilfe der Bevölkerungsdichte den jeweiligen Landkreisen zugeteilt. Eine Dichte über 485 EW/km^2 wurde dabei als hoch, zwischen 221 und 485 als mittel und unter 220 als niedrig klassifiziert. Die drei Urbanisierungsgrade bildeten zusammen mit den vier Schularten (Gymnasium, Realschule, Gemeinschaftsschule und Werkrealschule) zwölf Strata.

Da lediglich ein Datensatz baden-württembergischer Schulen und deren Klassenanzahl vorliegt, wurde ein zweistufiges Ziehungsverfahren angewendet, in dem zunächst die Schule und danach per Zufallsauswahl eine achte Klasse dieser Schule gezogen wurde. Die Anzahl

der zu ziehenden Schulen in jedem Stratum wurde proportional zur Grundgesamtheit angepasst (*probability proportional to size*). Um zu vermeiden, dass Schulen einer bestimmten Größe überproportional in der Ziehung enthalten sind, wurde neben den expliziten Stratifizierungsvariablen die Schulgröße als implizierte Stratifizierungsvariable berücksichtigt. Dazu wurden sie innerhalb eines Stratums der Größe nach geordnet und eine systematische Ziehung mithilfe von Samplingintervallen durchgeführt (z. B. Lohr 2010). Um die eingeschränkte Repräsentativität in den einzelnen Strata auszugleichen, erfolgte eine nachträgliche Proportionalisierung mithilfe einer Design-Gewichtung. Die Design-Gewichte werden mittels Inverse der Ziehwahrscheinlichkeit $p^{(j,k)}$ des einzelnen Lernenden gebildet und anschließend auf die Anzahl der Teilnehmenden mit einem Mittelwert von 1 normalisiert. Die normalisierten Gewichte ergeben sich aus $w_{NORM}^{(j,k)} = w^{(j,k)} \times \frac{1698}{\sum w^{(j,k)}}$ mit $w^{(j,k)} = 1 / p^{(j,k)}$.

Die finale Ziehung beinhaltete 1.689 Lernende in 86 Klassen. Tabelle 1 zeigt ausgewählte demografische Charakteristika der Stichprobe. Das Durchschnittsalter der Lernenden lag bei 13,88 Jahren (SD = 0.74). Die DIF-Analyse wurde anhand des Geschlechts, des Migrationsstatus, der Muttersprache sowie des Bildungshintergrundes der Eltern durchgeführt.

Tabelle 2 zeigt deskriptive Statistiken aller verwendeten Variablen. Die abhängige Variable WLE repräsentiert die geschätzte Personenfähigkeit aus dem einparametrischen Rasch-Modell (Kapitel 5) und WLE500 seine Transformation auf einen Mittelwert von 500 mit einer Standardabweichung von 100, wie es in internationalen Schulleistungsstudien wie PISA üblich ist – die transformierte Variable wird auch in der Regressionsanalyse verwendet. Bei der Variable Geschlecht repräsentiert 0 weibliche und 1 männliche Lernende – 55,1 Prozent der Stichprobe sind männlich. Lernende besitzen einen Migrationshintergrund, wenn mindestens einer der beiden Elternteile im Ausland geboren ist (41,1 Prozent).

Der sozioökonomische Status wurde mithilfe der Bücherfrage (*Bücher zu Hause*) approximiert. Auf die Frage, wie viele Bücher im Haushalt der Lernenden vorhanden sind (ohne Schulbücher und Zeitschriften), konnten diese auf einer Skala von 1 (keine) bis 6 (mehrere Regalwände) antworten. Um die Schulleistungen der Lernenden zu kontrollieren, wurden diese nach ihren Lesefähigkeiten, Rechenfähigkeiten sowie nach ihren allgemeinen Schulleistungen (*Fähigkeiten insgesamt*) auf einer Skala von 1 (niedrig) bis 5 (hoch) befragt. Zudem wurde das allgemeine Interesse für Wirtschaftsthemen (*interesteco*) sowie die Frage, ob sie Wirtschaftswissen für wichtig erachten, auf einer Skala von 1 (stimme nicht zu) bis 4 (stimme völlig zu) miterhoben.

	N	Mittelwert	Standardabw.	Min	Max
WLE.500	1.687	500	100	87,308	884,689
WLE	1.687	0,095	0,984	-3,967	3,882
Geschlecht	1.687	0,551		0	1
Alter	1.561	13,88	0,740	12	19
Migrationsstatus	1.561	0,411		0	1
Bücher zu Hause	1.545	3,442	1,619	1	6
Lesefähigkeit	1.687	3,864	0,712	1	5
Rechenfähigkeit	1.681	3,512	0,850	1	5
Fähigkeiten gesamt	1.677	3,663	0,638	1	5
interesteco	1.659	2,525	0,758	1	4
interestimp	1.653	2,970	0,752	1	4
GYM	1.687	0,391		0	1
RS	1.687	0,303		0	1
WRS	1.687	0,145		0	1
GMS	1.687	0,161		0	1

Tabelle 2: Deskriptive Statistiken

5.2 Fehlende Werte

Aufgrund der freiwilligen Teilnahme am Wirtschaftskompetenztest können bei der Analyse und Behandlung von fehlenden Werten lediglich Antwortausfälle im Test selbst betrachtet werden (*„item non-response“*). Weisen Lernende bei über 50 Prozent der Items fehlende Werte auf, wurden sie von weiteren Analysen ausgeschlossen.

Obwohl beide Ansätze mit Problemen verbunden sind, werden fehlende Werte in Kompetenzmessungen häufig entweder als fehlend („*NA*“) oder als falsch („0“) kodiert. Viele Large-Scale Assessments setzen daher gemischte Verfahren ein: Die NAEP-Studie (National Assessment of Educational Progress) kodiert Antwortausfälle bei Multiple-Choice-Items als fehlend, während Antwortausfälle in offenen Items als falsch klassifiziert werden. Gemischte Verfahren können sich ebenso an den verschiedenen Untersuchungsphasen orientieren. In TIMSS und PISA werden Antwortausfälle während der Test- und Itemvalidierung als fehlend klassifiziert, während die Ausfälle bei Populations- und Regressionsberechnungen als falsch kodiert werden (vgl. Pohl et al. 2014, 1).

Die vorliegende Analyse verwendet ebenfalls ein gemischtes Verfahren, das sich jedoch an der Antwortdauer orientiert. Weist eine fehlende Antwort eine Dauer von mehr als drei Sekunden auf, wurde sie als falsch bzw. zu 0 umkodiert. In diesen Fällen kann davon ausgegangen werden, dass das Item zumindest gelesen und nicht geraten wurde. Alle restlichen Antwortausfälle blieben weiterhin als fehlend kodiert.

Eine Umcodierung von fehlenden Werten in falsche Antworten liefert in der Regel positiv verzerrte Itemwerte, während der gegenteilige Ansatz möglicherweise zu weiteren Itemausschlüssen führt. Simulationsstudien zeigten jedoch, dass die Umwandlung von Antwortausfällen in falsche Antworten zu schwerwiegenderen Verzerrungen bei Parameterschätzungen führt (z. B. Rose et al. 2010).

6 Psychometrische Eigenschaften des Tests

Dieser Abschnitt beschreibt die angewandte Methodik sowie deren Ergebnisse für die psychometrische Validierung des Tests. Diese beruht auf dem 1-PL-Modell (Rasch-Modell) sowie auf mehrparametrischen Modellen, auf deren Basis in den weiteren Schritten verschiedene Tests zur Beurteilung der Modellpassung durchgeführt werden. Faktoranalytische Methoden sind durch die folgende IRT-Analyse im Grunde obsolet, da mithilfe der Modellfitwerte bzw. über den Nachweis, dass die Daten auf das Rasch-Modell passen, die Eindimensionalität bereits nachgewiesen werden kann. Deshalb wird die Dimensionalität faktoranalytisch lediglich anhand einer Hauptkomponentenanalyse veranschaulicht. Vorrangiges Ziel weiterer Analysen ist es, jene Items zu extrahieren, die sowohl eine psychometrisch valide als auch inhaltskonforme Erfassung von ökonomischer Kompetenz ermöglichen.

6.1 Dimensionalität

Eine Hauptbedingung für die Durchführung der nachfolgenden Analyse auf Basis der Item Response Theory ist die Unidimensionalität des zu messenden Konstrukts, sodass alle Items die gleiche latente Fähigkeit messen. Die Reduktion der Dimensionen erfolgte mithilfe einer Hauptkomponentenanalyse, die das Prinzip der Varianzmaximierung verwendet (z. B. Abdi und Williams 2010). Da die Hauptkomponentenanalyse für kontinuierliche Variablen angelegt ist, wurde die Extraktion anhand der tetrachorischen Korrelationsmatrix vollzogen. Sowohl der Kaiser-Meyer-Olkin-Koeffizient (Kaiser und Rice 1974) als auch der Bartlett‘s Test auf Sphärizität (Bartlett 1951) zeigten die Eignung der Daten für faktoranalytische Methoden. Das Ergebnis zeigt für den ersten Faktor einen Eigenwert von 10.42, welcher 22.17 Prozent der Gesamtvarianz erklärt. Der zweite Faktor zeigt lediglich einen Eigenwert von 2.45 und bezeugt die Dominanz des ersten Faktors.

Grafisch kann die Dimensionalität mithilfe eines Screetests (Cattell 1966) veranschaulicht werden, der hier zusätzlich um eine Parallelanalyse ergänzt wird. Dabei werden die Eigenwerte, die die Höhe der erklärten Varianz beschreiben, auf der Ordinate und die

Faktoren bzw. Komponenten auf der Abszisse abgetragen. An der Knickstelle, d. h. an der Stelle, an der die Varianz Richtung Nullpunkt konvergiert, lässt sich die Dimensionalität einsehen. Im Rahmen der Parallelanalyse werden über 100 dem Analysedatensatz ähnelnde Zufallsdatensätze gebildet und einer Faktorenanalyse unterzogen. Von allen gewonnenen Eigenwerten wird der Mittelwert gebildet und mit den empirischen Daten verglichen. Es werden nur Faktoren als relevant angesehen, die einen höheren Eigenwert als jene aus dem Paralleltest aufweisen. Abbildung 3 zeigt die Unidimensionalität des Globalkonstrukts „Ökonomische Kompetenz“. Unter jenen Komponenten, die eine hohe Varianzaufklärung aufweisen, besitzt lediglich der erste Faktor einen höheren Eigenwert als die simulierten Daten aus dem Paralleltest.

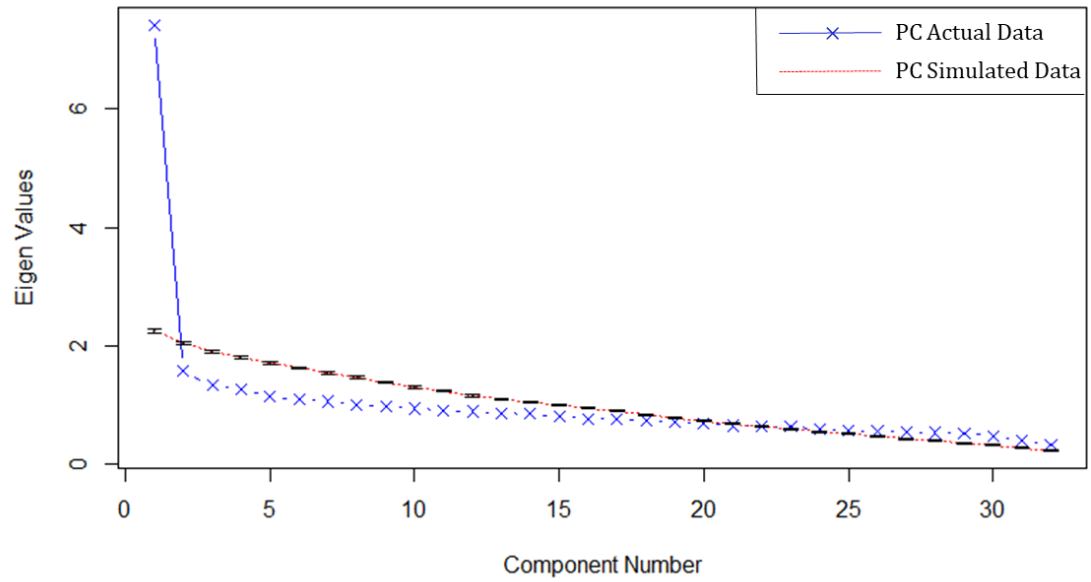

Abbildung 3: Hauptkomponenten- und Parallelanalyse

6.2 Testtheoretische Analysen

Da sich die Auswahl und Entwicklung der Items an der vorangegangenen WIKO-BW-Studie orientierte, kommt im ersten Schritt ein einparametrisches probabilistisches Testmodell („dichotomes Rasch-Modell“) zur Anwendung (Rasch 1960). Mehrparametrische IRT-Modelle werden in Abschnitt 6.2.4 diskutiert. Vorteile von Modellen der Item Response Theory für die Kompetenzdiagnostik werden in der Messung von Aufgabenschwierigkeit und Personenfähigkeit auf einer gemeinsamen Skala oder in der Vergleichbarkeit von Ergebnissen aus verschiedenen Studien (Stichprobenunabhängigkeit) gesehen (vgl. Hartig und Frey 2013, 49).

Das Rasch-Modell folgt einer logistischen Modellgleichung, die für dichotome Antwortformate Wahrscheinlichkeitsaussagen trifft. Die Modellgleichung für die richtige Lösung eines Items ist in Gleichung (1) dargestellt. θ_v beschreibt dabei die Fähigkeit von Person v und σ_i den Schwierigkeitsgrad von Item i.

$$P\left(\mathrm{X}_j = 1 \mid \theta_v, \sigma_i\right) = \frac{\exp(\theta_v - \sigma_i)}{1 + \exp(\theta_v - \sigma_i)} \quad (1)$$

Die geschätzten Parameter θ_v und σ_i werden auf einer gemeinsamen Logitskala abgebildet, sie können kontinuierliche Werte annehmen und bewegen sich in der Regel im Intervall [-4; 4]. Die Schätzung der Parameter erfolgt grundsätzlich durch Maximierung der Likelihoodfunktion (*Maximum Likelihood Estimation*), die sich aus der Multiplikation aller sich aus dem Rasch-Modell errechneten und nun in der Datenmatrix befindenden Wahrscheinlichkeiten bildet:

$$\max_{\theta,\sigma} \quad L = \prod_{v=1}^{N} \prod_{i=1}^{M} \frac{\exp\left(x_{vi}(\theta_v - \sigma_i)\right)}{1 + \exp(\theta_v - \sigma_i)} \quad (2)$$

Der MLE-Schätzer für die Personenfähigkeit (*Maximum Likelihood Estimation*) wird durch den von Warm (1989) eingeführten WLE-Schätzer *(Weighted Likelihood Estimation*) ersetzt, da dieser die Messwerte mit der individuellen Iteminformation gewichtet und somit eine präzisere Schätzung ermöglicht. Die Itemkalibrierung erfolgte mit der in *Educational Large Scale Assessments* geläufigen MMLE-Methode (*Marginal Maximum Likelihood Estimation*; z. B. Bock und Aitkin 1981). Für die Berechnung der Itemschwierigkeit und Personenfähigkeit sind die Rohwerte im Rasch-Modell eine suffiziente Statistik. Für die Itemanalyse sind neben den Schwierigkeitsparametern die Modellpassung und die Trennschärfe von entscheidender Bedeutung.

6.2.1 Modellpassung

Die Modellpassung für das 1-PL-Modell wurde anhand der FIT-Statistiken überprüft. Diese geben an, inwieweit die geschätzten Daten in die theoretische Modellvorhersage passen (Eckes 2005; Wright und Masters 1982). Dabei werden der Weighted-Mean-Square (INFIT) und der Unweighted Fit (OUTFIT) verwendet (vgl. Wright und Linacre 1994). Infitwerte berechnen sich, indem die quadratischen standardisierten Residuen mithilfe ihrer individuellen Varianzen gewichtet werden. Outfitwerte ergeben sich aus der standardisierten Summe der quadratischen Residuen über alle Beobachtungen hinweg. Fitwerte von 1,0 bedeuten eine perfekte Passung der beobachteten Daten in das theoretische Modell. Um eine angemessene

Validität zu gewährleisten, wurden kritische Grenzwerte bei 0.5 bzw. 0.6 und 1.5 festgelegt (z. B. Ames und Penfield 2015; Ayala 2009b).

6.2.2 Trennschärfe und interne Konsistenz

Da im Wirtschaftskompetenztest sowohl offene als auch geschlossene Antwortformate dichotomisiert wurden, wird die punktbiseriale Korrelation aus der klassischen Testtheorie als Trennschärfemaß verwendet (vgl. Guilford 1954). Die Trennschärfe misst den korrelativen Zusammenhang zwischen Itemantwort und dem Gesamttestwert. Ein Item ist somit genau dann trennscharf, wenn leistungsstarke Lernende das Item lösen und leistungsschwache Lernende eine falsche Antwort geben. Ein Koeffizient nahe null würde bedeuten, dass für es für die Lösungshäufigkeit irrelevant ist, ob Lernende Wirtschaftskompetenzen aufweisen oder nicht. Eine negative Trennschärfe deutet hingegen auf eine Fehlkonstruktion des Items hin, da gerade leistungsstarke Lernende dazu verleitet werden, die falsche Antwort zu wählen. Bei der Messung von Wirtschaftskompetenzen wird der Schwellenwert bei $r < 0.2$ (z. B. Itzlinger-Bruneforth et al. 2016; Walstad und Rebeck 2001) angesetzt, jedoch müssen durch Zusammenhänge zwischen Trennschärfe und Itemschwierigkeit Deckel- und Bodeneffekte berücksichtigt werden (vgl. Schelten 1997, 134). Um die Trennschärfekoeffizienten nicht zu überschätzen, wird durch Ausschluss des betrachteten Items vom Gesamttestwert das korrigierte Trennschärfemaß berechnet (vgl. Howard und Forehand 1962). Eine durch die Trennschärfe ausgedrückte hohe Korrelation zwischen Itemwert und Gesamttestwert deutet zudem auf eine hohe interne Konsistenz der Testskala hin (Streiner 2003). Cronbachs Alpha (Cronbach 1951) zeigte zudem einen hohen Wert ($\alpha = 0.81$).

6.2.3 Ergebnisse

Tabelle 3 zeigt die nach Schwierigkeit geordneten Ergebnisse aus dem Rasch-Modell und der klassischen Testtheorie. *RelFreq* beschreibt dabei die Lösungshäufigkeit, *rit_c* den korrigierten Trennschärfekoeffzienten und *sigma* den Schwierigkeitsparameter. Die Schätzungen wurden mithilfe des R-Pakets *TAM* (Robitzsch et al. 2018) errechnet. Die *EAP*-Reliabilität beträgt 0*.78.* Hinsichtlich relativer Lösungshäufigkeit sollten Items nicht von mehr als 95 Prozent und nicht von weniger als 5 Prozent der Lernenden gelöst werden (Itzlinger-Bruneforth et al. 2016) (Tabelle 3: Ergebnisse KTT und IRT [1-PL-Modell]).

			KTT		IRT			
Item	n	Quelle	RelFreq	rit_c	sigma	[SE]	Infit	Outfit
1	1689		0,738	0,258	-1,103	[0.058]	1,006	0,987
2	1619		0,634	0,164	-0,581	[0.055]	1,064	1,116
3	1556	Lusardi/Mitchell (2014)	0,638	0,223	-0,544	[0.056]	1,056	1,04
4	1554	Seeber et al. (2018)	0,641	0,257	-0,536	[0.056]	1,024	1,028
5	1560		0,624	0,345	-0,465	[0.055]	0,97	0,945
6	1636	Seeber et al. (2018)	0,598	0,039	-0,428	[0.054]	1,16	1,266
7	1550		0,611	0,383	-0,377	[0.055]	0,971	0,943
8	1587	Seeber et al. (2018)	0,595	0,355	-0,338	[0.054]	0,978	0,965
9	1565		0,591	0,379	-0,306	[0.055]	0,941	0,917
10	1584		0,569	0,397	-0,171	[0.054]	0,943	0,92
11	1558		0,562	0,405	-0,144	[0.054]	0,939	0,922
12	1566	Seeber et al. (2018)	0,557	0,365	-0,137	[0.054]	0,971	0,952
13	1620		0,55	0,376	-0,114	[0.053]	0,972	0,954
14	1570		0,52	0,313	-0,006	[0.054]	0,994	0,982
15	1613	Seeber et al. (2018)	0,534	0,409	0,009	[0.053]	0,921	0,902
16	1614		0,524	0,383	0,017	[0.053]	0,938	0,918
17	1628	OECD 2012	0,518	0,371	0,022	[0.053]	0,96	0,946
18	1602	Seeber et al. (2018)	0,49	0,337	0,171	[0.054]	0,985	0,975
19	1606	Seeber et al. (2018)	0,366	0,174	0,723	[0.056]	1,082	1,117
20	1629	Seeber et al. (2018)	0,338	0,411	0,899	[0.057]	0,926	0,887
21	1571		0,327	0,399	0,981	[0.059]	0,931	0,886
22	1619	Seeber et al. (2018)	0,289	0,371	1,121	[0.059]	0,946	0,918
23	1570		0,278	0,009	1,136	[0.06]	1,181	1,323
24	1623	Seeber et al. (2018)	0,29	0,315	1,161	[0.06]	0,985	0,986
25	1628	Seeber et al. (2018)	0,284	0,207	1,181	[0.06]	1,056	1,104
26	1629		0,257	0,37	1,313	[0.061]	0,945	0,898
27	1609	Seeber et al. (2018)	0,251	0,101	1,326	[0.062]	1,107	1,198
28	1618	Seeber et al. (2018)	0,224	0,174	1,496	[0.064]	1,068	1,201
29	1614	Seeber et al. (2018)	0,206	0,344	1,648	[0.067]	0,956	0,912
30	1628	Seeber et al. (2018)	0,176	0,212	1,799	[0.069]	1,03	1,087
31	1611		0,144	0,349	2,214	[0.079]	0,939	0,788
32	1634	Seeber et al. (2018)	0,102	0,132	2,456	[0.085]	1,036	1,348

Tabelle 3: Itemparameter KTT und IRT

Die Lösungshäufigkeit beträgt im Mittel 0.449 (SD = 0.179) und besitzt eine Reichweite von 0.102 bis 0.776, sodass kein Item nach diesem Kriterium ausschlusswürdig erscheint. Die errechneten Trennschärfekoeffizienten erreichen im Mittel einen Wert von .292 mit einer Spannweite von 0.007 bis 0.411. Die Items 6 und 23 weisen Trennschärfen nahe null auf und sind daher für die Messung von Wirtschaftskompetenzen ungeeignet. Es ist somit für die Lösung der Items nicht relevant, ob Lernende eine hohe oder eine niedrige Fähigkeit besitzen,

und deutet daher auf eine fehlerhafte Konstruktion der Items hin. Item 32 weist mit einer Lösungshäufigkeit von nur 10,2 % und einer Trennschärfe von 0.13 einen typischen Bodeneffekt auf – das Item zeigte jedoch in der vorangegangenen Pilotstudie mit einem *Convenience Sample* eine Trennschärfe von 0.31. Es ist daher anzunehmen, dass bei Kompetenzmessungen in höheren Klassenstufen die Lösungshäufigkeit und somit auch der Trennschärfekoeffizient steigt. Die errechneten Infitwerte besitzen im Mittel einen Wert von 0.99 mit einer Spannweite von 0.92 und 1.18 und bewegen sich somit im angesetzten Intervall zwischen 0.6 und 1.5. Die gegenüber Ausreißern sensitiveren Outfit-Werte erreichen im Mittel einen Wert von 1.01 (SD = 0.14) und erstrecken sich von 0.79 bis 1.35.

6.2.4 Modellgüte und Differential Item Functioning (DIF)

Die lokale stochastische Unabhängigkeit der Itemantworten, die für die Berechnung der WLE-Schätzer vorausgesetzt wird, kann mithilfe der Q3-Statistiken überprüft werden (vgl. Yen 1984). Demnach sollten die Residuen nur aufgrund der zugrunde liegenden latenten Fähigkeit positiv korrelieren bzw. sollten die Korrelationen verschwinden, nachdem die Personenfähigkeit konstant gehalten wird. Die Ergebnisse zeigten, dass sich die Korrelationen im Bereich von -0.31 und 0.07 bewegen – die gemittelte Korrelation beträgt -0.03. Für die Berechnungen wurde das R-Paket *sirt* (Robitzsch 2018) verwendet.

Ferner kann die Passung der Daten für mehrparametrische Modelle geprüft werden. Der allgemeine Fall eines mehrparametrischen Testmodells mit vier Parametern (Barton und Lord 1981; Magis 2013) folgt der Gleichung

$$P\left(\mathrm{X}_j = 1 \mid \theta_v, \sigma_i, \alpha_i, \gamma_i, \delta_i\right) = \gamma_i + (\delta_i - \gamma_i)\,\frac{\exp\left[\alpha_i(\theta_v - \sigma_i)\right]}{1 + \exp(\theta_v - \sigma_i)} \quad (3).$$

Dabei repräsentiert α_i den Diskriminationsparameter und äußert sich grafisch in einer variablen Steigung der *Item Characteristic Curve (ICC).* γ_i repräsentiert die untere Asymptote der *ICC* und unterstellt damit auch Lernenden mit niedriger Fähigkeit eine gewisse Wahrscheinlichkeit für die Lösung eines Items („Rateparameter"), während δ_i die Unaufmerksamkeit bzw. Leichtsinn in Form einer oberen Asymptote kleiner eins modelliert – Lernende mit hoher Fähigkeit haben eine reduzierte Wahrscheinlichkeit für die korrekte Lösung eines Items. Sowohl das 2-PL-Modell (δ_i = 1 und $\gamma_i = 0$) als auch das 3-PL-Modell (δ_i = 1; beide Birnbaum 1968) lassen sich aus Gleichung 4 ableiten. Abbildung 4 zeigt den unterschiedlichen Verlauf der *Item Characteristic Curves* für Item 11.

Der geschätzte Rateparameter im 4-PL-Modell γ_i beträgt 0.16, während die obere Asymptote δ_i bei 0.93 liegt. Die Ergebnisse für alle vier Modelle sind im Anhang in Tabelle A1 angehängt. Die Berechnungen erfolgten mithilfe des R-Pakets *mirt* (Chalmers 2012).

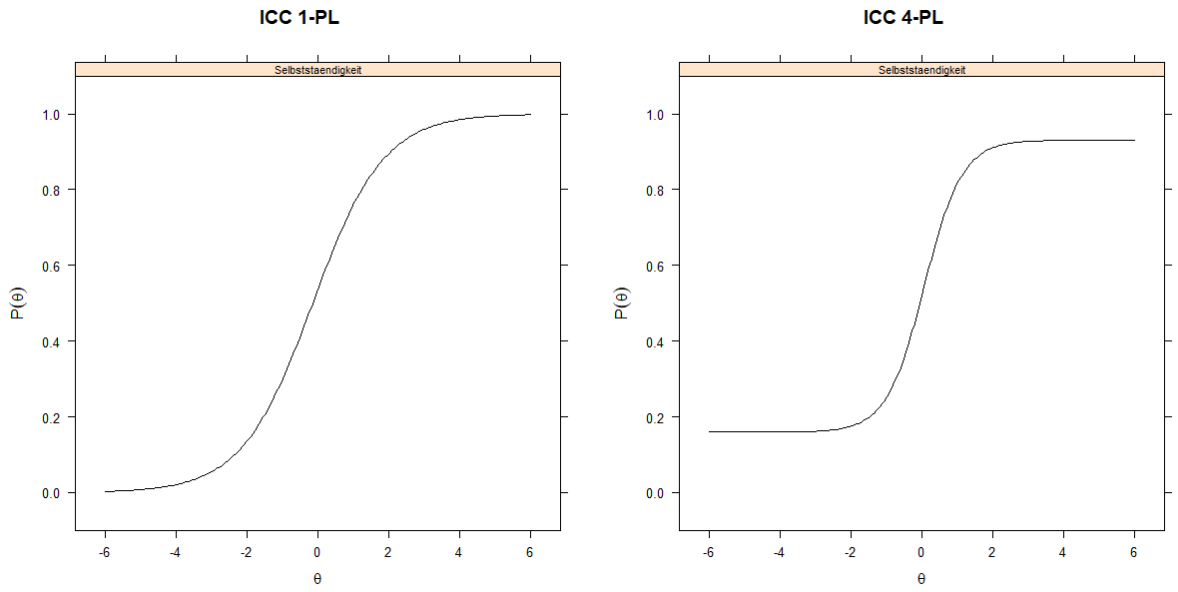

Abbildung 4: ICC für das 1-PL- und 4-PL-Modell

Für einen Vergleich der Modelle wird im ersten Schritt anhand der informationstheoretischen Maße *BIC* und *AIC* überprüft, welches Modell die Daten besser erklärt (Tabelle 4).

Modell	AIC	HQ	BIC	logLik
1-PL	60096,57	60162,95	60275,83	-30015,29
2-PL	59401,09	59529,83	59748,73	-29636,55
3-PL	59300,99	59494,1	59822,45	-29554,49
4-PL	59292,27	59549,75	59987,55	-29518,14

Tabelle 4: Vergleich testtheoretischer Modelle

Die Vergleiche zeigen nach dem *AIC*-Kriterium eine signifikant bessere Passung der Daten für die mehrparametrischen Modelle, jedoch unterscheiden sich die informationstheoretischen Maße nur gering. Für das Informationskriterium *BIC* weist das 2-PL-Modell die besten Werte auf.

Analog zu Kapitel 6.2.3 kann auch hier die Modellpassung anhand von FIT-Statistiken validiert werden. Da INFIT- und OUTFIT-Statistiken nur im Rahmen des Rasch-Modells verwendet

werden, wird stattdessen der FIT-Index $(S - \chi^2)$ (Orlando und Thissen 2003) eingesetzt. Dieser misst ebenfalls auf Basis des χ^2 – Ansatzes die residuale Abweichung von beobachteten zu theoretisch vorhergesagten Werten, jedoch basiert er im Vergleich zu anderen FIT-Indizes auf beobachteten Rohwerten und nicht auf (möglicherweise verzerrten) θ_v-Schätzern (für eine Diskussion vgl. Ames und Penfield 2015). Im Ergebnis zeigt sich eine bessere Passung für das 3- und 4-PL-Modell mit jeweils nur einer signifikanten[1] Abweichung, während das 2-PL-Modell vier und das 1-PL-Modell sieben signifikante Abweichungen aufweisen (Tabelle 5).

[1] $p < 0.01$

	1-PL		2-PL		3-PL		4-PL	
Item	(S-χ2)	*p*	(S-χ2)	*p*	(S-χ2)	*p*	(S-χ2)	*p*
1	23,856	0,249	23,499	0,265	24,569	0,175	24,205	0,148
2	100,69	0	42,265	0,006	32,893	0,047	25,919	0,21
3	37,219	0,016	32,966	0,047	15,012	0,661	15,424	0,565
4	32,028	0,058	26,307	0,195	24,984	0,202	24,2	0,234
5	19,727	0,539	13,741	0,843	14,204	0,82	14,321	0,708
6	283,04	0	64,182	0	47,019	0,001	44,948	0,002
7	41,135	0,005	23,216	0,228	17,808	0,535	16,925	0,528
8	26,709	0,181	24,017	0,242	21,839	0,292	21,673	0,247
9	27,053	0,169	21,86	0,291	20,6	0,3	16,363	0,633
10	40,223	0,007	30,161	0,05	30,437	0,033	29,981	0,026
11	39,084	0,01	17,961	0,525	16,796	0,537	16,5	0,558
12	18,08	0,644	14,143	0,823	14,714	0,741	14,454	0,699
13	23,33	0,327	22,729	0,302	23,96	0,198	23,641	0,167
14	15,764	0,783	15,279	0,76	14,943	0,78	15,496	0,691
15	34,69	0,031	25,374	0,149	19,288	0,438	14,599	0,748
16	27,472	0,156	18,06	0,583	17,52	0,555	16,685	0,545
17	22,156	0,391	16,93	0,658	14,397	0,76	14,796	0,735
18	22,523	0,37	17,33	0,631	17,43	0,561	18,121	0,448
19	58,453	0	18,789	0,658	17,149	0,702	16,895	0,717
20	27,639	0,188	13,356	0,896	12,824	0,847	12,622	0,814
21	30,096	0,116	18,064	0,583	14,263	0,768	15,894	0,664
22	37,653	0,02	28,324	0,102	25,548	0,181	24,98	0,161
23	145,21	0	8,676	0,997	10,707	0,968	10,574	0,957
24	19,061	0,642	18,453	0,62	18,625	0,609	16,775	0,668
25	35,105	0,038	17,845	0,715	13,33	0,897	15,839	0,668
26	28,071	0,173	20,64	0,419	19,452	0,493	20,842	0,346
27	86,105	0	22,51	0,43	20,925	0,464	25,371	0,188
28	80,496	0	42,385	0,006	34,344	0,033	31,868	0,045
29	32,719	0,049	21,896	0,346	22,704	0,304	23,314	0,224
30	19,644	0,544	18,728	0,662	17,976	0,651	17,664	0,61
31	21,747	0,414	13,521	0,854	10,765	0,931	10,967	0,896
32	22,81	0,298	25,471	0,227	20,789	0,41	20,261	0,442

Tabelle 5: Vergleich der (S-χ2)-Indizes

Inwieweit die gezeigten Unterschiede eine Bevorzugung der mehrparametrischen Modelle rechtfertigen, liegt letztlich in der Entscheidung des Testanwenders. Zu klären ist auch, inwieweit die bessere Passung mehrparametrischer Modelle deren testökonomische Restriktionen aufwiegen. Einschränkend muss erwähnt werden, dass die praktische Bedeutsamkeit der üblichen FIT-Indizes in Vergleichsstudien angezweifelt wird (z. B. Sinharay et al. 2011). Für Anwendungen empfiehlt es sich daher, mehrere FIT-Indizes heranzuziehen – dies kann ebenso im Rahmen der Trennschärfeanalyse durch das Hinzuziehen mehrparametrischer Modelle realisiert werden.

Zusätzlich dazu geht durch die höhere Anzahl zu schätzender Parameter eine höhere Streuung einher, zum anderen ist mit der Schätzung ein höherer rechnerischer und somit

testökonomischer Aufwand verbunden. Die Personenfähigkeit wird in Modellen mit zwei oder mehr Parametern mit dem Diskriminationsparameter gewichtet, das erschwert folglich die Interpretation.

Zuletzt wurde geprüft, inwieweit die im IÖK dargestellten Kompetenzbereiche eigene Dimensionen darstellen könnten. Dazu wurde ein dreidimensionales Rasch-Modell (Adams et al. 1997) berechnet und mit dem eindimensionalen Modell aus Abschnitt 6.2.3 verglichen. Der Likelihood-Ratio-Test zeigte dabei eine signifikant bessere Erklärung der Daten durch das eindimensionale Modell. Die $\chi 2$-verteilte Teststatistik betrug 121.32 mit 5 Freiheitsgraden.

Differential Item Functioning (DIF)

Ein wichtiges Validitätskriterium bildet der Vergleich unter Subgruppen. In IRT-basierten Ansätzen weist ein Item genau dann *DIF* auf, wenn Subgruppen unterschiedliche Wahrscheinlichkeiten haben, das Item korrekt zu lösen (Lord 1980). Dies äußert sich im 1-PL-Modell in parallel verlaufenden *Item Characteristic Curves (uniform DIF)* für die betrachteten Teilpopulationen. Für die Berechnung des Effekts werden die jeweiligen Subgruppen getrennt skaliert und die daraus errechneten Parameter auf statistische Unterschiede getestet. Inwieweit ein *DIF*-Effekt vorliegt, kann mithilfe der verbreiteten *ETS*-Klassifikation, die den Effekt nach Signifikanz und Stärke beurteilt, untersucht werden.

Dort wird die *Mantel-Haenszel-Statistik* (MH-Ansatz; vgl. Mantel and Haenszel 1959) auf einer Skala ausgedrückt, die negative Werte aufweist, wenn das Item für die Fokusgruppe schwieriger zu lösen ist. Positive Werte bedeuten eine höhere Schwierigkeit für die Referenzgruppe im Vergleich zur Fokusgruppe[2]. Dabei wird die Mantel-Haenszel *odds ratio* α_{MH} mithilfe der Formel $\Delta - DIF_{MH} = -2.35 \ln(\alpha_{MH})$ (Holland und Thayer 1986) auf die logistische Definition der Deltaskala transformiert. Basierend auf Effektstärke und Signifikanzniveau unterteilt die *ETS*-Klassifikation in drei Kategorien, in denen die auf die Delta-Metrik transformierte *odds ratio* $\Delta - DIF_{MH}$ die Effektstärke und die Mantel-Haenszel-Statistik MH_{χ^2} die Signifikanz auf Basis einer χ^2 -Verteilung beurteilt (z. B. Zwick 2012). Tabelle 6 zeigt die Einteilung in einen vernachlässigbaren, einen moderaten und einen starken *DIF*-Effekt.

[2] Die Fokusgruppe stellt dabei die potenziell benachteiligte Gruppe dar (z. B. Nicht-Muttersprachler).

Kategorie	$\lvert\Delta$ - $DIF\rvert$-Bereich und $MH_{\chi 2}$	Effektgröße
A	$\lvert\Delta$ - $DIF\rvert < 1.0$ und $MH_{\chi 2} < 3.84$	vernachlässigbar
B	$1 \leq \lvert\Delta$ - $DIF\rvert < 1.5$ und $MH_{\chi 2} > 3.84$	moderat
C	$\lvert\Delta$ - $DIF\rvert \geq 1.5$ und $MH_{\chi 2} > 3.84$	stark

Tabelle 6: ETS-Klassifikation

Bei der Messung von Wirtschaftskompetenzen werden Items ausgeschlossen, die einen starken *DIF*-Effekt aufweisen (Kategorie C). Dazu wurden Variablen mit mehr als zwei Ausprägungen (*Muttersprache* und *Bücher zu Hause*) dichotomisiert. Lernende, die bilingual aufwuchsen, wurden zu jener Gruppe hinzugerechnet, die kein Deutsch in ihrer Kindheit gesprochen hatten. Abbildung 5 zeigt die Ergebnisse der DIF-Analyse für die Variablen *Geschlecht, Migrationshintergrund, Muttersprache* und *Bücher zu Hause*. Die Ordinate repräsentiert dabei die auf die Delta-Metrik transformierte *odds ratio* und dessen kategoriale Grenzen. Signifikante Werte sind schwarz gekennzeichnet. Die Berechnungen wurden mithilfe des R-Pakets *difR* (Magis et al. 2010) durchgeführt.

Die Beurteilung unterschiedlicher Messergebnisse unter Subgruppen auf Basis der ETS-Richtlinien ersetzt in diesem Beitrag herkömmliche Modellgeltungstests, da Tests wie beispielweise der LR-Andersen-Test (Andersen 1973) in der Literatur mitunter kritisch beurteilt werden. Zum einen wird dort die Modellgeltung als Nullhypothese und die Ablehnung als Alternativhypothese bestimmt – die zu testenden Hypothesen sind somit im Vergleich zu üblichen Signifikanztests vertauscht, was eine Ausweitung der Irrtumswahrscheinlichkeit nötig machen würde. Welche und wie viele Split-Kriterien angelegt werden sollten, unterliegt zudem der Beliebigkeit (Gärtner et al. 2009). Dadurch wird ein Nachteil des 1-PL-Modells sichtbar: Die strengen Annahmen halten häufig in praktischen Umsetzungen nicht stand.

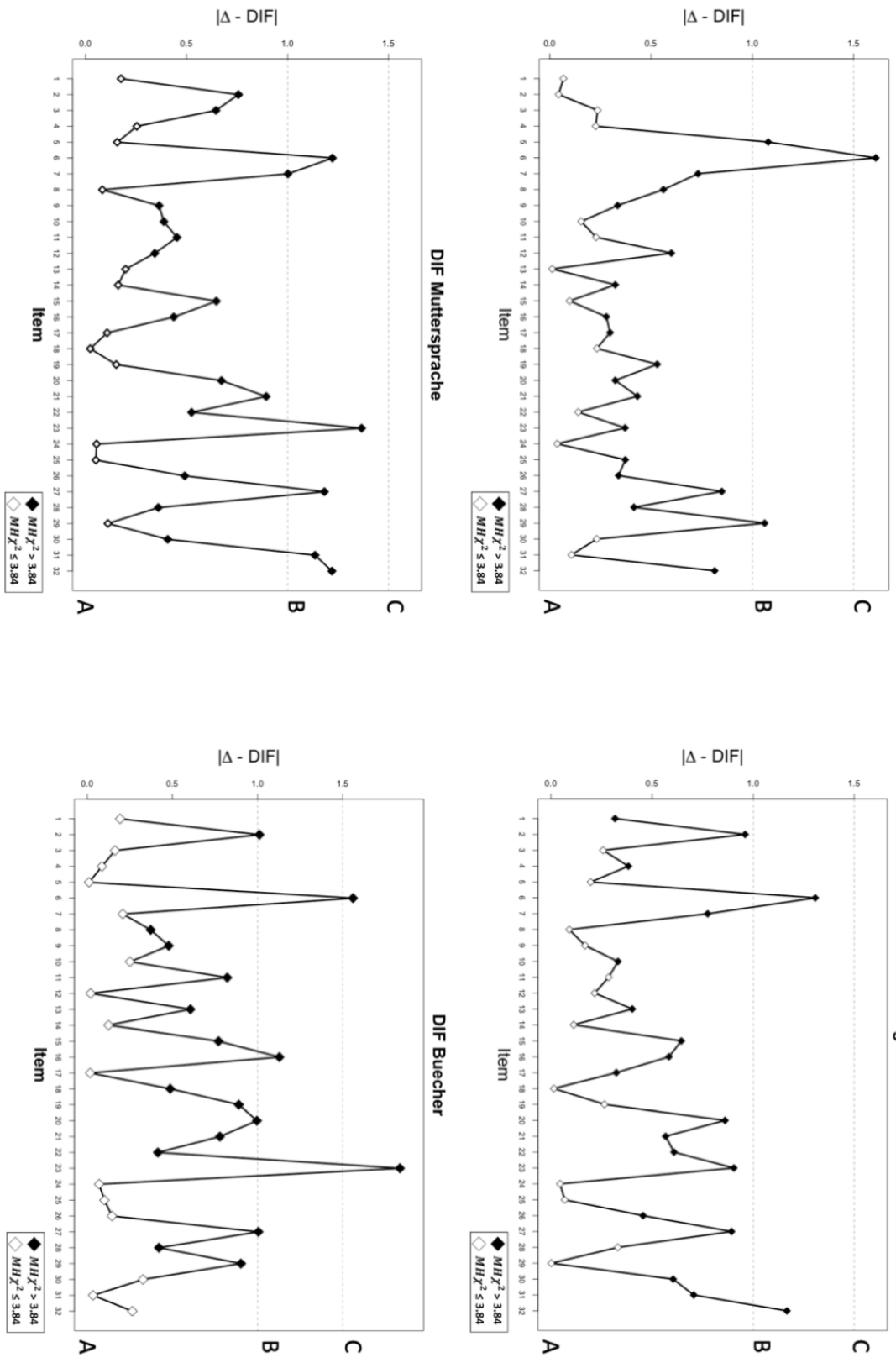

Abbildung 5: Differential Item Functioning

Die Analyse von Subgruppen nach Geschlecht führt zu einem Ausschluss eines Items mit einem signifikanten Effekt von 1.55 zuungunsten der männlichen Testteilnehmer. Zwei weitere Items weisen moderate DIF-Effekte auf, bleiben aber im Itempool enthalten. In den Subgruppen unterteilt nach Muttersprache zeigen fünf Items, in den Subgruppen nach Migrationsstatus lediglich zwei Items einen moderaten DIF-Effekt. Teilnehmende mit schwächer ausgeprägtem Bildungshintergrund (DIF Buecher) werden bei der Bearbeitung zweier Items (6 und 23) systematisch benachteiligt.

6.3 Gruppenvergleiche

Um zu prüfen, inwieweit die aus dem Test gewonnenen Personenfähigkeitswerte erwartbare Gruppenunterschiede aufweisen, werden im letzten Schritt korrelative Zusammenhänge mithilfe eines hierarchischen Regressionsmodells geprüft. Im Zentrum stehen Zusammenhänge zu Basisdemografika (Geschlecht, Migrationsstatus, Bildungshintergrund, Alter), Schulformzugehörigkeiten sowie Interessens- und Einstellungsdimensionen.

Die Mehrebenenstruktur der Daten wird im Folgenden durch die Verwendung eines Random-Intercept-Modells berücksichtigt, welches eine Variation der Mittelwerte auf Schulebene erlaubt. Um die Konfidenzintervalle der geschätzten Koeffizienten nicht zu unterschätzen, wird die Standardfehlerberechnung ebenfalls an die Mehrebenenstruktur angepasst. Hinsichtlich des potenziellen Problems Multikollinearität waren alle genutzten Prädiktorvariablen mit $r < 0.5$ korreliert. Eine multiple Regression mit dem ganzen Variablenset zeigte einen maximalen Variance Inflation Factor von 3,1 (Mittelwert = 1,7). Er liegt damit unterhalb einer kritischen Grenze von 5 (Mansfield und Helms 1982). Ein weiteres potenzielles Problem liegt darin, dass die durch das psychometrische Messmodell ermittelten Kompetenzwerte häufig messfehlerbehaftet sind und in der Folge zu verzerrten Parameterschätzungen führen (Robitzsch et al. 2016). Diesem Umstand wurde im Regressionsmodell durch die Verwendung von Plausible Values (Lüdtke und Robitzsch 2017) begegnet. Dabei wurden auf Basis des WLE-Schätzers (*WLE500*) und weiterer Kovariaten 20 Plausible Values mithilfe eines latenten Regressionsmodells imputiert. Für fehlende Werte in Kovariaten selbst wurde ein MICE-Verfahren (Multiple Imputation of Chained Equation; van Buuren und Groothuis-Oudshoorn 2011) gewählt. Das hierarchische Regressionsmodell (Raudenbush und Bryk 2010) folgt der Gleichung

$$\theta_{i,j}^{PV} = \beta_{0,j} + \beta_{i,j} X_{i,j} + \varepsilon_{i.j} \qquad (4),$$

in der $\theta_{i,j}^{PV}$ die aus dem latenten Regressionsmodell imputierte ökonomische Kompetenz des Lernenden *i* in Schule *j*, $X_{i,j}$ einen Kovariatenvektor und $\varepsilon_{i.j}$ den adjustierten Fehlerterm repräsentiert, wohingegen $\beta_{0,j} = \gamma_{00} + u_{0,j}$ für eine Komposition aus dem Kompetenzmittelwert γ_{00} und seiner schulspezifischen Abweichung $u_{0,j}$ steht. Tabelle 7 zeigt Ergebnisse für das Random-Intercept-Modell. Alle nicht kategorialen Variablen sind mittelwertzentriert. Komplexere Modellspezifikationen werden in Oberrauch und Kaiser (2018) diskutiert.

Abhängige Variable: Ökonomische Kompetenz (transformierte Logit-Skala)

	(1)	(2)	(3)	(4)	(5)
Geschlecht	12.732*	13.247*	10.717	12.754*	11.127*
	(5.72)	(5.54)	(5.65)	(5.30)	(5.15)
Alter	-14.433***	-12.091**	-13.287***	-13.997***	-10.767**
	(4.24)	(3.97)	(4.02)	(4.15)	(3.73)
Migrationsstatus	-23.384***	-23.083***	-23.115***	-21.502***	-21.373***
	(4.32)	(4.20)	(4.19)	(4.22)	(4.09)
Bücher zu Hause	9.169***	8.135***	7.494***	7.704***	5.629**
	(1.61)	(1.73)	(1.62)	(1.65)	(1.79)
GYM		101.499***			94.214***
		(8.50)			(8.61)
RS		45.656***			40.674***
		(7.74)			(7.76)
GMS		21.268*			14.936
		(9.53)			(9.45)
Rechenfähigkeit			13.413***		11.614***
			(2.85)		(2.79)
Lesefähigkeit			12.032***		8.631**
			(3.17)		(2.97)
Fähigkeit allg.			4.071		0.304
			(4.80)		(4.66)
interesteco				11.735**	10.090**
				(3.72)	(3.57)
interestimp				12.918***	11.427***
				(3.03)	(2.92)
constant	490.466***	443.068***	491.978***	490.028***	448.697***
	(7.62)	(7.39)	(7.26)	(7.35)	(7.39)
Observations	1687	1687	1687	1687	1687
R-sqr (total)	0.09	0.34	0.12	0.12	0.37

* p<0.05, ** p<0.01, *** p<0.001; R^2-Werte basieren auf Fishers Z-Transformation; Kovariaten wurden mithilfe von 20 Imputationen in das Modell eingespeist. Erklärte Varianz von fehlenden Werten: 0.001. Standardfehler sind cluster-robust; Intraklassenkorrelation: 0.27

Tabelle 7: Ergebnisse für das Random-Intercept-Modell

Mit Ausnahme von Modell (3) zeigen alle Modelle moderate und signifikante Gender-Gaps zugunsten der männlichen Lernenden, was mit bisherigen Ergebnissen aus einer vorausgegangenen Querschnittsstudie in Baden-Württemberg (Seeber et al. 2018) und anderen einschlägigen Arbeiten (z. B. Walstad 2013; Davies et al. 2005) korrespondiert. Ferner führt unter Berücksichtigung aller Kontrollvariablen ein Alterszuwachs um ein Jahr zu einem niedrigeren Kompetenzwert um 10,77 Punkte. Dabei ist festzustellen, dass ältere Lernende innerhalb einer Klassenstufe teilweise aus Nicht-Muttersprachlern und Wiederholenden bestehen. Der Migrationsstatus von Lernenden geht mit einem niedrigeren Kompetenzwert um 21,37 Punkte einher und bleibt über alle Modelle hinweg relativ robust. Die Büchervariable weist zwar moderate und signifikante Effekte aus, ist jedoch mit den Schulformvariablen konfundiert – die Intraklassenkorrelation mit der Büchervariable als abhängige Variable betrug 0,24. Bei der Analyse von Schulformeffekten (mit der Werkrealschule als Basis) zeigte die Zugehörigkeit zu einem Gymnasium (GYM) den stärksten Gesamteffekt und betrug fast eine ganze Standardabweichung (Modelle 2 und 5), gefolgt von Lernenden auf der Realschule (RS) mit einem Nettoeffekt von 40,47 Punkten. Der Effekt der Gemeinschaftsschule (GMS) wird insignifikant, sobald mit Selbsteinschätzungs- und Interessenvariablen kontrolliert wird. Ebenso wie die Interessensdimensionen haben die (selbst eingeschätzte) Rechen- und Mathematikfähigkeit lediglich einen schwachen Effekt auf die ökonomische Kompetenz, während die allgemeine Schulleistung keinen signifikanten Zusammenhang aufweist.

7 Zusammenfassung und Diskussion

Der vorliegende Beitrag zeigte die Entwicklung und die Ergebnisse eines Wirtschaftskompetenztests für 7. Klassenstufen auf Basis des Integrationsmodells ökonomischer Kompetenz. Ziel des Beitrags war die Untersuchung der psychometrischen Eigenschaften des Testinstruments, welche als zufriedenstellend klassifiziert werden können. Entsprechende Vorarbeiten, die zum Teil in Seeber et al. (2018) und Hentrich et al. (2017) geleistet wurden, stellten anhand von Expertenbeurteilungen und curricularen Analysen sicher, dass zentrale Inhalte des zu messenden Konstrukts abgebildet werden. Empirisch wird die Messung eines unidimensionalen Globalkonstrukts, wie es im Kompetenzmodell von Seeber et al. (2012) definiert wurde, bestätigt – faktoranalytische Untersuchungen der Dimensionalität wurden ebenso bereits in Seeber et al. (2018) erbracht. Hinsichtlich der Diskriminationsfähigkeit des Tests und seiner internen Konsistenz weisen die Items nach Kriterien der KTT größtenteils mittlere Trennschärfen auf. Im Rahmen der Modellpassung

untersuchte der Beitrag alternative IRT-Modelle mit bis zu vier Parametern. Dabei zeigte sich, dass sich mit der Schätzung zusätzlicher Parameter die Modellpassung verbesserte. Die DIF-Analyse, die im Zusammenhang mit der Testung ökonomischen Wissens erstmals von Walstad und Robson (1997) eingesetzt wurde, erwies sich als wichtiger Baustein zur Berücksichtigung der Konstruktvalidität. Auch im vorliegenden Test mussten zwei Items unberücksichtigt bleiben, da sie sonst bestimmte Subgruppen diskriminiert hätten. Die Analyse belegte für mehrere Items signifikante Unterschiede hinsichtlich des Differential Item Functioning, jedoch erreichte das Ausmaß auf der Delta-Metrik zumeist keine kritischen Werte, sodass das Instrument insgesamt befriedigende psychometrische Eigenschaften aufweist.

Es steht nun ein kompaktes Instrument für die 7. Jahrgangsstufe zur Verfügung[3], welches die Handlungs- und Beurteilungsperspektive als Ausgangspunkt definiert, sich von mathematischen Kompetenzen abgrenzt und anstatt reiner Wissenserhebungen ein (globales) Kompetenzkonstrukt abbildet. Die in Abschnitt 6 errechneten Korrelate korrespondieren größtenteils mit bisherigen Kompetenzerhebungen sowohl im deutschsprachigen Raum als auch mit den Ergebnissen internationaler Testungen ökonomischen (und finanziellen) Wissens. Dies gilt insbesondere für den gezeigten moderaten Gender-Gap zugunsten männlicher Lernender, welcher durch die vorherige Analyse des Differential Item Functioning nicht auf Itemverzerrungen zurückzuführen ist. Dies gilt auch für den Einfluss des Migrationsstatus‘ und der Schulformzugehörigkeit.

Selbstverständlich sind neben den bereits diskutierten methodischen Herausforderungen mit der Erhebung von Kompetenzen in der Domäne Wirtschaft Einschränkungen verbunden. Erstens zeigt sich anhand der Itemverteilung gemäß Kompetenzmodell die Schwierigkeit, adäquate Items für die Rolle des Wirtschaftsbürgers zu entwickeln. Auch wenn Jugendliche zweifelsohne weniger häufig als Wirtschaftsbürger denn beispielsweise als Konsument auftreten, so ist es insbesondere im Kompetenzbereich „Entscheidung und Rationalität“ bisher nicht gelungen, ein adäquates Item für dieses Feld im Kompetenzmodell zu etablieren. Die gemessenen Kompetenzen sollten sich jedoch unabhängig von der zugrunde gelegten Rolle äußern.

Zweitens konnte aufgrund der zugrunde liegenden Komplexität lediglich eine Annäherung an den Validitätsbegriff geleistet werden. Cronbach (1976, 447) merkte bereits an, dass es einen

[3] Das Testlet wurde bisher noch nicht veröffentlicht, da es noch in der Langzeitstudie eingesetzt wird. Für wissenschaftliche Zwecke kann es bei der Projektleitung von WIKO-BW angefordert werden.

validen Test als solchen gar nicht gäbe, sondern lediglich Interpretationen von Daten, die einer bestimmten Prozedur entsprängen. Beispielsweise konnte eine konvergente Validität, nach der das Testergebnis mit Ergebnissen aus einem Test, der ähnliche Inhalte misst, korrelieren sollte, im Rahmen dieses Beitrags nicht geleistet werden. Ebenso konnte wegen des Querschnittscharakters der Daten keine prognostische Validität nachgewiesen werden. Die Frage, inwieweit Kinder mit Fachunterricht höhere Kompetenzwerte aufweisen, wird jedoch im Rahmen weiterer Untersuchungen zu beantworten sein.

Drittens wurde einmal mehr die Misere deutlich, ein aus fachdidaktischer Sicht komplexes Konstrukt empirisch zu erfassen (z. B. Loerwald und Schnell 2016). Dies trifft insbesondere auf die ökonomische Domäne zu, die nicht trennscharf von anderen Domänen abgegrenzt werden kann. Ökonomische Kompetenzen enthalten auch domänenfremde Dispositionen wie mathematische Fähigkeiten oder die Reflexion über politische und systemische Rahmenbedingungen (Seeber et al. 2012, 51). Aus diesem Grund untersuchen Arbeiten entweder Teilfacetten ökonomischer Kompetenz oder entwerfen komplexere Messmodelle, wie es im Rahmen der ECOS-Studien unternommen wurde (Macha 2015). Angesichts des hohen Stellenwertes der Handlungsperspektive und dessen Rollenbeschreibungen im Lehrplan des neuen Faches „Wirtschaft, Berufs- und Studienorientierung“ in Baden-Württemberg erschien es jedoch sinnvoll, das Instrument für die künftige Identifikation von Facheffekten am IÖK und dessen Rollen auszurichten.

Literaturverzeichnis

Abdi, H./Williams, L. J. (2010): Principal component analysis. In: WIREs Comp Stat 2 (4), 433–459. DOI: 10.1002/wics.101.

Achtenhagen, F./Winther, E. (2006): Möglichkeiten des Kompetenzaufbaus und seiner Erfassung. In: G. Minnameier & E. Wuttke (Hrsg.), Berufs- und wirtschaftspädagogische Grundlagenforschung. Lehr-Lern-Prozesse und Kompetenzdiagnostik; Festschrift für Klaus Beck. Frankfurt am Main: Lang, 345-360.

Adams, R. J./Wilson, M./Wu, M. (1997): Multilevel item response models: An approach to errors in variables regression. In: Journal of Educational and Behavioral Statistics, 22, 47-76.

Allgood, S./Walstad, W. B./Siegfried, J. J. (2015): Research on Teaching Economics to Undergraduates. In: Journal of Economic Literature 53 (2), 285–325. DOI: 10.1257/jel.53.2.285.

Ames, A. J./Penfield, R. D. (2015): An NCME Instructional Module on Item-Fit Statistics for Item Response Theory Models. In: Educational Measurement: Issues and Practice 34 (3), 39–48. DOI: 10.1111/emip.12067.

Anastasi, A./Urbina, S. (1997): Psychological testing, 7. ed., Upper Saddle River, NJ: Prentice Hall.

Andersen, E. B. (1973): A goodness of fit test for the rasch model. In: Psychometrika 38 (1), 123–140. DOI: 10.1007/BF02291180.

Ayala, R. J. de (2009): The theory and practice of item response theory, New York: Guilford Press (Methodology in the social sciences).

Bank, V./Retzmann, T. (2012): Fachkompetenz von Wirtschaftslehrern. Grundlagen und Befunde einer Weiterbildungsbedarfsanalyse, Schwalbach/Ts.: Wochenschau Verlag.

Bartlett, M. S. (1951): The Effect of Standardization on a chi square Approximation in Factor Analysis. In: Biometrika (38), 337–344.

Barton, M. A./Lord, F. M. (1981): An Upper Asymptote For The Three-Parameter Logistic Item-Response Model*. In: ETS Research Report Series 1981 (1), i-8. DOI: 10.1002/j.2333-8504.1981.tb01255.x.

Beck, K./Krumm, V. (1998): Wirtschaftskundlicher Bildungs-Test (WBT). Handanweisung, Göttingen.

Becker, W./Greene, W./Rosen, S. (1990): Research on High School Economic Education. In: The Journal of Economic Education 21 (3), 231–245. DOI: 10.1080/00220485.1990.10844670.

Birnbaum, A. (1968): Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. In: F. M. Lord und M. R Novick (Hg.): Statistical theories of mental test scores, Reading: Addison-Wesley.

Bock, R. D./Aitkin, M. (1981): Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. In: Psychometrika 46 (4), 443–459. DOI: 10.1007/BF02293801.

Brown, J. D. (1996): Testing in language programs, Upper Saddle River, NJ: Prentice Hall Regents.

Bucher-Koenen, T./Lusardi, A./Alessie, R. J. M./van Rooij, M. C. J. (2016): How financially literate are women? An overview and insights, Global Financial Literacy Excellence Center (WP 2016-1).

Cattell, R. B. (1966): The Scree Test For The Number Of Factors. In: Multivariate behavioral research 1 (2), 245–276. DOI: 10.1207/s15327906mbr0102_10.

Chalmers, R. P. (2012): mirt: A Multidimensional Item Response Theory Package for the R Environment. In: J. Stat. Soft. 48 (6). DOI: 10.18637/jss.v048.i06.

Clark, L. A./Watson, D. (1995): Constructing validity: Basic issues in objective scale development. In: Psychological Assessment 7 (3), 309–319. DOI: 10.1037/1040-3590.7.3.309.

Cronbach, L. J. (1946): Response Sets and Test Validity. In: Educational and Psychological Measurement 6 (4), 475–494. DOI: 10.1177/001316444600600405.

Cronbach, L. J. (1951): Coefficient alpha and the internal structure of tests. In: Psychometrika 16 (3), 297–334.

Cronbach, L. J. (1976): Test validation. In: Thorndike, R. L. (Hg.): Educational measurement, 2. ed, Washington: American Council on Education, 443–507.

Cronbach, L. J./Meehl, P. E. (1955): Construct validity in psychological tests. In: Psychological Bulletin 52 (4), 281–302. DOI: 10.1037/h0040957.

Magis, D./Béland, S./Tuerlinckx, F./De Boeck, P. (2010): A general framework and an R package for the detection of dichotomous differential item functioning. In: Behavior Research Methods (42), 847–862.

Davies, P./Mangan, J./Telhaj, S. (2005): Bold, reckless and adaptable? Explaining gender differences in economic thinking and attitudes. In: British Educational Research Journal 31 (1), 29–48. DOI: 10.1080/0141192052000310010.

Eckes, T. (2005): Evaluation von Beurteilungen. In: Zeitschrift für Psychologie / Journal of Psychology 213 (2), 77–96. DOI: 10.1026/0044-3409.213.2.77.

Erner, C./Goedde-Menke, M./Oberste, M. (2016): Financial literacy of high school students. Evidence from Germany. In: The Journal of Economic Education 47 (2), 95–105. DOI: 10.1080/00220485.2016.1146102.

Gärtner, M./Heine, J.-H./Hofer, S. (2009): Das Rasch Modell. Modellprüfung und Informationskriterien. Multivariate Statistik bei psychologischen Fragestellungen, 28.01.2009.

Green, B. F. (2013): Construct Validity of Computer-Based Tests. In: Howard Wainer und Henry I. Braun (Hg.): Test Validity, Hoboken: Taylor and Francis.

Guilford, J. P. (1946): New Standards For Test Evaluation. In: Educational and Psychological Measurement 6 (4), 427–438. DOI: 10.1177/001316444600600401.

Guilford, J. P. (1954): Psychometric methods. [Place of publication not identified]: Mcgraw Hill ([McGraw-Hill Series in Psychology.]).

Hambleton, R. K./Rovinelli, R. J. (1986): Assessing the Dimensionality of a Set of Test Items. In: Applied Psychological Measurement 10 (3), 287–302. DOI: 10.1177/014662168601000307.

Hartig, J./Frey, A. (2013): Sind Modelle der Item-Response-Theorie (IRT) das „Mittel der Wahl“ für die Modellierung von Kompetenzen? In: Z Erziehungswiss 16 (S1), 47–51. DOI: 10.1007/s11618-013-0386-0.

Haynes, S. N./Richard, D. C. S./Kubany, E. S. (1995): Content validity in psychological assessment: A functional approach to concepts and methods. In: Psychological Assessment 7 (3), 238–247. DOI: 10.1037/1040-3590.7.3.238.

Heath, J. A. (1989): An econometric model of the role of gender in economic education. In: The American Ecnomic Review 79 (2), 226–230.

Henning, G. (1992): Dimensionality and construct validity of language tests. In: Language Testing 9 (1), 1–11. DOI: 10.1177/026553229200900102.

Hentrich, S./Rolfes, T./Seeber, G. (2017): Entwicklung und Validierung eines Modells zur Messung ökonomischer Kompetenzen Jugendlicher. In: Arndt, H. (Hg.): Perspektiven der Ökonomischen Bildung. Disziplinäre und fächerübergreifende Konzepte, Zielsetzungen und Projekte, Schwalbach/Ts.: Wochenschau Verlag, 140–153.

Holland, P. W./Thayer, D. T. (1986): Differential Item Functioning And The Mantel-Haenszel Procedure. In: ETS Research Report Series 1986 (2), i-24. DOI: 10.1002/j.2330-8516.1986.tb00186.x.

Holland, P. W./Thayer, D. T. (2013): Differential Item Performance and the Manthel-Haenszel Procedure. In: Wainer, H./Braun, H. I. (Hg.): Test Validity, Hoboken: Taylor and Francis.

Howard, K. I./Forehand, G. A. (1962): A Method for Correcting Item-Total Correlations for the Effect of Relevant Item Inclusion. In: Educational and Psychological Measurement 22 (4), 731–735. DOI: 10.1177/001316446202200407.

Itzlinger-Bruneforth, U./Kuhn, J.-T./Kiefer, T. (2016): Testkonstruktion. In: Breit, S./Schreiner, C. (Hg.): Large-Scale Assessment mit R. Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung, 1. Aufl., Wien: facultas.

Kaiser, H. F./Rice, J. (1974): Little Jiffy, Mark Iv. In: Educational and Psychological Measurement 34 (1), 111–117. DOI: 10.1177/001316447403400115.

Kaiser, T./Lutter, A. (2015): Empirische Forschung zu financial literacy – Zugänge – Befunde – Desiderata. In: Zeitschrift für Didaktik der Gesellschaftswissenschaften 6 (2), 77–95.

Kaiser, T./Menkhoff, L. (2017): Does Financial Education Impact Financial Literacy and Financial Behavior, and If So, When? In: The World Bank Economic Review 31 (3), 611–630. DOI: 10.1093/wber/lhx018.

Kotte, D./Lietz, P. (1998): Welche Faktoren beeinflussen die Leistung in Wirtschaftskunde? In: Zeitschrift für Berufs- und Wirtschaftspädagogik 94, 421–434.

Lawshe, C. H. (1975): A quantitative approach to content validity. In: Personnel Psychology 28 (4), 563–575. DOI: 10.1111/j.1744-6570.1975.tb01393.x.

Leiser, D./Ganin, M. (1996): Economic Participation and Economic Socialisation. In: Lunt, P./Furnham, A. (Hg.): The Economic Beliefs and Behaviours of Young People. Cheltenham, 93–109.

Li, Y./Jiao, H./Lissitz, R. W. (2012): Applying Multidimensional Item Response Theory Models in Validating Test Dimensionality: An Example of K–12 Large-scale Science Assessment. In: Journal of Applied Testing Technology 2012 (13) (zuletzt geprüft am 31.03.2019).

Loerwald, D./Schnell, C. (2014): Tests als Instrumente zur Individualdiagnostik in der ökonomischen Bildung. Konzeption, Validierung und Auswertung von Testaufgaben für die Sekundarstufe I in Niedersachsen. In: Retzmann, T. (Hg.): Ökonomische Allgemeinbildung in der Sekundarstufe I und Primarstufe. Konzepte, Analysen, Studien und empirische Befunde, Jahrestagung 2013 der Deutschen Gesellschaft für Ökonomische Bildung an der Universität Erlangen-Nürnberg, Schwalbach/Ts.: Wochenschau Verlag, 294–306.

Loerwald, D./Schnell, C. (2016): Diagnostik im Dilemma zwischen fachdidaktischen Ansprüchen und empirischen Anforderungen. Zur (vermeintlichen) Trivialität von Testitems. In: Zeitschrift für Didaktik der Gesellschaftswissenschaften 7 (1), 57–73.

Lohr, S. L. (2010): Sampling: Design and Analysis, 2. Aufl., Boston, MA: Brooks/Cole.

Lord, F. M. (1980): Applications of Item Response Theory To Practical Testing Problems, NJ: Lawrence Erlbaum Associates.

Lüdtke, O./Robitzsch, A. (2017): Eine Einführung in die Plausible-Values-Technik für die psychologische Forschung. In: Diagnostica 63 (3), 193–205. DOI: 10.1026/0012-1924/a000175.

Lusardi, A./Mitchell, O. (2011): Financial literacy around the world: An overview. Hg. v. National Bureau of Economic Research, Cambridge, MA.

Lusardi, A./Mitchell, O. S. (2014): The economic importance of financial literacy. Theory and evidence. In: Journal of Economic Literature 52 (1), 5–44. DOI: 10.1257/jel.52.1.5.

Macha, K. (2015): Ökonomische Kompetenz messen. Theoretisches Modell und Ergebnisse der Economic Competencies Study (ECOS). Zugl.: Siegen, Univ., Diss., 2015. Berlin: LIT Verl. (Ökonomische Bildung, 8).

Macha, K./Schuhen, M. (2011): Modellierung ökonomischer Kompetenz in einer Pilotstudie zu ECOS. In: Siegener Beiträge zur Ökonomischen Bildung (2), 1–29.

Macha, K./Schuhen, M. (2013): ECOS – Ein unter Gendergesichtspunkten fairer Test allgemeiner ökonomischer Kompetenzen. In: Retzmann, T. (Hg.): Ökonomische Allgemeinbildung in der Sekundarstufe II. Konzepte, Analysen und empirische Befunde, Schwalbach/Ts: Wochenschau Verlag, 140–152.

Magis, D. (2013): A Note on the Item Information Function of the Four-Parameter Logistic Model. In: Applied Psychological Measurement 37 (4), 304–315. DOI: 10.1177/0146621613475471.

Mansfield, E. R./Helms, B. P. (1982): Detecting Multicollinearity. In: The American Statistician 36 (3a), 158–160. DOI: 10.1080/00031305.1982.10482818.

Mantel, N./Haenszel, W. (1959): Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. In: JNCI: Journal of the National Cancer Institute (22), 719–748. DOI: 10.1093/jnci/22.4.719.

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

Messick, S. (1980): Test validity and the ethics of assessment. In: American Psychologist 35 (11), 1012–1027. DOI: 10.1037//0003-066X.35.11.1012.

Müller, K./Fürstenau, B./Witt, R. (2007): Ökonomische Kompetenz sächsischer Mittelschüler und Gymnasiasten. In: Zeitschrift für Berufs- und Wirtschaftspädagogik 103 (2), 227–247.

Nickolaus, R./Gschwendtner, T./Geißel, B. (2008): Entwicklung und Modellierung beruflicher Fachkompetenz in der gewerblich-technischen Grundbildung. In: Zeitschrift für Berufs- und Wirtschaftspädagogik (104(1)), 48–73.

Oberrauch, L./Kaiser, T. (2018): Economic compentence in early secondary school: Evidence from a large-scale assessment in Germany (zuletzt geprüft am 06.04.2019).

OECD (2014a): PISA 2012 Results: Students and Money. Financial Literacy Skills for the 21st Century.

OECD (2014b): PISA 2012 Technical Report.

OECD INFE (2012): Supplementary questions. Optional survey questions for the OECD INFE financial literacy core questionnaire, Beirut.

Orlando, M./Thissen, D. (2003): Further Investigation of the Performance of S - X2: An Item Fit Index for Use With Dichotomous Item Response Theory Models. In: Applied Psychological Measurement 27 (4), 289–298. DOI: 10.1177/0146621603027004004.

Pohl, S./Gräfe, L./Rose, N. (2014): Dealing with omitted and not-reached items in competence tests. Evaluating approaches accounting for missing responses in item response theory models. In: Educational and Psychological Measurement 74 (3), 423–452. DOI: 10.1177/0013164413504926.

Rasch, G. (1960): Probabilistic models for some intelligence and attainment tests. Kopenhagen: Danish Institute for Educational Research.

Raudenbush, S. W./Bryk, A. S. (2010): Hierarchical linear models. Applications and data analysis methods, 2. ed., [Nachdr.]. Thousand Oaks, Calif.: Sage Publ (Advanced quantitative techniques in the social sciences, 1).

Retzmann, T./Frühauf, F. (2014): „Financial Fitness for Life“ – Reichweite und Grenzen der US-amerikanischen Testreihe für die finanzielle Allgemeinbildung. In: Retzmann, T. (Hg.): Ökonomische Allgemeinbildung in der Sekundarstufe I und Primarstufe. Konzepte, Analysen, Studien und empirische Befunde, Jahrestagung 2013 der Deutschen Gesellschaft für Ökonomische Bildung an der Universität Erlangen-Nürnberg, Schwalbach/Ts.: Wochenschau Verlag

Robitzsch, A. (2018): Supplementary item response theory models. R package version 2.7-50.

Robitzsch, A./Giang, P./Takuya, Y. (2016): Fehlende Daten und Plausible Values. In: Breit, S./Schreiner, C. (Hg.): Large-Scale Assessment mit R. Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung, 1. Aufl., Wien: facultas.

Robitzsch, A./Kiefer, T./Wu, M. (2018): TAM: Test analysis modules. R package version 2.12-18.

Rose, N./Davier, M. von/Xu, X. (2010): Modeling nonignorable missing data with item response theory (IRT). In: ETS Research Report Series 2010 (1), i-53. DOI: 10.1002/j.2333-8504.2010.tb02218.x.

Rubio, D. M./Berg-Weger, M./Tebb, S. S./Lee, E. S./Rauch, S. (2003): Objectifying content validity: Conducting a content validity study in social work research. In: Social Work Research 27 (2), 94–104. DOI: 10.1093/swr/27.2.94.

Rumpold, H./Greimel-Fuhrmann, B. (2016): Wirtschaftswissen in der Sekundarstufe I. Entwicklung eines Erhebungsinstruments für die Zielgruppe von Schüler/inne/n der achten Schulstufe. In: Zeitschrift für ökonomische Bildung (5), 119–149.

Salehi, M./Tayebi, A. (2012): Differential Item Functioning: Implications for Test Validation. In: JLTR 3 (1). DOI: 10.4304/jltr.3.1.84-92.

Schelten, A. (1997): Testbeurteilung und Testerstellung. Grundlagen der Teststatistik und Testtheorie für Pädagogen und Ausbilder in der Praxis. 2., durchges. und erw. Aufl., Stuttgart: Steiner.

Schnell, C. (2017): Hat der Migrationshintergrund einen Einfluss auf die Schülerleistung im Fach Wirtschaft? Ergebnisse einer Studie in Niedersachsen. In: Zeitschrift für ökonomische Bildung (6), 98-120.

Schnell, C. (2016): „Lautes Denken“ als qualitative Methode zur Untersuchung der Validität von Testitems. Erkenntnisse einer Studie zur Diagnose des ökonomischen Fachwissens von Schülerinnen und Schülern der Sekundarstufe I. In: Zeitschrift für ökonomische Bildung (5), 26-49.

Schuhen, M./Schürkmann, S. (2014): Construct validity of financial literacy. In: International Review of Economics Education 16, 1–11. DOI: 10.1016/j.iree.2014.07.004.

Schumann, S./Eberle, F. (2014): Ökonomische Kompetenzen von Lernenden am Ende der Sekundarstufe II. In: Z Erziehungswiss 17 (S1), 103–126. DOI: 10.1007/s11618-013-0459-0.

Schumann, S./Oepke, M./Eberle, F. (2011): Über welche ökonomischen Kompetenzen verfügen Maturandinnen und Maturanden? Hintergrund, Fragestellungen, Design und Methode des Schweizer Forschungsprojekts OEKOMA im Überblick. In: Faßhauer, U./Aff, J./Fürstenau, B./Wuttke, E. (Hg.): Lehr-Lernforschung und Professionalisierung. Perspektiven der Berufsbildungsforschung, Farmington Hills: Opladen, 51–63.

Schürkmann, S./Schuhen, M. (2013): Kompetenzmessung im Bereich financial literacy. Ergebnisse zum Umgang mit Online-Rechnern aus der FILS-Studie. In: Zeitschrift für ökonomische Bildung (01), 73–89.

Seeber, G./Körber, L./Hentrich, S./Rolfes, T./Haustein, B. (2018): Ökonomische Kompetenzen Jugendlicher in Baden-Württemberg. Testergebnisse für die Klassen 9, 10 und 11 der allgemeinbildenden Schulen, hg. v. Stiftung Würth, Künzelsau: Swiridoff.

Seeber, G./Retzmann, T. (2017): Financial Literacy – Finanzielle (Grund)Bildung – Ökonomische Bildung. In: Vierteljahreshefte zur Wirtschaftsforschung, 86 (3/2017), 69-80.

Seeber, G./Retzmann, T./Remmele, B./Jongebloed, H.-C. (2012): Bildungsstandards der ökonomischen Allgemeinbildung. Kompetenzmodell – Aufgaben – Handlungsempfehlungen, Schwalbach /Ts.: Wochenschau Verlag.

Shepard, L. A. (1993): Chapter 9: Evaluating Test Validity. In: Review of Research in Education 19 (1), 405–450. DOI: 10.3102/0091732X019001405.

Siegfried, J. J./Fels, R. (1979): Research on Teaching College Economics: A survey. In: Journal of Economic Literature (17(3)), 923–969.

Sinharay, S./Haberman, S. J./Jia, H. (2011): Fit Of Item Response Theory Models: A Survey Of Data From Several Operational Tests. In: ETS Research Report Series 2011 (2), i-80. DOI: 10.1002/j.2333-8504.2011.tb02265.x.

Soper, J. C./Walstad, W. B. (1987): Test of economic literacy. Examiner's Manual, 2. Aufl., New York: Joint Council on Economic Education.

Stewart-Brown, S./Tennant, A./Tennant, R./Platt, S./Parkinson, J./Weich, S. (2009): Internal construct validity of the Warwick-Edinburgh Mental Well-being Scale (WEMWBS): a Rasch analysis using data from the Scottish Health Education Population Survey. In: Health and quality of life outcomes 7, 15. DOI: 10.1186/1477-7525-7-15.

Streiner, D. L. (2003): Starting at the beginning: an introduction to coefficient alpha and internal consistency. In: Journal of personality assessment 80 (1), 99–103. DOI: 10.1207/S15327752JPA8001_18.

Thompson, B./Daniel, L. G. (1996): Factor Analytic Evidence for the Construct Validity of Scores: A Historical Overview and Some Guidelines. In: Educational and Psychological Measurement 56 (2), 197–208. DOI: 10.1177/0013164496056002001.

van Buuren, S./Groothuis-Oudshoorn, K. (2011): mice: Multivariate Imputation by Chained Equations in R. In: J. Stat. Soft. 45 (3). DOI: 10.18637/jss.v045.i03.

Verbraucherzentrale Bundesverband (2006): Alltagskompetenz im Test – Umfrage an Berliner Schüler. Online: www.vzbv.de/mediapics/auswertung_schuelerumfrage_weltverbrauchertag_2006.pdf (zuletzt geprüft am 23.06.2017).

Wainer, H./Braun, H. I. (Hg.) (2013): Test Validity. Hoboken: Taylor and Francis. Online: http://gbv.eblib.com/patron/FullRecord.aspx?p=1272892.

Walstad, W. B. (2013): Economic Understanding in US High School Courses. In: American Economic Review 103 (3), 659–663. DOI: 10.1257/aer.103.3.659.

Walstad, W. B./Rebeck, K. (2001): Test of economic literacy. Examiner's manual, 3rd ed., New York, NY: National Council on Economic Education.

Walstad, W. B./Robson, D. (1997): Differential Item Functioning and Male-Female Differences on Multiple-Choice Tests in Economics. In: The Journal of Economic Education 28 (2), 155–171. DOI: 10.1080/00220489709595917.

Warm, T. A. (1989): Weighted likelihood estimation of ability in item response theory. In: Psychometrika 54 (3), 427–450. DOI: 10.1007/BF02294627.

Weinert, F. E. (2001): Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In: Weinert, F. E. (Hg.): Leistungsmessungen in Schulen, Weinheim: Beltz, 17–31.

Wright, B. D./Linacre, J. M. (1994): Reasonable Mean-Square Fit Values. In: Rasch Measurement Transactions 8 (3), 370. Online: https://www.rasch.org/rmt/rmt83b.htm.

Wright, B. D./Masters, G. N. (1982): Rating Scale Analysis. Rasch measurement, Chicago, Ill.: Mesa Pr.

Würth, R./Klein, H. J. (2001): Wirtschaftswissen Jugendlicher in Baden-Württemberg. Eine empirische Untersuchung, Künzelsau: Swiridoff (Schriften des Interfakultativen Instituts für Entrepreneurship an der Universität Karlsruhe (TH), 4).

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

Zwick, R. (2012): A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement. Educational Testing Service, Princeton (Research Report, ETS RR-12-08).

Anhang:

Tabelle A1: Parameter und Standardfehler 1PL - 4PL-Modell (nach Ausschluss der Items 6 und 23)

		1-PL		2-PL				3-PL						4-PL							
Itemnr.	n	σ	SE	α	SE	σ	SE	α	SE	σ	SE	γ	SE	α	SE	σ	SE	γ	SE	δ	SE
1	1689	-1.121	0.062	0.829	0.079	-1.34	0.123	1.238	0.35	-0.164	0.494	0.402	0.143	1.136	0.355	-0.343	0.667	0.351	0.204	1	0.003
2	1619	-0.591	0.059	0.511	0.065	-1.067	0.163	0.508	0.065	-1.062	0.185	0.003	0.023	2.352	0.918	-1.386	0.115	0.001	0.016	0.721	0.026
3	1556	-0.553	0.06	0.542	0.067	-0.957	0.15	3.728	1.104	1.049	0.083	0.54	0.02	3.701	1.142	1.036	0.086	0.541	0.02	1	0.005
4	1554	-0.545	0.06	0.71	0.072	-0.743	0.104	0.722	0.074	-0.719	0.133	0.004	0.029	1.148	0.297	-1.055	0.196	0.005	0.066	0.847	0.055
5	1560	-0.473	0.06	0.983	0.082	-0.499	0.07	1.399	0.242	0.147	0.206	0.253	0.076	1.494	0.365	0.196	0.186	0.278	0.074	0.996	0.042
7	1550	-0.384	0.06	0.995	0.082	-0.4	0.067	1.806	0.297	0.361	0.13	0.302	0.049	1.841	0.563	0.323	0.14	0.299	0.059	0.992	0.055
8	1587	-0.345	0.059	0.927	0.077	-0.381	0.068	0.924	0.097	-0.354	0.184	0.008	0.068	1.009	0.222	-0.473	0.229	0.005	0.066	0.956	0.081
9	1565	-0.312	0.059	1.11	0.086	-0.303	0.059	1.606	0.247	0.232	0.144	0.225	0.058	4.643	2.196	0.179	0.086	0.327	0.036	0.879	0.027
10	1584	-0.176	0.058	1.101	0.084	-0.175	0.058	1.752	0.299	0.375	0.134	0.232	0.054	1.644	0.294	0.309	0.153	0.209	0.064	0.999	0.007
11	1558	-0.146	0.059	1.166	0.088	-0.136	0.056	1.371	0.218	0.083	0.192	0.093	0.084	1.872	0.679	0.069	0.144	0.159	0.086	0.93	0.062
12	1566	-0.139	0.059	0.977	0.079	-0.148	0.063	1.264	0.243	0.274	0.229	0.165	0.088	1.231	0.234	0.221	0.24	0.147	0.094	0.999	0.006
13	1620	-0.117	0.058	0.981	0.078	-0.123	0.061	1.691	0.346	0.504	0.151	0.252	0.059	1.524	0.31	0.419	0.176	0.221	0.07	0.999	0.007
14	1570	-0.007	0.058	0.798	0.072	-0.01	0.072	0.816	0.076	0.007	0.095	0.003	0.023	0.973	0.223	-0.291	0.296	0.002	0.022	0.894	0.106
15	1613	0.01	0.058	1.251	0.09	0.01	0.052	1.272	0.093	0.029	0.057	0.001	0.011	2.201	0.694	-0.099	0.131	0.086	0.076	0.856	0.043
16	1614	0.016	0.058	1.186	0.087	0.016	0.054	1.681	0.262	0.4	0.121	0.169	0.051	2.269	0.781	0.341	0.111	0.212	0.054	0.933	0.057
17	1628	0.021	0.057	0.987	0.078	0.021	0.061	0.998	0.081	0.038	0.073	0.002	0.017	1.214	0.219	-0.221	0.185	0.001	0.015	0.895	0.075
18	1602	0.173	0.058	0.906	0.075	0.193	0.066	1.707	0.317	0.798	0.108	0.245	0.04	2.286	0.791	0.666	0.137	0.27	0.039	0.919	0.068
19	1606	0.734	0.06	0.419	0.063	1.562	0.249	1.294	0.381	1.915	0.198	0.255	0.032	2.552	2.03	1.172	0.496	0.28	0.025	0.693	0.182
20	1629	0.913	0.061	1.218	0.09	0.835	0.068	1.255	0.095	0.834	0.067	0.001	0.005	1.394	0.209	0.591	0.234	0.001	0.007	0.879	0.11
21	1571	0.997	0.063	1.216	0.092	0.913	0.072	1.256	0.096	0.908	0.07	0	0.003	1.749	0.288	0.353	0.168	0.001	0.01	0.737	0.075
22	1619	1.139	0.063	1.058	0.085	1.141	0.09	1.703	0.307	1.235	0.081	0.096	0.027	1.72	0.291	1.23	0.084	0.098	0.026	0.999	0.015
24	1623	1.18	0.064	0.856	0.077	1.378	0.121	1.261	0.233	1.435	0.11	0.084	0.032	2.996	1.357	0.69	0.181	0.129	0.022	0.622	0.082
25	1628	1.199	0.064	0.506	0.068	2.167	0.284	1.845	0.542	1.897	0.153	0.194	0.022	2.084	0.661	1.883	0.142	0.202	0.02	0.999	0.018
26	1629	1.334	0.066	1.108	0.089	1.296	0.095	1.151	0.096	1.272	0.091	0.001	0.005	1.445	0.482	0.736	0.295	0.004	0.042	0.743	0.145
27	1609	1.347	0.066	0.275	0.067	4.289	1.021	1.682	0.766	2.483	0.364	0.205	0.021	3.803	3.401	2.178	0.221	0.22	0.014	0.993	0.081
28	1618	1.519	0.068	0.408	0.07	3.346	0.557	2.161	0.517	2.048	0.158	0.165	0.014	4.82	3.727	1.306	0.291	0.17	0.012	0.551	0.125
29	1614	1.674	0.071	1.111	0.095	1.622	0.115	1.212	0.197	1.574	0.109	0.008	0.024	1.389	0.508	1.283	0.564	0.017	0.03	0.835	0.294
30	1628	1.827	0.073	0.629	0.08	2.74	0.318	1.291	0.346	2.236	0.228	0.085	0.023	1.524	0.976	1.608	0.996	0.088	0.03	0.646	0.39
31	1611	2.247	0.082	1.318	0.118	1.945	0.128	1.376	0.126	1.884	0.124	0	0.001	1.825	0.425	1.037	0.406	0	0.001	0.535	0.158
32	1634	2.492	0.089	0.348	0.093	6.456	1.654	2.149	0.545	2.53	0.246	0.076	0.009	3.656	2.497	1.735	0.608	0.078	0.008	0.446	0.261

Hinweis: Berechnungen erfolgten ohne Fixierung der unteren und oberen Asymptoten mithilfe des R-Pakets mirt; Standardfehlerberechnung erfolgte auf Basis der *central difference approximatio("Oakes"); EM-Cyclen: 20000*

Items: auf Anfrage