

FABIO FORTUNATI / NINA JOHANNA WELSANDT /
HERMANN JOSEF ABS / ESTHER WINTHER

Die Entwicklung eines authentischen technologiebasierten Tests zur Erfassung wirtschaftlicher Kompetenz

Psychometrische Eigenschaften des Testinstruments TBA-EL

The Development of an Authentic Technology-Based Test for Measuring Economic Competence

Psychometric Properties of the Test Instrument TBA-EL

KURZFASSUNG: Erste schulische Grundlagen für den systematischen Erwerb ökonomischer Kompetenz werden in den Lerngelegenheiten der Sekundarstufe I gelegt. Kern dieses Beitrages ist es, (1) die Entwicklung eines authentischen technologiebasierten Testinstruments zu beschreiben, mit dem ökonomische Kompetenz in der Jahrgangsstufe 8 erfasst werden kann und (2) die Untersuchung der psychometrischen Eigenschaften dieses Instruments. Von besonderem Interesse ist, wie sich durch spezifische Designkriterien die Testfairness sichern lässt. Darüber hinaus wird der Einfluss personenbezogener Merkmale auf die Testleistung betrachtet. Die Ergebnisse zeigen eine gute psychometrische Qualität der Kompetenzmessung als auch geringe DIF-Effekte bezogen auf persönliche Merkmale. Die erzielten Evidenzen dienen zum einen der langfristigen Verbesserung des Testinstruments und verdeutlichen zum anderen, wie mithilfe eines authentischen und technologiebasierten Testdesigns die Testfairness eines Instruments gestärkt werden kann.

Schlagwörter: Authentisches Assessment, ökonomische Kompetenz, Testfairness

ABSTRACT: The first school-based foundations for the systematic acquisition of economic literacy are laid in the learning opportunities of lower secondary school. The core of this paper is (1) to describe the development of an authentic technology-based test instrument that measures economic literacy in grade 8 and (2) to investigate the psychometric properties of the instrument. Of particular interest, in addition to test fairness, is the influence of person-related characteristics on test performance. The results show a good psychometric quality of the competency measurement as well as low DIF effects related to personal characteristics. The evidence obtained serves on the one hand for the long-term improvement of the test instrument and on the other hand illustrates how the test fairness of an instrument can be strengthened with the help of an authentic and technology-based test design.

Keywords: Authentic Assessment, Economic Literacy, Test Fairness

1. Problemstellung und Zielsetzung

Trotz der von der KMK betonten Relevanz einer grundlegenden Wirtschaftskompetenz bei Schüler:innen (KMK, 2006) zur Erklärung sozialer, politischer und gesellschaftlicher Phänomene ist die schulische Verankerung von wirtschaftlichen Lerngelegenheiten zwischen den Bundesländern höchst heterogen (FORTUNATI/WINTHER, ACCEPTED WITH REVISIONS, 2023B; GÖKBUDAK, 2021; WEBER, 2023). Für den systematischen Erwerb wirtschaftlicher Kompetenz ist eine schulische Verankerung wirtschaftlicher Lerngelegenheiten jedoch unerlässlich (KAISER ET AL., 2020; G. SEEBER ET AL., 2022; SIEGFRIED/ACKERMANN, 2020). Der Forschungsstand an Evidenz über die wirtschaftliche Kompetenz erweist sich bei Schüler:innen der Sekundarstufe I als lückenhaft. Die meisten Testinstrumente zur Messung wirtschaftlicher Kompetenz adressieren einzelne Inhaltsbereiche in der Sekundarstufe II, da dort eine curriculare Verankerung in den meisten Bundesländern vorhanden ist (ACKERMANN, 2021; ENGARTNER ET AL., 2018; SIEGFRIED/ACKERMANN, 2020).

Neben einer alltäglichen Relevanz von wirtschaftlicher Kompetenz weisen erste Befunde im Bereich der kaufmännischen Berufsausbildung darauf hin, dass eine ökonomische Literalität im Sinne eines ökonomischen Grundwissens einen hohen prädiktiven Einfluss auf den Erwerb berufsspezifischer kaufmännischer Kompetenzen aufweist (BAETHGE-KINSKY ET AL., 2016; NICKOLAUS ET AL., 2013; WINTHER, 2010; WINTHER/ACHTENHAGEN, 2009). Vergleichende Ergebnisse zwischen deutschen und schweizerischen Auszubildenden in den kaufmännischen Berufen Industrie- und Logistikkaufmann demonstrieren, dass die schweizerischen Auszubildenden bei ähnlichen individuellen Merkmalen sowohl höhere berufsspezifische als auch domänenverbundene Kompetenzen aufweisen (SCHUMANN ET AL., 2017). In Bezug auf die ökonomische Literalität von Schülerinnen und Schülern vor Beginn der Berufsausbildung in Deutschland zeigen sich speziell in der Sekundarstufe II Auffälligkeiten. So können vergleichende Befunde zwischen Deutschland und der Schweiz deutliche Unterschiede zwischen den Kompetenzniveaus zugunsten der schweizerischen Schülerschaft feststellen (SIEGFRIED/ACKERMANN, 2020). Ähnliches zeigt sich auch bei Befunden innerhalb des Bereichs der Berufsbildung (EBERLE ET AL., 2016).

Gleichwohl sind die erworbenen Schulabschlüsse nach Beendigung der Sekundarstufe I die erste Zugangsvoraussetzung für den Übergang in das Erwerbsleben. Aufgrund der heterogenen Verankerung ökonomischen Wissens in den Bundesländern ist Evidenz über den Kompetenzstand bei Beginn der kaufmännischen Berufsausbildung von besonderer Relevanz zur Einschätzung des Vorwissens der Auszubildenden. Erkenntnisse im Bereich der Sekundarstufe I sind rar und zumeist aus dem internationalen Forschungsfeld der Economic Literacy (SOPER/WALSTAD, 1987; WALSTAD ET AL., 2013) oder Financial Literacy (FÖRSTER ET AL., 2018; OECD, 2014, 2020; WALSTAD/REBECK, 2017). Bezogen auf die Sekundarstufe I existieren im deutschsprachigen Raum Erhebungen, die jedoch inhaltlich auf überwiegend volkswirtschaftliche Kenntnisse (Oberrauch & Kaiser, 2020) oder solche im Bereich der Financial Literacy (CHEN ET AL., 2019; RUMPOLD /GREIMEL-FUHRMANN, 2016; WUTTKE ET AL., 2020) ab-

stellen und curricular eng an ein spezifisches Bundesland gebunden sind (G. SEEBER ET AL., 2022). Auffällig dabei ist, dass trotz der gestiegenen technischen Möglichkeiten die Testinstrumente in der ökonomischen Bildung selten authentischen Handlungssituationen abbilden (WELSANDT/ABS, 2023). Die Bedeutung von Authentizität ist sowohl in der Kompetenzforschung im allgemeinbildenden Schulwesen als auch in der Berufsbildungsforschung von Bedeutung, da die Kontextualität von Kompetenz adressiert wird. Aus kompetenztheoretischer Sicht ist eine möglichst authentische Gestaltung der Testumgebung maßgeblich, da Kompetenz situativ gezeigt wird und nicht losgelöst vom Kontext zu betrachten ist (ACKERMANN, 2019; KLIEME/HARTIG, 2007; WINTHER, 2010). Authentizität im Kontext der Kompetenzmessung bedeutet dabei nicht, eine reale Situation ungefiltert abzubilden, sondern didaktisch so aufzubereiten und zu inszenieren, dass eine (berufs-)typische Situation dargestellt wird, mit der ein spezifisches Kompetenzziel verwoben ist (WINTHER/ACHTENHAGEN, 2009, 2013). Die Entwicklung und Verwendung von computerbasierten digitalen Testumgebungen ermöglicht es, neuartige Operationalisierungsmöglichkeiten zu nutzen, um (1) das latente Konstrukt näher an der Lebenswelt der Zielgruppe anzubinden und (2) durch die veränderte Gestaltung von Stimuli und Antwortformaten in Testaufgaben die Konstruktrepräsentation zu verbessern (GOLDHAMMER/KRÖHNE, 2020).

In der beruflichen Kompetenzforschung lässt sich seit den 2010er-Jahren eine deutliche Zunahme in der Verwendung von authentischen, technologiebasierten Testumgebungen zu konstatieren (WELSANDT/ABS, 2023). Exemplarisch ist hier die ASCOT-Studie zu nennen, in der berufliche Kompetenzen für den gewerblich-technischen, den kaufmännischen als auch den medizinischen Bereich mittels Computersimulationen erhoben wurden (BECK ET AL., 2016). Darüber hinaus stellen Erkenntnisse aus der Berufsbildungsforschung und dem Bereich der Hochschuldidaktik wiederkehrend fest, dass die Messung von Kompetenzen in authentischen Testumgebungen Performanz in spezifischen Situationen erbringen zu können, positiv beeinflussen kann (DEEMER ET AL., 2022; HAERAZI ET AL., 2019). Darüber hinaus konnten positive Effekte für die Motivation (GULIKERS, BASTIAENS, KIRSCHNER/KESTER, 2006; NICOL ET AL., 2014), das Autonomieerleben (RAYMOND ET AL., 2013) als auch Metakognition (WU ET AL., 2015) und Selbstreflexion der eigenen Leistung festgestellt werden (VANAKI/MEMARIAN, 2009). Die authentische Abbildung realer Lebenssituationen stellt somit ein bedeutsames Designelement für die ökologische Validität eines Testinstruments dar.

Ziel dieses Beitrages ist es, den Entwicklungsprozess und die Validierung eines authentischen, technologiebasierten Testinstruments (TBA) zur Messung ökonomischer (Grund-)Kompetenz anhand des Logic-Assessment-Models (KLOTZ, 2015; KLOTZ ET AL., 2015) aufzuzeigen. Die Validierung des authentischen Testinstruments *Technology-based Assessment of Economic Literacy* (TBA-EL¹) dient zum einen der Anbindung an das Prinzip der Handlungsorientierung in der beruflichen Bildung und zum anderen

1 Förderkennzeichen: Diese Studie wird im Rahmen der in Nordrhein-Westfalen mit Mitteln des Ministeriums für Schule und Bildung finanzierten Studie „ECON 2022“ (Aktenzeichen 526–6.08.C)1–1 540'99) gefördert.

dazu empirische Evidenz für die psychometrische Qualität des Testinstruments auf Basis der Feldtestdaten des Projekts ECON 2022 zu gewinnen. Darüber hinaus ist es von besonderem Interesse zu prüfen, ob ein authentisches Testdesign Einflüsse von individuellen und institutionellen Merkmalen und Unterschiede in der Schwierigkeit auf Itemebene bei gleicher Personenfähigkeit reduzieren kann. Bisherige Testinstrumente in der ökonomischen Bildung bspw. diagnostizieren national wie international häufig Geschlechtereffekte (HAPP ET AL., 2019), die auch auf das Testdesign zurückzuführen sind (ACKERMANN/SIEGFRIED, 2019). Ebenfalls können Effekte hinsichtlich weiterer personenbezogener Merkmale wie dem Migrationshintergrund, der familiären Verwendung der Landessprache (KÜHLING-THEES ET AL., 2020; WALSTAD/REBECK, 2017) oder der sozioökonomische Hintergrund (OECD, 2020; REICHERT-SCHLAX ET AL., 2022) als auch institutionelle Merkmale wie die Schulform beobachtet werden (SEEBER ET AL., 2022; G. SEEBER ET AL., 2018). Unklar bleibt, inwieweit ein authentisches Testdesign den diskriminierenden Einfluss von Hintergrundvariablen auf die Testleistung begrenzen kann.

Dieser Beitrag knüpft an diesem Forschungsdesiderat an und fokussiert dabei auf zwei zentrale Forschungsfragen, welche (1) die Überprüfung der empirischen Qualität des Testinstruments als auch (2) den Einfluss von personenbezogenen und institutionellen Merkmalen auf das Testinstrument untersucht:

- *Forschungsfrage 1:* Ist das Testinstrument TBA-EL hinsichtlich seiner psychometrischen Eigenschaften geeignet, das Konstrukt ökonomische Kompetenz in der Sekundarstufe I zu messen?
- *Forschungsfrage 2:* Hinsichtlich welcher personenbezogenen und institutionellen Merkmale zeigt das Testinstrument TBA-EL Unterschiede bei gleicher Personenfähigkeit?

Im Folgenden werden im Beitrag bisherige Testinstrumente und deren Spezifikation thematisiert sowie die Authentizität als bedeutendes Konstruktionsmerkmal zur Kompetenzmessung in den Blick genommen. Im Weiteren wird die Konstruktion des Testinstruments TBA-EL anhand des Evidence-centered-Designs (ECD) dargestellt sowie die psychometrischen Eigenschaften und die Testfairness des Testinstruments geprüft.

2. Wirtschaftliche Kompetenz und Testkonstruktion

2.1 Kurzüberblick über bisherige Testinstrumente und Einflussfaktoren auf Schülerleistungen

In den letzten Jahrzehnten sind im Forschungsfeld der ökonomischen Bildung national wie international zahlreiche Testinstrumente entwickelt worden, die sich hinsichtlich des konzeptionellen Verständnisses von wirtschaftlicher Kompetenz, dem Testdesign und der Zielgruppe zuweilen erheblich unterscheiden (WELSANDT/ABS, 2023). Als richtungsweisendes Testinstrument gilt der Test of Economic Literacy (TEL) (SOPER/

WALSTADT, 1987; WALSTAD ET AL., 2013) sowie dessen Weiterentwicklung und die deutsche Adaption (BECK & KRUMM, 1998). Der TEL setzt den Fokus auf ein Verständnis von grundlegendem volkswirtschaftlichem Wissen und gesamtwirtschaftlicher Zusammenhänge. Aufgrund der breiten Rezeption wird das Testinstrument noch heute in nationalen wie internationalen Studien eingesetzt (FÖRSTER ET AL., 2015; KÜHLING-THEES ET AL., 2021; ZLATKIN-TROITSCHANSKAIA ET AL., 2019; ZLATKIN-TROITSCHANSKAIA ET AL., 2020).

In Bezug auf die inhaltliche Konzeption von Testinstrumenten zeigt sich anknüpfend an die Einführung verbindlicher Bildungsstandards und der Kompetenzorientierung eine Öffnung des Verständnisses von ökonomischer Bildung hin zum Einbezug von ökonomischen Alltagserfahrungen (bspw. Financial Literacy, Verbraucherbildung), betriebswirtschaftlichem Vorwissen (EBERLE, 2015; SCHUMANN/EBERLE, 2011; S. SEEGER ET AL., 2015; TRAMM/SEEGER, 2006) als auch der sozioökonomischen Bildung (ACKERMANN, 2019). Neben der inhaltlichen Öffnung kann im Hinblick auf die Testdurchführung und das Testdesign eine zunehmende Handlungsorientierung bspw. über offene Aufgabenformate und computergestützte Testsysteme beobachtet werden (bspw. PISA Financial Literacy).

Seit dem Jahr 2000 wurden mehr als 23 Testinstrumente zur Messung wirtschaftlicher Kompetenz entwickelt, deren Zielgruppe zumeist in der Sekundarstufe II zu verorten ist (WELSANDT/ABS, 2023).

Für die Sekundarstufe I existieren nur wenige Testinstrumente. Zu nennen ist insbesondere PISA Financial Literacy (OECD, 2014, 2017, 2020), das finanzielle Literalität anhand von Inhaltsbereichen, Kontexten und kognitiven Prozessen erhebt. Der Fokus liegt dabei auf dem privaten Umgang mit Geld sowie der Altersvorsorge und dem Vermögensaufbau. Der Test of Financial Literacy (TFL) von WALSTAD/REBECK (2017) basiert auf den National Standards for Financial Literacy (WALSTAD ET AL., 2017) und bezieht weitere Inhalte der Verbraucherbildung wie den Einfluss von Marketing auf das eigene Kaufverhalten o.Ä. mit ein. In Deutschland gibt es für die Sekundarstufe I nur wenige Testinstrumente zur Messung wirtschaftlicher Kompetenz. Hier sind im Speziellen die Arbeiten von MACHA (2015) hervorzuheben. Dort wird wirtschaftliche Kompetenz für Schüler:innen der Sekundarstufe I in möglichst alltäglichen Handlungssituationen modelliert (MACHA, 2015; MACHA/SCHUHEN, 2011, 2012, 2013). Mit der Einführung des Unterrichtsfachs Wirtschaft/Berufs- und Studienorientierung (WBS) für die Sekundarstufe I im Jahr 2016 wurde die bis dato umfangreichste Befragung von Schüler:innen zur ökonomischen Bildung anhand einer repräsentativen Stichprobe im Längsschnitt durchgeführt (G. SEEGER ET AL., 2022; SEEGER ET AL., 2018). Zu kritisieren ist hier jedoch ein Mangel an Authentizität im Testdesign, da die Items keinem spezifischen Handlungsstrang mit Lebensweltbezug der Schüler:innen folgen. In der Gesamtschau zeigt sich jedoch, dass bis auf Ausnahmen wenig Evidenz für die ökonomische Kompetenzstruktur in der Sekundarstufe I in Deutschland vorliegt. Es fehlt vor allem an Testinstrumenten, die mehrere Teilbereiche ökonomischer Kompetenz und zusätzlich Handlungssituation technologiegestützt und authentisch abbilden (SEEGER ET AL., 2015).

Für eine evidenzbasierte Testwertinterpretation der Schülerleistungen stellt neben der inhaltlichen Ausrichtung sowie der psychometrischen Qualität des Testinstruments auch die Sensitivität der Testleistung bezogen auf Gruppenzugehörigkeiten ein wichtiges Merkmal zur Beurteilung dar. Bezogen auf die Testfairness ist zu beobachten, dass Teilnehmerinnen in Testinstrumenten der ökonomischen Bildung inhaltsübergreifend niedrigere Testergebnisse aufweisen als männliche Testteilnehmer (FÖRSTER ET AL., 2015; SCHUMANN/EBERLE, 2014; SCHUMANN ET AL., 2017; SOPER/WALSTAD, 1987). Als eine mögliche Erklärung gilt das Testdesign, insbesondere das Itemformat (ACKERMANN/SIEGFRIED, 2019). So zeigen TAYLOR und LEE (2012) sowie RIENER und WAGNER (2018), dass Testteilnehmerinnen bei Multiple-Choice Items bessere und bei Single-Choice Items niedrigere Werte erzielen als Testteilnehmer. Darüber hinaus konnte im Bereich der MINT-Bildung ein Vorteil der weiblichen Schülerinnen bei offenen Aufgabenformaten festgestellt werden (BELLER/GAFNI, 2000; BOLGER /KELLAGHAN, 1990; REARDON ET AL., 2018). Wie ACKERMANN und SIEGFRIED (2019) anmerken, ist die Variation des Itemdesigns in Tests der ökonomischen Bildung gering und besteht häufig ausschließlich oder überwiegend aus Single-Choice Items (BECK, 1993; SCHUMANN/EBERLE, 2014). Eine weitere Quelle für eine potenzielle Verletzung der Testfairness stellt der Migrationshintergrund von Schüler:innen dar. Auch hier belegen zahlreiche Large-scale Assessments der ökonomischen Bildung im weitesten Sinne signifikante Unterschiede zulasten von Schüler:innen, die einen Migrationshintergrund aufweisen (OECD, 2014, 2017, 2020). Ähnliches ist beim sozioökonomischen Hintergrund der besuchten Schulform und der familiäre Verwendung der Landessprache festzustellen (REICHERT-SCHLAX ET AL., 2022).

2.2 Authentizität als Konstruktionsmerkmal bei der Testentwicklung in Large-scale Assessments

Im Bereich der Large-scale Assessments kann im letzten Forschungsjahrzehnt beobachtet werden, dass diese nicht mehr primär vereinzelte Wissensaspekte einer Domäne adressieren, sondern komplexere Fähigkeiten wie das domänenspezifische Problemlösen oder Kollaboration mitberücksichtigen. Hintergrund bilden hier „work-ready skills“, bspw. Employability Skills sowie 21st Century Skills (OECD, 2013, 2016a, 2016b, 2019a), deren Ursprung im Bereich der beruflichen Bildungsforschung (siehe bspw. die ASCOT und ASCOT+-Initiative) und im situativen Lernen zu finden sind (GREENO, 1998; GREENO ET AL., 1996). Der Ansatz, berufliche Kompetenzen anhand möglichst authentischer Anforderungssituationen zu messen, findet zunehmend Eingang in die Kompetenzmessung der schulischen Allgemeinbildung (BLÖMEKE ET AL., 2015; PELLEGRINO, 2014; PELLEGRINO/HILTON, 2012) sowie der Hochschulbildung (ANDREWS/HIGSON, 2014; BLÖMEKE ET AL., 2013; COLTHORPE ET AL., 2021; SCHULTZ ET AL., 2022; VILLARROEL ET AL., 2018). So kann festgestellt werden, dass Studierende sich nach Abschluss ihres Studiums nicht ausreichend für die Arbeitswelt vorbereitet (ELLSTRÖM/ELLSTRÖM, 2014) und unsicher fühlen (GERHOLZ/RÜSCHEN, 2011). Darüber hinaus wird

ein fehlender Anwendungsbezug der im Studium erworbenen Kompetenzen kritisiert (PLUMP, 2010). Insbesondere im internationalen Kontext ist ein zunehmender Einsatz von arbeitsweltorientierten Assessments zu beobachten (WU ET AL., 2015). Diese bilden über simulative Methoden mögliche zukünftige Tätigkeiten ab, in denen erworbene Kompetenzen praxisbezogen demonstriert werden sollen (bspw. Problemlösen, Kollaboration, Kommunikation), um Employability Fähigkeiten entwickeln zu können (ANDREWS/HIGSON, 2014; SOTIRIADOU ET AL., 2020). Die Wahrnehmung der Studierenden, ob ein Assessment authentisch gestaltet ist, ist entscheidend für das Engagement bei der Bearbeitung (BLOXHAM, 2015; COLTHORPE ET AL., 2021; GULIKERS, BASTIAENS/KIRSCHNER, 2006; JAMES/CASIDY, 2018; SAMBELL ET AL., 2013).

Authentizität im Bereich der Assessmentforschung kann nach BENNER (2012) und RAYMOND ET AL. (2013) als Realismus, Kontextualisierung und Problematisierung beim Lehren und Messen von fachlichen Inhalten verstanden werden. *Realismus* bezeichnet in dieser Definition die Anschlussfähigkeit einer Testumgebung an für das zu messende Konstrukt bedeutsame Lebens- oder Arbeitssituationen eines Individuums, in denen *kontextualisiert* domänenspezifisches Wissen angewendet werden kann. Dabei müssen die Testaufgaben *problemorientiert* konstruiert werden, um bei den Getesteten ein Bewusstsein für die Bedeutsamkeit eines Problems hervorzurufen und einen Handlungsimpuls zu erzeugen. Dabei muss die authentische Testumgebung sowohl die Anforderungen der Lebens- oder Arbeitswelt als auch Unterrichtsprozesse des Lehrens und Lernens in die Domäne integrieren (GULIKERS ET AL., 2004; WIGGINS, 1990, 1998).

Eine authentische Testumgebung hat in formativen Assessments einen positiven Einfluss auf die Entwicklung komplexerer kognitiver Fähigkeiten, das Engagement und die Motivation der Testteilnehmer:innen bei der Bearbeitung des Assessments als auch der Fähigkeit zur Selbstregulation, Metakognition und Selbstreflexion (AFRIANA/FESTIYED, 2020; DEEMER ET AL., 2022). Aus Validitätsperspektive kann angeführt werden, dass mithilfe einer authentischen Testumgebung die externe Konstruktvalidität erhöht werden kann, da die gemessene Kompetenz lebensweltlich erhoben (CARE/KIM, 2018) und auch fachdidaktisch ein Anschluss über eine simulierte Realität ermöglicht wird (PELLEGRINO ET AL., 2016). Auch aus psychometrischer Perspektive steigt die Bedeutung von authentischen Testumgebungen, um komplexere Kompetenzstrukturen adäquater abbilden zu können (CARE/KIM, 2018; ERCIKAN/OLIVERI, 2016; MISLEVY, 2016; PELLEGRINO ET AL., 2016).

Zur Frage, inwiefern eine Testumgebung als authentisch betrachtet werden kann, existieren in der Literatur zahlreiche Konzeptionen (für einen tiefergehenden Einblick VOLLAVORELL, 2018). Dieser Beitrag bezieht sich auf das breit rezipierte Modell GULIKERS ET AL. (2004), da es sich auch auf die summative Testung bezieht, die im Rahmen dieser Studie relevant ist. Sie beschreiben fünf Merkmale, die Authentizität in einer Testumgebung konstituieren:

1. Eine authentische Aufgabe in einer Testumgebung besteht aus Handlungsanweisungen, welche die avisierten Tätigkeiten des zugrundeliegenden Konstrukts in der Lebenswelt simulieren.

2. Der physische oder virtuelle Kontext einer Testumgebung spiegelt die Art und Weise wider, wie Kompetenzen in der Lebenswelt angewendet werden könnten.
3. Soziale Prozesse, die bedeutsam für das Demonstrieren von Kompetenz in spezifischen Situationen sind, werden miteinbezogen.
4. Der Output der Testumgebung (bspw. die gezeigte Performanz) repräsentiert ein möglichst lebensweltnahes Abbild, lässt über mehrere Indikatoren Rückschlüsse auf das zugrunde liegende Konstrukt zu und ist objektiv überprüfbar.
5. Eine vorausgehende Entwicklung von Kriterien zur Evaluierung des Assessments und Orientierung an Standards, die indizieren, welche Fähigkeiten und Wissensstrukturen für spezifische Leistungsniveaus zu erwarten sind.

Bezogen auf die ökonomische Bildung stellen WELSANDT/ABS (2023) fest, dass 22 von 26 untersuchten Testinstrumenten seit 1990 einen lebensweltlichen Bezug aufweisen, ohne jedoch zwingend zielgruppenorientiert zu sein. Zusätzlich kann konstatiert werden, dass die lebensweltlichen Bezüge der Testinstrumente keiner stringenten Narration folgen und sich zumeist auf spezifische Items beschränken (bspw. bei PISA Financial Literacy).

Gleichwohl ist die Entwicklung von authentischen Testumgebungen mit Herausforderungen verbunden. Der gewählte situative Kontext muss für das gewählte Konstrukt ausreichend bedeutsam sein, um dieses adäquat abbilden zu können (MCARTHUR, 2023; PELLEGRINO ET AL., 2016; WIGGINS, 1990). Dies wirft die Frage auf, inwiefern in authentischen Settings gewährleistet werden kann, wie die zu messende Kompetenz in Gänze und in verschiedenen Komplexitätsniveaus abgebildet werden kann (MISLEVY, 2018). ERCIKAN/OLIVERI (2016) adressieren dies, indem ex ante die Passung konstruierter Items an die zur Messung intendierten Konstruktfacetten überprüft wird (siehe hierzu Construct Mapping, WILSON/DRANEY, 2004). Darüber hinaus ist bei der Entwicklung anzumerken, dass mit einer zunehmenden Komplexität der Testumgebung zumeist ein erhöhter „Cognitive Load“ verbunden ist, der spezifische Gruppen beispielsweise durch ein erhöhtes Sprachniveau diskriminieren und selbst ein schwierigkeitsgenerierendes Merkmal darstellen kann (EHRICH ET AL., 2021; MINKLEY ET AL., 2021; ZU ET AL., 2021). Aus technologischer Perspektive kann angeführt werden, dass durch die Verwendung von computergestützten digitalen Testformaten die Authentizität von Testaufgaben durch den Einsatz von audiovisuellen Medien förderlich und auch in Large-scale Assessments verwendet werden kann (GOLDHAMMER ET AL., 2016; GOLDHAMMER/KRÖHNE, 2020). Zu beachten ist jedoch, dass die Testumgebung für Zielgruppe adäquat und nicht überfordernd ist. Dies kann durch eine vorab durchgeführte Usability-Befragung geprüft werden (SANGMEISTER ET AL., 2018).

3. Methodisches Vorgehen

3.1 Design des Testinstruments

Der Entwicklungsprozess des Testinstruments orientiert sich am Logic-Assessment-Modell von KLOTZ (2015), welches den testtheoretischen Aspekt der Assessment-Triangle (PELLEGRINO ET AL., 2001) mit der durch Beschreibung der Konstruktionsschritte evidenzgenerierenden Perspektive des Evidence-centered-Designs von MISLEVY UND RICONSCENTE (2005) integriert (siehe Tabelle 1).

Im Folgenden soll anhand der Konstruktionsschritte des Logic-Assessment-Modells die Entwicklung des TBA-EL aufgezeigt werden. Kapitel 3.1 bezieht sich auf die Konstruktionsschritte 1 und 2, während Kapitel 3.2 und 4 das Scoring und die verwendeten psychometrischen Messverfahren thematisieren.

Tabelle 1: Logic-Assessment-Modell nach Klotz (2015)

Konstruktionsebene	Bearbeitungsschritte	Zeitplan
1. Itemkonstruktion		Q2 – Q3 2020
Domänenanalyse	Analyse von Fachbüchern, bereits bestehender Testinstrumente, Ordnungsdokumenten etc.	
Domänenmodellierung	Konstruktion des Modells und Prüfung der Repräsentanz in Curricula der Sek I	
Modellspezifikation	Entwicklung der Items und Construct Map	
2. Testkonstruktion		Q4 2020 – Q2 2021
Itemauswahl und Testkomposition	Narrative Einbettung und Anordnung der Items in authentische Handlungssituationen	
Validierung und Anpassung	Expertenbefragung	
Durchführung des Feldtests	Erhebung an 20 Schulen, N=816	
3. Scoring		Q3 2021
Definition von Kategorien	Entwicklung eines detaillierten Codebooks	

Konstruktionsebene	Bearbeitungsschritte	Zeitplan
Bewertung der Kategorien	Aufgabenkorrektur und deren Überprüfung	
4. Psychometrische Messverfahren		Q3 – Q4 2021
Wahl des Messmodells	Probabilistische Testtheorie	
Überprüfung der empirischen Itemgüte	IRT-basierte Itemprüfung	
Modellprüfung	Dimensionalität des Tests	

5. Rückschlüsse auf das Zielkonstrukt (Stichproben- und Personenebene)

Itemkonstruktion – Domänenanalyse, Domänenmodellierung und Construct Mapping

Die projektspezifische Definition ökonomischer (Grund-)Kompetenz orientiert sich inhaltlich sowohl an ACKERMANN (2019) als auch an die Definition von Economic Literacy und Economic Numeracy als domänenverbundene Kompetenz im Verhältnis zur berufsspezifischen kaufmännischen Kompetenz von WINTHER (2010). Die Definition orientierend am kontextspezifischen kognitiven Kompetenzkonzept (HARTIG/KLIEME, 2006) und der prozessualen, handlungsregulativen Differenzierung domänenspezifischer Kompetenzen in Teildimensionen von GELMAN/GREENO (1989). Die gewählte Definition betrachtet weder affektive noch motivationale Faktoren. Ökonomische (Grund-)Kompetenz wird wie folgt verstanden:

Die ökonomische Kompetenz bezieht sich auf die erfolgreiche Bewältigung ökonomisch geprägter Handlungssituationen, die im persönlich-finanziellen, beruflich-unternehmerischen und gesellschaftlich-volkswirtschaftlichen Lebensbereich unter Einbezug der Nachhaltigkeit zu verorten sind. Diese zeigt sich durch das Wissen und die Fähigkeiten, das Anforderungsspektrum einer ökonomischen Problemstellung in einem spezifischen ökonomischen Kontext zu erfassen, zu verstehen, systematisch zu analysieren, Lösungen zu entwickeln, zu beurteilen, begründet handeln und reflektieren zu können. (FORTUNATI/WINTHER, ACCEPTED WITH REVISIONS, 2023A, S. 7)

Ausgehend von der projektspezifischen Kompetenzkonzeption wurde die ökonomische Domäne auf Grundlage des Evidence-centered-Designs analysiert und in Anlehnung an ACKERMANN (2021) ein Domänenmodell entwickelt (FORTUNATI/WINTHER, ACCEPTED WITH REVISIONS, 2023B).

Das Domänenmodell differenziert sich in drei Lebensbereiche, in denen Individuen mit ökonomischen Situationen konfrontiert sind. Im *persönlich-finanziellen Lebensbereich* werden ökonomische Alltagssituationen in der Konsumenten- und Verbraucherrolle wie bspw. Konsumententscheidungen und deren Folgeabschätzungen adressiert. Darüber hinaus ist der verantwortungsvolle Umgang mit den eigenen persönlichen Finanzen (alltägliche Mittelverwendung) und der Aufbau von Privatvermögen von Relevanz. Der *beruflich-unternehmerische Lebensbereich* bezieht sich auf ökonomische Problemstellungen, mit denen Individuen in der Arbeitswelt konfrontiert sind. Der Lebensbereich lässt sich in berufsallgemeine und berufsübergreifende Handlungssituationen differenzieren. Berufsallgemeine Handlungssituationen umfassen Inhalte, die für nahezu alle arbeitenden Personen von Relevanz sind (bspw. Arbeitsrecht, betriebliche Mitbestimmung, Tarifkonflikt, AGG o. Ä.) und nicht an spezifische Berufsfelder gebunden sind. Berufsübergreifende Handlungssituationen beziehen sich auf Inhalte, die in Berufsfeldern – hier spezifisch in kaufmännisch-verwaltenden Berufen benötigt – werden. Dies umfasst bspw. Grundlagen der Unternehmensführung, Rechnungswesen und Kostenkalkulationen. Davon abzugrenzen sind berufsspezifische Handlungssituationen. Der *gesellschaftlich-volkswirtschaftliche Lebensbereich* adressiert gesamtwirtschaftliche Problemstellungen, die alle Einwohner:innen eines Landes betreffen (bspw. Arbeitsmarkt- und Sozialpolitik, Fiskal- und Finanzmarktpolitik, Steuern, internationale Wirtschaftsbeziehungen, Migrationsfragen etc.). Die Problemstellungen zeichnen sich durch einen hohen Abstraktions- und Komplexitätsgrad aus und sind zumeist eng mit politischen Kontexten verknüpft.

Als Querschnittsdimension fungiert das ganzheitliche Konzept von *Nachhaltigkeit*, das aus gesellschaftlicher und fachdidaktischer Sicht an Bedeutung gewinnt (CORSTEN/ROTH, 2012; GRUNDMANN, 2017; HÜBNER, 2017; SCHANK/LORCH, 2018). Ökonomische Problemstellungen werden damit um eine ökologische, ökonomische und soziale Dimension als auch eine intra- und intergenerationelle Facette erweitert, die es ermöglicht, gesellschaftliche Debatten anhand von nachhaltigkeitsinduzierten Zielkonflikten beurteilen und bewerten zu können.

Anknüpfend an die Literacy-Konzeption von WINTHER (2010) lassen sich ökonomische Handlungssituationen über einen *sprachlich-argumentativen* oder *mathematisch-analytischen Zugang* modellieren. Winther beschreibt Economic Literacy und Numeracy als domänenverbundene Bereiche wirtschaftlichen Wissens, die Grundfertigkeiten des (wirtschafts-)beruflichen Handelns darstellen. Unter der Economic Literacy werden text- und bildsprachliche Kenntnisse sowie ein Verständnis von grundlegenden ökonomischen Konzepten in wirtschaftlichen Alltagskontexten verstanden. Die Economic Numeracy lässt sich vor dem Hintergrund der Beherrschung grundlegender mathematischer Kenntnisse und Fähigkeiten in zwei Bereiche gliedern: (1) Bewältigung von Rechenoperationen mit wirtschaftlichen Bezügen und (2) Interpretation und Beurteilen der Aussagekraft sowie des Zustandekommens betriebswirtschaftlicher Realwerte als auch ökonomischer Modellierungen.

Davon ausgehend wurde die curriculare Repräsentanz des Domänenmodells anhand der Analyse von N=31 Curricula der ökonomischen Bildung der Sekundarstufe

I aus 10 verschiedenen Bundesländern und verschiedener Schulformen überprüft. Als eine notwendige Bedingung für die inhaltliche Validität des Testinstruments wird die curriculare Validität, also das Alignment zwischen dem intendierten Curriculum und dem Test verstanden (AERA ET AL., 2014; PELLEGRINO, 2002, 2012; PELLEGRINO ET AL., 2001). Hierfür wurde im Rahmen der curricularen Analyse geprüft, inwieweit sich typische Muster zwischen den curricularen Inhalten sowie der curricular intendierten Wissensrepräsentation und der Wissenserwerbsprozesse identifizieren lassen. Dies dient dem Zweck, neben einer ausreichend inhaltlichen Passung zwischen Testinstrument und Lehrplänen auch eine möglichst enge Anbindung an curricular intendierten kognitiven Prozessen zu erreichen, um den curricular vorgesehenen Kompetenzerwerb präziser abbilden zu können.

Bezüglich der Itemkonstruktion wurde sich an den zwei maßgeblichen Prinzipien der Aufgabenentwicklung (Spezifität und Kognition) im Bereich der beruflichen Assessments orientiert und knüpft an das Konzept der aus kaufmännisch-verwaltender Perspektive domänenübergreifenden Handlungssituationen der Economic Literacy und Numeracy an (WINTHER, 2010).

1. Designprinzip *kognitive Verarbeitungsprozesse*: In Anlehnung MARZANO/KENDALL (2007) sowie WINTHER (2010) wurden drei kognitive Anforderungsniveaus entwickelt, die den Einsatz unterschiedlich komplexer kognitiver Prozesse für das Lösen ökonomischer Problemstellungen repräsentieren. *Anforderungsniveau 1* beschreibt das Wiedererkennen/Erinnern von Fakten und domänenspezifischer Konzepte, welches nötig ist, um Handlungssituationen innerhalb der Domäne bewältigen zu können. *Anforderungsniveau 2* beschreibt das Verstehen und Analysieren von domänenspezifischen Konzepten sowie das Anwenden von prozeduralen Fähigkeiten zur Problemlösung. *Anforderungsniveau 3* setzt voraus, dass für die Lösung einer Aufgabe die Fähigkeit, ökonomische Daten sinnvoll interpretieren, ökonomische Sachverhalte bewerten und Entscheidungen treffen zu können, benötigt wird.
2. Designprinzip *Spezifität*: In Anlehnung an GELMAN/GREENO (1989) wird zwischen domänenspezifischen und domänenverbundenen Inhalten unterschieden. Je spezifischer die Items sind, desto mehr ökonomisches Wissen aus mehreren Teilbereichen ist vonnöten, um die Aufgabe lösen zu können. Unspezifische Items greifen auf generisches Wissen und Fähigkeitsstrukturen zurück, die jedoch eine Voraussetzung für das Lösen domänenverbundener Problemstellungen darstellen. Die Unterscheidung hinsichtlich generischer und domänenspezifischer Kompetenzen kann maßgeblich mit der Kontextualisierung der Aufgabenstellung zusammenhängen. So zeigen bspw. Befunde von HERING ET AL. (2020) und HERING ET AL. (2021), dass der Transfer von generischen mathematischen Kompetenzen im Kontext der kaufmännischen Berufsbildung insbesondere von der Kontextualisierung der Aufgaben abzuhängen scheint. Für die Konstruktion der Items wurde ein Bewertungsraster entwickelt. *Anforderungsniveau 1* entspricht dabei Items, die keine spezifischen unterrichtsbezogenen Vorkenntnisse zur Lösung eines Items voraussetzen, während dies für *Anforderungsniveau 2* wesentlich ist. Für *Anforderungsniveau 3* ist von Bedeu-

tung, dass für das Lösen einer ökonomischen Situation Wissen aus mehreren ökonomischen Teilbereichen notwendig ist.

In Anlehnung an das Konzept der Lernprogression von WILSON (2009) wird bei der Itemkonstruktion über die schwierigkeitsgenerierenden Merkmale der Kognition und Spezifität eine Construct-Map entwickelt (siehe Tabelle 2). Es wird eine spezifische Profizienz des erworbenen Wissens postuliert. Weiterhin wird angenommen, dass sich diese durch eine zunehmende Wissensbasis nicht nur quantitativ erweitert, sondern auch qualitativ in Form von Entwicklungsstufen manifestiert. Die nächsthöhere Entwicklungsstufe schließt dabei die Wissens- und Fähigkeitsbestände der niedrigeren Stufe mit ein. Ziel ist es dabei auf Grundlage der theoretischen Annahmen zur Entwicklung der ökonomischen Kompetenz und der empirisch zu beobachtenden Itemschwierigkeiten Rückschlüsse für die Testwertinterpretation in Bezug auf das Kompetenzniveau der Testteilnehmenden ziehen zu können.

Tabelle 2: Construct-Map und Itemformate

Items	Nr.	Iteminhalte	Spezifität	Kognition	Aufgabenformat	Beantwortungsformat
FT11	1	Preisberechnung Einkaufszettel, Grundrechenarten	1	2	Auswahl	Single-Choice
FT12	2	Bedürfnisse & Bedarf	2	1	Auswahl	Multiple-Choice
FT13	3	Wirtsch. Unterschied Bio-Produkte/konv. Produkte	3	2	Auswahl	Multiple-Choice
FT14	4	Knappheitskonzept	2	3	Zuordnung	Drag & Drop
FT21	5	Influencer-Marketing	2	1	Zuordnung	Drag & Drop
FT22	6	Wirkung von digitalen Marketingstrategien	2	3	Offenes Format	Kurzaufsatz
FT23	7	Nutzen digitaler Marketingstrategien aus Unt.-Sicht	3	2	Auswahl	Multiple-Choice
FT31	8	Definition Nachhaltigkeit	2	3	Offenes Format	Kurzaufsatz
FT32	9	Facetten von Nachhaltigkeit	3	2	Zuordnung	Drag & Drop

Items	Nr.	Iteminhalt	Spezifität	Kognition	Aufgabenformat	Beantwortungsformat
FT33	10	Fair-Trade-Produkte	1	1	Offenes Format	Ergänzung
FT34	11	Fair-Trade-Konzept	3	2	Auswahl	Multiple-Choice
FT35	12	Informationsquellen zu Produktinformationen	2	1	Auswahl	Multiple-Choice
FT41	13	Berechnung Jahreszinsen	1	2	Auswahl	Single-Choice
FT42	14	Konzept Zinseszins	1	2	Auswahl	Single-Choice
FT43	15	Berechnung unterjährige Zinsen	1	2	Auswahl	Single-Choice
FT44	16	Zölle & Auswirkungen auf Unternehmen	3	2	Auswahl	Multiple-Choice
FT45	17	Gewinnkonzept	2	2	Auswahl	Single-Choice
FT46	18	Kaufkraft	2	2	Auswahl	Multiple-Choice
FT51	19	Zentrale Begriffe Kaufvertrag	2	1	Auswahl	Single-Choice
FT52	20	Prozentrechnen In-App-Kauf (verminderter GW)	1	2	Offenes Format	Wertberechnung
FT53	21	Prozentrechnen In-App-Kauf (Preisdifferenz)	1	2	Offenes Format	Wertberechnung
FT54	22	Gefahren des In-App-Kaufs	2	3	Offenes Format	Kurzaufsatz
FT61	23	Subtraktion Preisunterschied	1	2	Offenes Format	Wertberechnung

Items	Nr.	Iteminhalte	Spezifität	Kognition	Aufgabenformat	Beantwortungsformat
FT62	24	Prozentrechnen Angebotsvergleich (vermehrter GW)	1	2	Offenes Format	Wertberechnung
FT63	25	Ursachen Kostenvorteile für Online-Shopping	3	2	Auswahl	Multiple-Choice
FT64	26	Wirkung von AGBs bei Kaufverträgen	2	1	Auswahl	Single-Choice
FT65	27	Währungsumrechnung	2	2	Auswahl	Single-Choice
FT66	28	Institutionen des Verbraucherschutzes	2	1	Auswahl	Multiple-Choice
FT71	29	Bezahlen mit Kleingeld	2	2	Offenes Format	Kurzaufsatz
FT72	30	Funktionen des Geldes	2	1	Auswahl	Single-Choice
FT73	31	Kaufvertrag rechtswirksam?	2	3	Offenes Format	Kurzaufsatz
FT74	32	Bezahlen mit EC-Karte	2	1	Offenes Format	Kurzaufsatz
FT75	33	Kaufvertrag Botengang	2	3	Offenes Format	Kurzaufsatz
FT81	34	Preisbildung	2	2	Auswahl	Single-Choice
FT82	35	Wirtschaftskreislauf	2	2	Zuordnung	Drag & Drop

Testkonstruktion – Itementwicklung und authentische Testumgebung

Anknüpfend an die von GULIKERS ET AL. (2004) beschriebenen Merkmale eines authentischen Assessments wurde das Testinstrument in seiner Narration als eine „Reise durch den Supermarkt“ entwickelt. Es gliedert sich in acht verschiedene Einheiten, die

jeweils unterschiedliche ökonomische Problemstellungen multiperspektiv adressieren (siehe Tabelle 3). Die technische Umsetzung der Testumgebung erfolgte computergestützt mithilfe des CBA-Itembuilders (KRÖHNE, 2023).

Tabelle 3: Überblick über die Inhalte und adressierten ökonomischen Lebensbereiche des Testinstruments TBA-EL

Situationen	Fachwissenschaftliche Inhalte	Ökonomischer Lebensbereich
„Der Einkaufszettel“	Basales Rechnen, Bedürfnisse & Bedarf, Allokationsprinzip, Einsatz von Produktionsmitteln, Knappheit	PF, BU, GV, NH
„Die Smartwatch“	Einfluss von personalisierter Werbung auf den Kunden (aus Verbraucher- und Unternehmensperspektive)	PF, BU
„Die Projektarbeit“	Konzept der Nachhaltigkeit im ökonomischen Kontext am Beispiel von Fairtrade, Verbraucherinformationsmöglichkeiten	PF, BU, GV, NH
„Wirtschaften in der Welt“	Konzept Zinsen & Zinseszins, Auswirkungen staatlicher Maßnahmen auf heimische Produzenten, Gewinnkonzept, Kaufkraft & Inflation	BU, GV
„Der In-App-Kauf“	Bürgerliches Recht (Geschäftsfähigkeit), Prozentrechnen bei Rabatten, Mobiles Bezahlen, Ver- und Überschuldung bei Jugendlichen	PF
„Die Bluetooth-Kopfhörer“	Betriebswirtschaftliches Verstehen von Kostenvorteilen des Onlinehandels ggü. dem stationären Handel, AGB, Währungsumrechnung, Verbraucherrechte bei Konfliktsituationen	PF, BU
„An der Kasse“	Zahlungsarten, Funktionen des Geldes, bürgerliches Recht (Kaufvertrag), Sicherheit beim elektronischen Bezahlen	PF, GV
„Nach dem Einkauf“	Preisbildung, Wirtschaftskreislauf	GV

PF = Persönlich-finanziell, BU = Beruflich-unternehmerisch, GV = Gesellschaftlich-volkswirtschaftlich, NH=Nachhaltigkeit

Die Schüler:innen begleiten dabei „Kim und Juri“, zwei Jugendliche, durch den Supermarkt. Die Abschnitte werden durch begleitende Videos mit Erzähler eingeführt, um einen kontinuierlichen Erzählstrang zu erzeugen. Jede Einheit enthält zwei bis sechs Items. Das Testinstrument besteht insgesamt aus 35 Items, die ausgewogen die domänenbezogenen Inhalte und kognitiven Anforderungen repräsentieren (siehe Tabelle 2). Das Testinstrument ist in eine technologiebasierte Umgebung eingebettet und erzeugt durch den Einbezug der Probanden in die Narration bspw. durch das gezielte Einfordern von Hilfeleistung für die fiktiven Jugendlichen eine handlungsorientierte Testumgebung. Den Schüler:innen stehen weitere Tools (bspw. ein virtueller Notizblock, ein Taschenrechner und ein Hilfebutton zur Erklärung von Funktionen) zur Verfügung. Hinsichtlich des Itemformats wurden sowohl geschlossene Single-Choice, Multiple-Choice und Drag & Drop Items als auch offene Formate wie das Formulieren von Argumenten oder das begründete Entscheiden verwendet (siehe Abb. 1). Darüber hinaus wurden in einem Begleitfragebogen Kontextfaktoren der Schüler:innen erhoben.

Abbildung 1 zeigt den Aufbau des Testinstruments. Die Testumgebung besteht aus drei verschiedenen Instrumenten. Neben dem Testinstrument zur Messung ökonomischer Kompetenz (TBA-EL), auf dessen der Fokus des Beitrages liegt, werden auch nachhaltigkeitsbezogene Einstellungen (ökonomische, ökologische und soziale Nachhaltigkeitsfacette) im Testinstrument eingebettet erhoben. Das dritte Instrumentarium besteht aus einem Fragebogen zu ökonomischen Kontextfaktoren. Der QR-Code in Abbildung 1 zu einem erklärenden You-Tube-Video des Testinstruments, welches einen Einblick in die Testumgebung und die verschiedenen Itemformate gibt.

Technologiebasiertes Assessment ECON 2022
Zielgruppe: Schüler*innen aus Jahrgang 8

Instrumentarium 1
Test ökonomischer Literalität

Administration:
Authentische Aufgaben in ökonomischen Handlungssituationen

Stichprobe: N=816
Itemanzahl: 35 (8 thematische Units)

Antwortformate:

Single Choice	11 Items
Multiple Choice	8 Items
Drag & Drop	8 Items
Offen	9 Items

Instrumentarium 2
Nachhaltiges Konsumbewusstsein

Administration:
Messung der persönlichen Relevanz zu Nachhaltigkeitsfacetten nach Beendigung einer bezugnehmenden Unit im Test

Stichprobe: N=693
Itemanzahl: 18 (4 Blöcke)

Antwortformat:
Schieberegler von 0% (gar nicht wichtig) bis 100% (sehr wichtig)

Instrumentarium 3
Begleitefragebogen zu ökonomischen Kontextfaktoren

Inhalt:
Items zu Konstrukten wie z. B. Informationsquellen, Lerngelegenheiten, Umgang mit Geld, Selbstwirksamkeit

Stichprobe: N=693
Itemanzahl: (13 Konstrukte)

Antwortformat:

4-Stufige Skalen	10 Items
5-stufige Skalen	2 Items
6-stufige Skala	1 Item

Abbildung 1: Überblick über das Assessment ECON 2022 und ein exemplarisches Item (Auswirkungen eines Importzolls), eigene Darstellung

Das exemplarische Item in Abbildung 1 entstammt der Einheit 4 „Wirtschaften in der Welt“. Im einführenden Video zur Einheit wird den Schüler:innen ein Blogbeitrag einer ecuadorianischen Fair-Trade-Kaffeeproduzentin vorgestellt und knüpft in der Narration an die vorherige Unit an (Merkmal 2). Das Item thematisiert die Auswirkungen der Einführung eines Importzolls auf den Import von Maschinen aus dem Ausland und die Auswirkungen auf die heimischen Kaffeeproduzenten und adressiert damit zum einen den gesellschaftlich-volkswirtschaftlichen als auch den beruflich-unternehmerischen Lebensbereich. Die Multiperspektivität spiegelt sich im Lösungsraum des Items wider. Von den vier Antwortmöglichkeiten sind zwei korrekt. Jede richtige Antwort bezieht sich auf einen der adressierten Lebensbereiche des Domänenmodells (Merkmal 1 und 4). Aufgrund der Notwendigkeit zur Lösung der Aufgabenstellung, Wissensbestände aus mehreren ökonomischen Teilbereichen aktivieren zu können, ist das Item dem Anforderungsniveau 3 der Spezifität zuzuordnen. Kognitionsbezogen adressiert das Item konzeptionelles Wissen mit dem Ziel des Verstehens und der Analyse von Auswirkungen staatlicher Maßnahmen auf wirtschaftliche Akteure und ist als Anforderungsniveau 2 in der Kategorie Kognition kategorisiert. Soziale Prozesse (Merkmal 3) werden in der Testumgebung nicht berücksichtigt, da sie für das Demonstrieren ökonomischer Kompetenz in diesem Kontext nicht bedeutsam sind.

Die Bewertung der entwickelten Testaufgaben bezogen auf die schwierigkeiten generierenden Merkmale erfolgte sowohl durch die Testentwickler:innen als auch durch N=25 Expert:innen mit Expertise in Schule / Bildungsadministration, Fachexperten und/oder Testentwicklung. Ferner wurden auch die Adäquanz der Aufgabenstellung und der verwendeten Distraktoren für die Jahrgangsstufe 8 erfragt. Zu jedem Item konnten die Expert:innen einen schriftlichen Kommentar verfassen, der die obigen Bewertungen weiter ausführt und Hinweise für Optimierungen geben kann. Die Ergebnisse der Expert:innenbefragung sind in den weiteren Testentwicklungsprozess eingeflossen (siehe hierzu WELSANDT ET AL., SUBMITTED, 2023).

3.2 Datenerhebung und Datenanalyse

Die Feldtesterhebung in ECON 2022 erfolgte in einem Querschnittsdesign und wurde zwischen April und Juni 2021 an 20 verschiedenen Schulen in Nordrhein-Westfalen in der Jahrgangsstufe 8 computergestützt durchgeführt. Die Erhebung fand in Kooperation mit der International Civic and Citizenship Education Study (ICCS) statt. Insgesamt nahmen 816 Schülerinnen und Schüler an der Erhebung teil. Diese wurde mithilfe eines externen Dienstleisters (International Association for the Evaluation of Educational Achievement) administriert und dauerte pro Sitzung eine Stunde exklusive des Begleitfragebogens. Die Daten zu den personenbezogenen Merkmalen der Schüler:innen wurden in der ICCS-Befragung erhoben und konnten durch eine Zuordnung mithilfe einer individuellen Teilnehmernummer dem ECON-Datensatz exakt zugeordnet werden. Die Erhebung basiert auf einer stratifizierten Zufallsstichprobe. Die Strata berücksichtigen im Feldtest den Anteil der Schüler:innen mit Migrationshintergrund sowie die Schulform. Im Mittel waren die Teilnehmenden 14,5 Jahre alt. Bezogen auf die Gesamtpopulation ist das Geschlechterverhältnis ausgewogen verteilt. Der Anteil an Schüler:innen mit Migrationshintergrund ist in der Gesamtpopulation mit 40 % (Stand Juli 2022) höher als in der Stichprobe (siehe Tabelle 4). Bei der Schulformzugehörigkeit sind Schüler:innen der Hauptschule überproportional vertreten (MSB, 2022).

Tabelle 4: Beschreibung der Stichprobe

Merkmale		Stichprobe			Gesamt-population
		N	Fehlende Werte	In Prozent	In Prozent
Geschlecht		816		-	-
	Weiblich	396	-	48,40 %	48,39 %
	Männlich	412		50,50 %	51,61 %
	Divers	9		1,10 %	

Merkmale		Stichprobe		Gesamt-population	
		N	Fehlende Werte	In Prozent	In Prozent
Schulform	Hauptschule	141	-	17,30 %	5,10 %
	Schulart mit mehreren Bildungsgängen	76		9,30 %	5,70 %
	Realschule	189		23,10 %	21,14 %
	Gesamtschule	176		21,50 %	29,56 %
	Gymnasium	234		28,60 %	33,42 %
Migrationshintergrund	Kein Migrationshintergrund	431	121	52,80 %	60,00 %
	Migrationshintergrund	265		32,40 %	40,00 %
Familiäre Verwendung der Landessprache	Keine familiäre Verwendung der Landessprache	203	109	25 %	-
	Familiäre Verwendung der Landessprache	504		62 %	
Sozioökonomischer Hintergrund (SÖH)	Niedriger SÖH	161	112	20 %	
	Mittlerer SÖH	400		49 %	-
	Hoher SÖH	143		18 %	

Die Testzeit betrug 60 Minuten zuzüglich der Bearbeitung eines Begleitfragebogens im Anschluss. Die Schülerantworten wurden nach einem detaillierten Scoring-Guide gecodet. Offene Antwortformate wurden von drei Ratern bearbeitet. Die Interrater-Reliabilität wurde mit Krippendorffs Alpha ermittelt und liegt zwischen 0,63 und 0,84 je nach Item. Es wurde auf Test- und Itemebene auf Interrater-Effekte untersucht (LINACRE, 1994), dabei konnten keine signifikanten Auswirkungen einzelner Rater auf die Itemschwierigkeit festgestellt werden ($\text{Chi-Quadrat}=3,15$; $\text{df}=2$; $p=.207$). Auf Itemebene konnte nur bei einem Item ein signifikanter Unterschied zwischen zwei Ratern festgestellt werden, das einen DIF über 0,426 aufweist. Fehlende Schülerantworten wurden als „falsche Antworten“ gewertet, da ein ungelöstes Item in einem Leistungstest als „nicht wissen“ interpretiert wurde (POHL ET AL., 2014).

Für die Analyse der Daten wird ein polytomes IPL-IRT-Modell, Multidimensional-random-Coefficients-Multinomial-Logit-Modell (MCMLM) (ADAMS ET AL., 1997) gewählt und mit dem Programm ACER ConQuest (ADAMS ET AL., 2018) skaliert. Bei der

Datenerhebung konnte aufgrund eines Erfassungsfehlers der Testsoftware beim ersten Item FT11 die Schülerantworten nicht reliabel zu den dargebotenen Antwortoptionen des Items zugeordnet werden, sodass Item FT11 von der Analyse ausgeschlossen werden musste. In die Analyse der Feldtestdaten können somit nur 35 der 36 Items berücksichtigt werden. Tabelle 5 stellt übersichtlich dar, welche Analyseverfahren für das Bestimmen der psychometrischen Qualität des Testinstruments verwendet wurden.

Die Beantwortung der ersten Forschungsfrage bezieht sich auf die Überprüfung der psychometrischen Qualität des Testinstruments. Hierfür werden zunächst die Personen- als auch die Itemparameter wie auch die Messgenauigkeit bestimmt. Dazu werden die Parameter mit der Marginal-Maximum-Likelihood-Methode (MML) geschätzt (ADAMS ET AL., 1997). Als Parameter werden die Itemschwierigkeit sowie die Personenfähigkeitsschätzer (WLE) und deren Verteilung ermittelt. Darüber hinaus wird mittels einfaktorieller Varianzanalyse (ANOVA) geprüft, ob sich die Itemschwierigkeit hinsichtlich der Aufgabentypen und Inhaltsbereiche unterscheiden (MOOSBRUGGER/KELAVA, 2012, 2020; WINTHER, 2010).

Die Messgenauigkeit des Tests kann anhand des Standardfehlers der einzelnen Personenparameter sowie der Reliabilitätskoeffizienten der probabilistischen und klassischen Testtheorie geprüft werden. Das zu messende Konstrukt gilt als zuverlässig schätzbar, wenn (1) die Standardfehler der Personenparameter gering und (2) die Reliabilitätskoeffizienten hoch sind (EAP/PV & $WLE \geq .70$; Cronbachs $\alpha \geq .70$) (DÖRING/BORTZ, 2016; FREY, 2012). Die präzise Schätzung der Personenfähigkeiten ist von unmittelbarer Bedeutung für die valide Testwertinterpretation (AERA ET AL., 2014). Darüber hinaus wird mit der Person-Separation-Reliabilität (WLE) geprüft, ob die Reproduzierbarkeit der Personenparameter gewährleistet ist. Zudem soll mit der Messung der Item-Separation-Reliabilität untersucht werden, ob der Test tatsächlich zwischen leichten und schwierigen Items unterscheiden kann. Ebenfalls soll mit der Bestimmung der testcharakteristischen Kurve (TCC) untersucht werden, ob ein Zusammenhang zwischen den summierten Personen-Testwerten und der latenten Personenfähigkeit existiert (ROST, 2004). Die TCC wird konzeptionell als die Regression der summierten Antwortscores der Testteilnehmer verstanden und kann grafisch als die Summe aller itemcharakteristischen Kurven (ICC) betrachtet werden. Der empirische Zusammenhang muss einen streng monotonen, hohen Zusammenhang aufweisen.

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

Tabelle 5: Analyseebenen und methodisches Vorgehen

Analyseebene	Itemparameter	Personenparameter	Reliabilität	Itemhomogenität
Analysemethoden	Personen-Item-Map		Kennwerte der IRT	Itemfit-Werte
	Verteilung der Itemschwierigkeiten	Verteilung der Personenparameter	Kennwerte der KTT	Grafische Analyse (ICC)
	Itemschwierigkeit nach Aufgabentyp	Testcharakteristische Kurve (TCC)	Testinformationskurve (TIF)	DIF-Analysen
	Itemschwierigkeit nach Inhaltsbereich			

Für die Beantwortung von Forschungsfrage 2 wird mithilfe des Differential-Item-Functioning- (DIF) geprüft, ob bei gleicher Personenfähigkeit Unterschiede in der Lösungswahrscheinlichkeit von Items hinsichtlich eines personenbezogenen Merkmals bestehen. Dies dient zur Ermittlung der Testfairness über verschiedene Subgruppen hinweg (PAEK, 2002; TERESI ET AL., 2008). Zunächst wird auf Testebene geprüft, ob signifikante Gruppenunterschiede hinsichtlich der Hintergrundvariablen zu finden sind. Im zweiten Schritt wird der Interaktionsterm auf Signifikanz geprüft, darüber hinaus wird auf Itemebene untersucht, ob der DIF-Schätzer einzelner Items einer Subgruppe sich signifikant von der absoluten Itemschwierigkeit unterscheidet (PAEK, 2002). Hierfür wurde mittels des Wald-Tests ein Chi-Quadrat-Wert für einen Freiheitsgrad ermittelt, um so auf Itemebene Signifikanz ermitteln zu können (KIRSCH, 2021). Signifikante DIF-Unterschiede können als Indiz für das Verletzten der Testfairness gewertet werden. Für die DIF-Analysen werden die Merkmale Geschlecht, der Migrationshintergrund, der sozioökonomische Hintergrund als auch die zu Hause gesprochene Sprache der Schülerinnen und Schüler verwendet. Der Umgang mit Verletzung der Testfairness durch DIF-Effekte in einzelnen Subpopulationen einer Stichprobe wird in der Literatur kontrovers diskutiert. In der Forschungslandschaft gibt es keine allgemein anerkannten Grenzwerte für die Tolerierbarkeit von DIF-Effekten. Vielmehr existieren unterschiedliche Klassifizierungsschemata, die DIF-Effekte zu kategorisieren versuchen. In diesem Artikel wird sich auf das Klassifizierungsschema von PAEK/WILSON (2011) bezogen, das DIFs nach der Stärke und ihrer statistischen Signifikanz bewertet (PAEK/WILSON, 2011). Dabei stellen DIF-Effekte der Kategorie A ($|DIF| < 0,426$ oder $p > 0,05$) keine problematischen Items dar, während der Einsatz von Items der Kategorie B ($0,426 < |DIF| < 0,628$; $p < 0,05$) und Kategorie C ($|DIF| > 0,628$; $p < 0,05$) als begründenswert zu bewerten ist. Darüber hinaus ist zu bedenken, dass nicht jeder DIF-Effekt eine Verletzung der Testfairness darstellen muss, so kann bspw. ein DIF-Effekt bezogen auf das Vorwissen von Schüler:innen zu einer Thematik wünschenswert sein, da das Testinst-

rument somit instruktionssensitiv reagiert. Daher ist eine notwendige Elimination von Items aufgrund von DIF-Effekten nicht zwingend ratsam und sollte auch aus Perspektive der Abwägung von Konstruktvalidität entschieden werden. Auf Grundlage der Personen- und Itemparameter sowie der Informationen zur Itemhomogenität werden auffällige Items vom Text unter Berücksichtigung von Überlegungen zur Inhaltsvalidität und Reliabilität des Testinstruments exkludiert oder überarbeitet.

3.3 Beschreibung der personenbezogenen und institutionellen Merkmale

Der sozioökonomische Hintergrund wird über den höchsten Bildungsabschluss der Eltern und der Anzahl an Büchern im Haushalt gebildet und ordinal skaliert. Die Klassifizierung des Bildungsabschlusses erfolgt über den International Standard Classification of Education (ICES-Level). Die neun ISCED-Level werden in Anlehnung an den AES (2020) in drei Kategorien gruppiert (1) Personen bis einschließlich Sekundarbereich 1 (ISCED 0–2), (2) Personen mit Erreichen des Sekundarbereichs 2 und postsekundärer, nicht tertiärer Bildung (ISCED 3 und 4) und (3) Personen mit mindestens einem Abschluss auf dem Niveau eines kurzen tertiären Bildungsprogramms (ISCED 5–8). Die Anzahl der Bücher eines Haushalts ist ebenfalls ordinal mit drei Ausprägungen codiert. Mittels konfirmatorischer Faktorenanalyse wurde festgestellt, dass beide Variablen auf einen Faktor laden, sodass die Variablen zusammengefasst werden können (Bartlett-Test: Chi-Quadrat (1) = 58,715, $p < .001$; KMO = 0,500; AVE bei einem Faktor = 0,647). Die Variable Schulform wird ebenfalls ordinal skaliert. Der Migrationshintergrund ist nominal skaliert und schließt Personen ein, bei denen mindestens ein Elternteil oder die Testperson selbst im Ausland geboren ist. Das Merkmal der familiären Verwendung der Landessprache definiert, ob Deutsch im häuslichen Umfeld die häufigste gesprochene Sprache darstellt oder nicht. Unter Geschlecht werden die Merkmale männlich, weiblich und divers verstanden, wobei das Merkmal divers von keiner Testteilnehmer:in gewählt wurde.

4. Ergebnisse

4.1. Ergebnisse zu Forschungsfrage 1 – Psychometrische Eigenschaften des Testinstruments

Die Befunde zu Forschungsfrage 1 beziehen sich auf die Analyse der Item- und Personenparameter, der Reliabilität sowie der Itemhomogenität. Eine optimale Skalierung eines Testinstruments ist gegeben, wenn die Parameter des Modells und die dazugehörigen Items (1) ausreichend weit auf der Logit-Skala streuen und (2) eine möglichst hohe Varianz der Logit-Werte aufweisen.

Die durch die MCMLM-Methode geschätzten Personen- und Itemparameter werden auf eine gemeinsame Logit-Skala transformiert (Personen-Item-Map). Die Personen-

Item-Map (Wright Map) zeigt die Anzahl der Fälle des Personenparameters und die Itemparameter (siehe Abbildung 2).

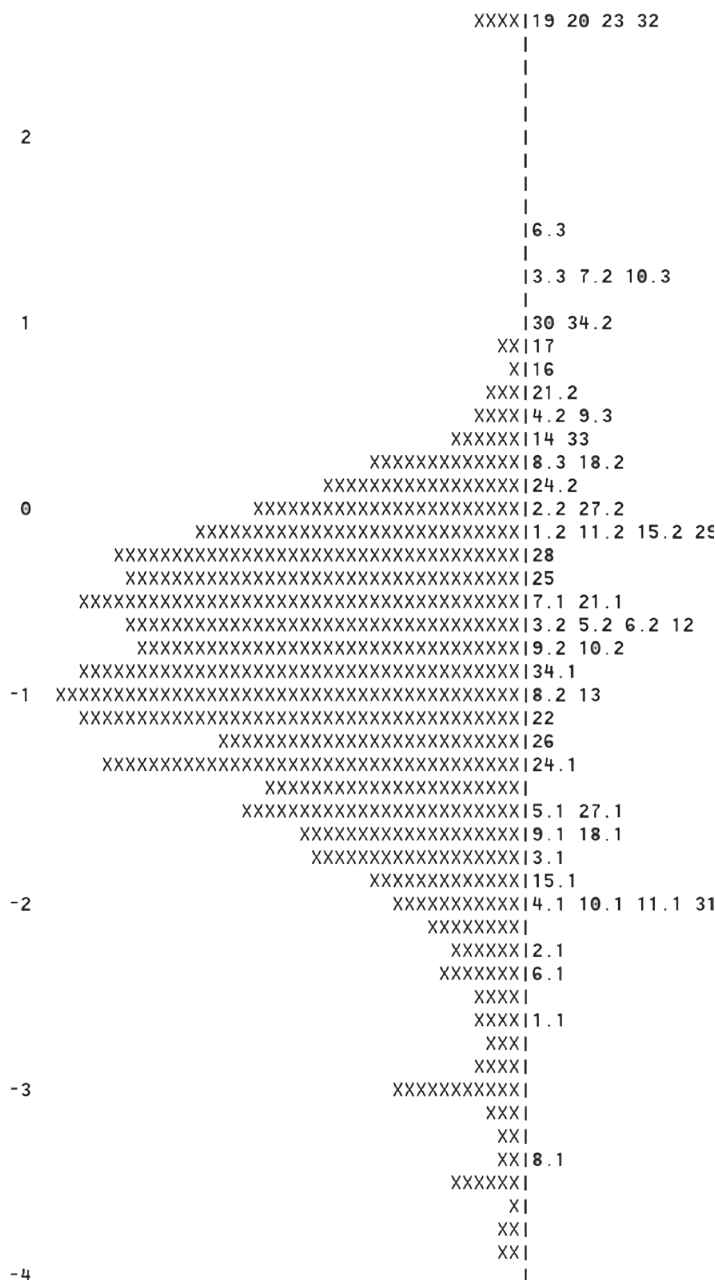


Abbildung 2: Personen-Item-Map für die Feldteststichprobe des Testinstruments TBA-EL, eigene Darstellung

Die Itemschwierigkeitsparameter streuen im Bereich von $-2,024$ und $6,106$ mit einer Spannweite von $8,130$. Der Mittelwert der Itemschwierigkeitsparameter ist aufgrund der Modellspezifikation auf 0 fixiert ($M=0$; $SD=1,591$). Der Kolmogoroff-Sminorv-Test zeigt zunächst keine Normalverteilung der Itemparameter an ($K-S=0,204$; $df=34$, $p<0,001$). Bei Exkludierung des Ausreißeritems $FT75$ mit der Itemschwierigkeit von $6,106$ Logits kann jedoch die Normalverteilung angenommen werden ($K-S=0,178$; $df=33$; $p=0,10$). Aus der Wright Map kann geschlossen werden, dass 22 von 34 Items eine negative Itemschwierigkeit aufweisen und somit als „eher einfach“ zu werten sind. Die 12 Items im positiven Logit-Bereich sind hingegen als „eher schwierig“ zu klassifizieren. Eine Betrachtung der Itemschwierigkeiten nach Aufgabentyp zeigt in einer ANOVA, dass keine signifikanten Gruppenunterschiede hinsichtlich der Aufgabentypen Single-Choice, Multiple-Choice und eines offenen Antwortformats bestehen ($F(3,164)=31,64$; $p=0,056$). Ebenfalls wurden keine signifikanten Gruppenunterschiede in der Itemschwierigkeit hinsichtlich der Inhaltsbereiche des Domänenmodells gefunden ($F(0,195)=1,95$; $p=0,824$). Somit kann geschlossen werden, dass bei der Lösung der Items mögliche Effekte durch die Komplexität des Aufgabenformats zu vernachlässigen sind. Darüber hinaus kann gezeigt werden, dass für die Inhaltsbereiche des Domänenmodells eine ausgewogene Verteilung hinsichtlich der inhaltlichen Spezifität der Aufgaben gegeben ist.

Die Personenparameter (wLE) streuen zwischen den Werten $-5,647$ und $0,960$ mit einer Spannweite von $6,607$ Logits. Der Mittelwert des Personenfähigkeitsparameters beträgt $-1,053$ mit einer Standardabweichung von $1,074$ Logits. Die Verteilung der Personenparameter ist somit linksschief und nicht normalverteilt ($K-S=0,127$; $df=816$; $p<0,001$). Die Nichtnormalverteilung des Personenfähigkeitsparameters könnte mit Probanden zusammenhängen, die hohe fehlende Werte beim Antwortverhalten aufzeigen. So haben 58 von 816 Schüler:innen mehr als die Hälfte der Items nicht bearbeitet. Eine Linksverschiebung des Personenfähigkeitsparameters bleibt auch bei Ausschluss der 58 Probanden bestehen, sodass dies als Indiz dafür interpretiert werden kann, dass der Test womöglich etwas zu schwer konzipiert wurde. Für den Feldtest kann ein sehr hoher, monotoner, nicht linearer, s-förmiger Zusammenhang zwischen den Personen-Testwerten und den Personenparametern festgestellt werden ($r=0,957$; $\alpha<0,001$).

Im Testinstrument nehmen die Standardfehler der Personenparameter geringe Werte an ($M=0,341$; $SD=0,144$) und liegen im Logit-Bereich von $0,292$ bis $1,473$. Die Schätzung der Personenparameter ist für den Logit-Bereich zwischen $-0,568$ und $0,617$ signifikant ($p < 0,05$). Die Standardfehler sind für den Bereich zwischen $-1,88$ Logits bis $0,95$ Logits am geringsten. Die Personenfähigkeiten kleiner 2 Logits werden weniger zuverlässig geschätzt. Diese deckt sich auch mit der grafischen Beurteilung der Testinformationsfunktion. Die maximale Testinformation kann im Logit-Bereich zwischen -2 und 1 beobachtet werden. Die Messgenauigkeit des Testinstruments für die globalen Personenparameter kann über den Reliabilitätskoeffizient der IRT (EAP/PV) und der KTT (Cronbachs Alpha) beurteilt werden. Die Ergebnisse der globalen Reliabilitätsparameter konstatieren für alle Kennzahlen Werte deutlich über $0,70$ an, was für eine hohe Reliabilität des Testinstruments spricht (siehe Tabelle 6). Die Höhe des wLE-Koeffizienten ($0,879$) lässt darauf

schließen, dass das Testinstrument zuverlässig zwischen fähigen und weniger fähigen Personen unterscheiden kann.

Tabelle 6: Probabilistische- und klassische Reliabilitätskennwerte

Probabilistische Kennwerte	Personenbezogen	EAP/PV	0,962
		MLE	0,879
		WLE	0,881
	Itembezogen	Item-Separation-Reliabilität	0,985
Klassische Testtheorie		Cronbachs Alpha	0,860

Zusammenfassend kann somit festgestellt werden, dass es eine ausreichende empirische Evidenz hinsichtlich der Messgenauigkeit des Instruments für das Konstrukt der ökonomischen Literalität gibt.

Die Annahme eines Partial-Credit-Rasch-Modells ist erfüllt, wenn (1) die Item-Infit-Werte (wMNSQ) idealerweise dem Erwartungswert 1 entsprechen, als auch signifikant sind ($|t| < 1,96$ bzw. $p < 0,05$) und (2) die Schwellenparameter der Antwortkategorien aufsteigend angeordnet sind. Für den Cut-Off-Wert für den Item-Infit schlagen ADAMS UND KHOO (1996) einen Wertebereich zwischen 0,75 und 1,33 vor, währenddessen in Large-scale Assessments wie PISA wMNSQ-Werte zwischen 0,85 und 1,15 als angemessen betrachtet werden (ADAMS/KHOO, 1996; KASTBERG ET AL., 2021; MOOSBRUGGER/KELAVA, 2012, 2020; OECD, 2014, 2019B, 2020).

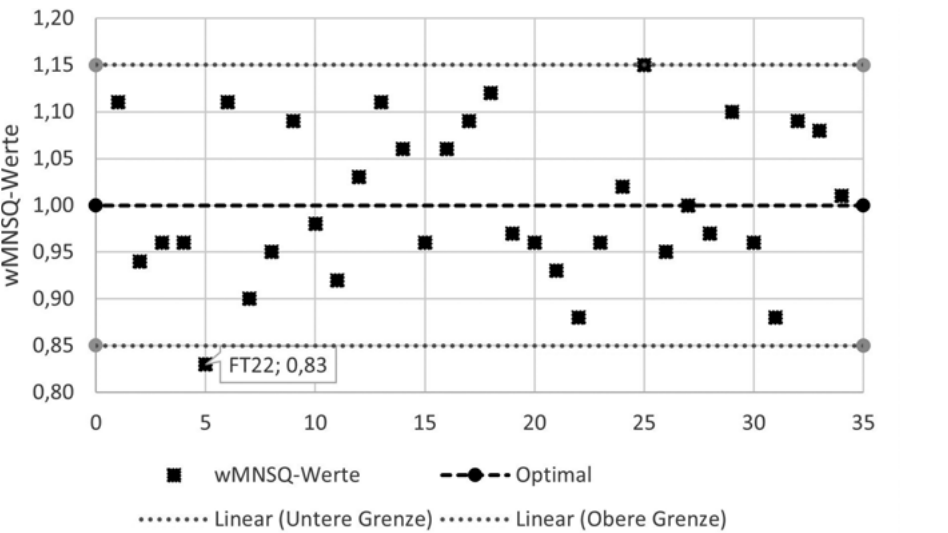


Abbildung 3: Itemparameter – wMNSQ der Testitems des Testinstruments TBA-EL, eigene Darstellung

34 von 35 Items des Testinstruments erfüllen den strengen WMNSQ-Wertebereich, lediglich Item FT22 weist einen Overfit auf (siehe Abbildung 3). Die T-Werte streuen in einem Bereich zwischen -5,20 und 5,4 (siehe Abbildung 4). Vier Items weisen einen T-Wert kleiner -1,96 auf, was auf eine zu hohe Trennschärfe spricht und als eher nicht problematisch betrachtet wird. Bei sechs Items kann ein T-Wert größer 1,96 festgestellt werden, was auf eine signifikante Abweichung und eine niedrige Trennschärfe schließen lässt. Für eine präzisere Beurteilung des Item-Infit wird auch die Trennschärfe der klassischen Testtheorie berücksichtigt, die nicht unter dem Wert von 0,20 liegen sollte. Items innerhalb des WMNSQ-Intervalls von 0,85 bis 1,15 wiesen Trennschärfen von größer 0,20 auf.

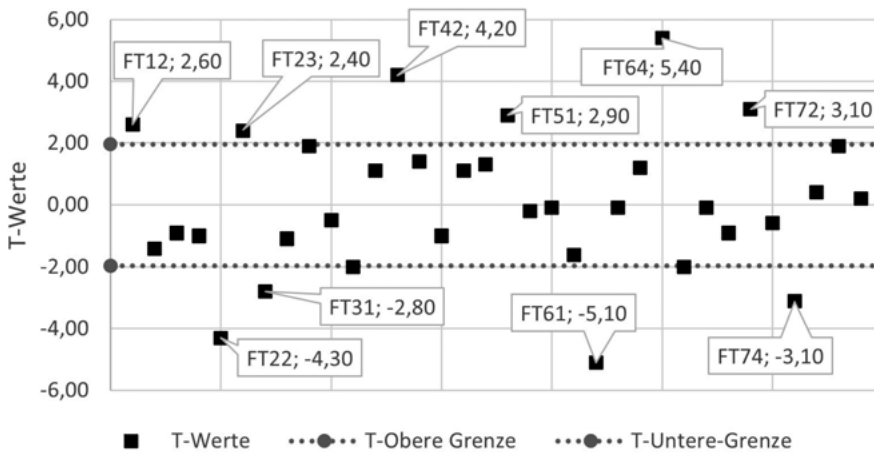


Abbildung 4: Itemparameter – T-Werte der WMNSQ des Testinstruments TBA-EL, eigene Darstellung

Fünf von sechs Items mit einem T-Wert größer 1,96 haben Trennschärfen von genau 0,20 (siehe Abb. 5). Nur Item FT64 zeigt eine Trennschärfe von kleiner 0,20 auf (0,18). Für die Hauptstudie lässt sich daher schließen, dass eine Prüfung der Distraktoren in besagten Items notwendig sein kann, um eine sprachlich oder inhaltlich bessere Abgrenzung der Antwortoptionen zu gewährleisten. Hinsichtlich der Schwellenparameter weisen lediglich zwei von 17 polytom codierten Items ungeordnete Schwellenparameter auf. Für beide betroffenen Items (FT23 und FT51) scheint die nur teilweise richtige Antwortoption für die Probanden schwerer zu erkennen zu sein als die vollständig richtige.

Mithilfe der Item-Characteristic-Curves (ICC) können die Items eines WMNSQ-Wertes auch grafisch analysiert werden. Idealerweise sollten die Kurven s-förmig verlaufen, da so gewährleistet ist, dass ein Item zwischen fähigen und weniger fähigen Testteilnehmern unterscheiden kann und der Informationsgehalt eines Items in Bezug auf die unterstellte Kompetenz am höchsten ist (MOOSBRUGGER/KELAVA, 2012, 2020).

Bei der grafischen Analyse der Items konnten bei vier Items sehr flache ICCs festgestellt werden. Alle Items haben Itemschwierigkeitswerte größer 2 Logits.

4.2 Ergebnisse zu Forschungsfrage 2 – Fairness des Testinstruments

Die Befunde zu Forschungsfrage 2 beziehen sich auf die Analyse der Testfairness. Hierzu wird untersucht, ob das Testinstrument gruppenspezifisch hinsichtlich auf persönliche Merkmale diskriminiert. Ziel sollte es daher sein, DIF-Effekte bei der Konstruktion eines Testinstruments möglichst gering zu halten. Leistungsunterschiede sollten durch die unterschiedlichen Personenfähigkeiten der Probanden erklärt werden können und nicht durch die Zugehörigkeit zu einer spezifischen Subpopulation. Nach der Klassifizierung von PAKK/WILSON (2011) sollte bei Items, die einen DIF-Effekt größer 0,426 aufweisen, in der Testrevision begründet werden, weshalb diese weiterverwendet oder gegebenenfalls modifiziert werden, um den DIF-Effekt zu verringern.

Auf Testebene zeigt die DIF-Analyse keine signifikanten Geschlechts- und Sprachunterschiede (siehe Tabelle 7). Beim Migrationshintergrund als auch beim sozioökonomischen Hintergrund konnte ein signifikanter DIF-Effekt festgestellt werden. Der DIF-Effekt beim Migrationshintergrund mit 0,282 als auch beim sozioökonomischen Hintergrund mit 0,429 ist als gering einzuschätzen.

Tabelle 7: DIF-Unterschiede der Subgruppen auf Testebene

Subgruppen	Z	DIF-Schätzer	Standardfehler	Chi-Quadrat	P
Kein MIG	1	0,142	0,053	7,27(1)	0,007
Mit MIG	2	-0,142	0,053		
Männlich	1	0,009	0,043	0,04(1)	0,8415
Weiblich	2	-0,009	0,043		
Keine familiäre Verwendung der Landessprache	1	-0,088	0,051	2,95(1)	0,0859
Familiäre Verwendung der Landessprache	2	-0,088	0,051		
Niedriger SÖH	1	-0,258	0,070	15,08(2)	0,001
Mittlerer SÖH	2	0,087	0,071		
Hoher SÖH	3	0,171	0,097		

Auf Itemebene wurde für den Interaktionsterm Item*Geschlecht kein signifikanter DIF-Unterschied festgestellt (Chi-Quadrat (df) = 32,18 (32), $p = 0,458$), sodass davon auszu-

gehen ist, dass das Testinstrument in Hinblick auf das Geschlecht nicht diskriminiert. Für die Merkmale Migrationshintergrund (Chi- Quadrat (df) = 55,20 (32), $p = 0,007$), familiäre Verwendung der Landessprache (Chi- Quadrat (df) = 79,18 (32), $p < 0,001$) als auch beim sozioökonomischen Hintergrund (Chi- Quadrat (df) = 104,95 (64), $p < 0,001$) konnten hingegen signifikante DIF-Unterschiede festgestellt werden. Zur Berechnung des DIF-Effekts wurde der DIF-Schätzer pro Item verdoppelt. In der Gesamtschau konnten 12 unterschiedliche Items identifiziert werden, die einen signifikanten DIF-Effekt über 0,426 aufweisen. Davon haben sechs Items bei mehr als einem personenbezogenen Merkmal einen DIF-Effekt (siehe Tab. 6). Hinsichtlich des Merkmals „Migrationshintergrund“ zeigen vier Items einen DIF-Effekt der Kategorie B und zwei Items der Kategorie C. Bezogen auf das Merkmal der familiären Verwendung der Landessprache kann bei drei Items einen DIF-Effekt der Kategorie B und bei zwei Items der Kategorie C festgestellt werden. Beim Merkmal des sozioökonomischen Hintergrundes werden drei Gruppen miteinander verglichen, sodass der DIF-Effekt eines Items hinsichtlich seiner Stärke je nach Gruppenvergleich variieren kann (siehe Tabelle 8).

Tabelle 8: DIF-Effekte der Subgruppen auf Itemebene

Items	Nr.	Iteminhalt	Absolute Item- schwie- rigkeit der Stichprobe	Migrati- onshinter- grund	Familiäre Verwen- dung der Landes- sprache	Sozioökono- mischer Hintergrund
FT13	3	Wirtsch. Unterschied Bio- Produkte/konv. Produkte	-1,017			B+**
FT31	8	Definition Nachhaltigkeit	0,548		B+*	B-**
FT45	17	Gewinnkonzept	1,091	B+*	B-**	B-/C-**
FT46	18	Kaufkraft	1,178			B-/C-**
FT52	20	Prozentrechnen In-App- Kauf (verminderter GW)	2,959	B+*		C-**
FT53	21	Prozentrechnen In-App- Kauf (Preisdifferenz)	3,463	B-*		
FT54	22	Gefahren des In-App-Kaufs	0,244			B+**
FT62	24	Prozentrechnen Angebots- vergleich (vermehrter GW)	2,759	C+**		

Items	Nr.	Iteminhalt	Absolute Item- schwie- rigkeit der Stichprobe	Migrati- onshinter- grund	Familiäre Verwen- dung der Landes- sprache	Sozioökono- mischer Hintergrund
FT64	26	Wirkung von AGBs bei Kaufverträgen	-0,288	B+**	C-**	
FT65	27	Währungsumrechnung	-1,087			B+**
FT71	29	Bezahlen mit Kleingeld	0,022			B+**
FT73	31	Kaufvertrag rechtswirksam?	1,282	C-*	B+**	
FT81	33	Preisbildung	0,595		C-**	B-/C-**

B = DIF-Effekt > 0,426; C = DIF-Effekt > 0,538; * $p < 0,05$; ** $p < 0,001$

5. Diskussion

5.1 Psychometrische Qualität des Testinstruments TBA-EL

In diesem Beitrag wird das authentische Testinstrument TBA-EL für Schülerinnen und Schüler der Sekundarstufe I vorgestellt. Ziel dieses Artikels ist es, zum einen die Qualität der psychometrischen Eigenschaften des Testinstruments zu untersuchen, um fundierte Entscheidungen bezüglich möglicher Verbesserungen für die repräsentative Hauptstudie vornehmen zu können. Zum anderen sollte herausgestellt werden, welchen Einfluss ein möglichst authentisches Testdesign auf Gruppenunterschiede entwickeln kann. Dies adressiert den in der Literatur konstatierten persistenten Einfluss der personenbezogenen Merkmale auf die Testleistung von Schüler:innen – insbesondere in den Testinstrumenten der ökonomischen Bildung. Vor diesem Hintergrund wird geprüft, inwieweit ein authentisches Testdesign dazu beiträgt, etwaige Unterschiede zu minimieren.

Hinsichtlich der psychometrischen Qualität des Testinstruments TBA-EL können in Bezug auf die Item- und Personenparameter zufriedenstellende Werte festgestellt und hohe Reliabilitätschätzer erzielt werden. Die Item-Infit-Werte liegen bei 34 von 35 Items innerhalb der strengeren PISA-Grenzwerte (siehe Abb. 4) und haben zumeist hohe Trennschärfen. In Bezug auf den Einsatz des Testinstruments in der Hauptstudie werden daher auffällige Items in ihrer Darstellung vereinfacht und die verwendeten Dis-traktoren sprachlich präzisiert. Positiv zu bewerten ist die hohe EAP/PV-Reliabilität von 0,962 und die zuverlässige Schätzung der Personenfähigkeiten (WLE: 0,881). Das Test-instrument streut in Bezug auf die Itemschwierigkeit ausreichend breit auf der Logits-

kala, sodass von einer adäquaten Verteilung zwischen leichten und schwierigen Items ausgegangen werden kann. Die linksschiefe Verteilung der Personenfähigkeiten zeigt jedoch an, dass das Instrument für eine Teilpopulation eine insgesamt zu hohe Schwierigkeit besitzt (siehe Abb. 3). Ursächlich für die linksschiefe Verteilung sind die Items mit hohem mathematischem Anteil. Bezogen auf die erste Forschungsfrage kann in der Gesamtschau geschlossen werden, dass das verwendete Testinstrument zur Messung ökonomischer Kompetenz in der Jahrgangsstufe zufriedenstellende psychometrische Eigenschaften aufweist und dass sich ökonomisch Testinhalte identifizieren lassen, die für die Hauptstudie in den Schwierigkeiten zu reduzieren sind, um das Testinstrument besser auf die Gesamtpopulation zu mappen.

5.2 Testfairness und Itemrevision

Bezogen auf die Testfairness wurden DIF-Analysen mit den Merkmalen Geschlecht, Migrationshintergrund, die familiäre Verwendung der Landessprache und des sozioökonomischen Hintergrundes durchgeführt. Die Ergebnisse zeigen hinsichtlich des Geschlechts keinerlei signifikanten DIF-Effekt an. Dies gilt ebenfalls bei der Analyse der Testleistung zwischen männlichen und weiblichen Teilnehmenden. In der Literatur sind zumeist Geschlechtereffekte zuungunsten weiblicher Teilnehmer festzustellen (KAISER ET AL., 2020; LUSARDI/MITCHELL, 2014; LUSARDI/TUFANO, 2015; OBERRAUCH, 2019; OBERRAUCH/BRAHM, 2022; OBERRAUCH/KAISER, 2020; OECD, 2014; G. SEEBER ET AL., 2022; WUTTKE ET AL., 2020). Die benutzten Testinstrumente sind jedoch dabei zumeist Abwandlungen des WBTS oder nutzen Items aus PISA Financial Literacy. Diese Testinstrumente bestehen hauptsächlich aus Single- und Multiple-Choice-Items und sind nicht in authentische Assessmentumgebungen eingebunden. Das Fehlen eines geschlechterspezifischen Effekts beim TBA of Economic Literacy lassen sich möglicherweise darauf zurückführen, dass (1) die breite Variabilität in den Aufgaben- und folglich in den Antwortformaten spezifische geschlechtsspezifische Präferenzen in den Aufgabenbearbeitungen nivelliert und dass (2) durch authentische Handlungssituationen und deren narrative Anker verschiedene Identifikationsmöglichkeiten mit den Testinhalten geschaffen werden. Insgesamt sind die Studien bezüglich des Einflusses des Formats eines Items auf die Testleistung widersprüchlich (ACKERMANN/SIEGFRIED, 2019; HAPP ET AL., 2019; HAPP ET AL., 2021; SIEGFRIED/WUTTKE, 2019): Einige Studien zeigen, dass die gewählten Itemformate und deren Verteilung Geschlechtereffekte reduzieren können, während sich dies in anderen Studien nicht bestätigen ließ. Auch scheinen sich international die Effekte zu unterscheiden (HAPP/FÖRSTER, 2019). Während in Deutschland geschlechtsspezifische Effekte in der ökonomischen Domäne häufiger zu beobachten sind, scheint dies in anderen Staaten bspw. der Schweiz und Japan weniger der Fall zu sein (ACKERMANN, 2019; SCHUMANN ET AL., 2017). Ein Problem stellt hier auch die fehlende Vergleichbarkeit der Ergebnisse dar, da die eingesetzten Testinstrumente sich vom Testdesign und Inhalt sowie der Zielgruppe unterscheiden. Eine empirisch belastbarere Aussage für das vorliegende Testinstrument TBA-EL kann erst mit den repräsentativen Daten der Hauptstudie getroffen werden.

Hinsichtlich des Migrationshintergrundes und des sozioökonomischen Hintergrundes ließen sich sowohl auf Test- als auch auf Itemebene signifikante DIF-Effekte identifizieren. Beide DIF-Effekte auf Testebene sind jedoch unter 0,426 und gemäß der Kategorisierung von PAKK UND WILSON (2011) als akzeptabel zu werten. Auf Itemebene zeigten sich für das Merkmal des Migrationshintergrundes drei signifikante DIF-Unterschiede: So ließ sich ein DIF-Effekt zuungunsten der Testteilnehmer:innen mit Migrationshintergrund bei einer Aufgabe ohne dargebotene Lösungsoption (FT73: Entscheidungsbegründung formulieren) feststellen, die kognitiv komplexe Prozesse anspricht und ein mehrschrittiges Verfahren zur Lösungsgewinnung erfordert. Zugunsten von Testteilernehmer:innen mit Migrationshintergrund konnten zwei Items identifiziert werden, die das korrekte Erkennen und Abrufen von Faktenwissen (in beiden Fällen Definitionen) erfordern. Vergleichbare Schlussfolgerungen zu DIF-Unterschieden, die sich auf den Migrationshintergrund zurückführen lassen, lieferte auch die PISA-Erhebung 2009 (DA DINIS COSTA/ARAÚJO, 2012).

Aus der empirischen Bildungsforschung ist bekannt, dass insbesondere der sozioökonomische Hintergrund der Teilnehmer:innen den Großteil des Unterschieds in der Testleistung zwischen Personen mit und ohne Migrationshintergrund erklärt (BEHR/FUGGER, 2020; BORGNA/CONTINI, 2014). Befunde zu Studien der Migrationsbewegungen in zentraleuropäische Staaten lassen den Schluss zu, dass im Mittel der sozioökonomische Status der Einwandernden geringer ist als der Einheimischen (JERRIM, 2015). Dies gilt insbesondere für die deutschsprachigen Länder und Dänemark (RINDERMANN/THOMPSON, 2016). Zudem unterscheiden sich die individuellen Hintergrundfaktoren der unterschiedlichen Einwanderungsgruppen erheblich (ARIKAN ET AL., 2017), sodass der gefundene geringe DIF-Unterschied im Testinstrument für gewisse Subgruppen deutlich höher oder auch niedriger ausfallen kann. Bezogen auf den sozioökonomischen Hintergrund zeigen sich zehn signifikante DIF-Unterschiede, von denen zwei auch beim Merkmal Migrationshintergrund auftreten (siehe Tab. 8). Analog zum Migrationshintergrund zeigt sich ein DIF-Effekt zuungunsten von Testteilnehmer:innen bei Items, die konzeptionelles Wissen erfordern und das Formulieren von Argumenten zur Lösungsgewinnung benötigen. Items, die Faktenwissen adressieren, fallen Schüler:innen mit einem niedrigeren sozioökonomischen Hintergrund häufiger leichter. Zudem zeigen sich DIF-Effekte bei drei mathematisch bezogenen Items, die größtenteils Probanden mit höherem SÖH bevorteilen.

Bezüglich des Merkmals der familiären Verwendung der Landessprache konnte auf Testebene kein signifikanter DIF-Effekt festgestellt werden, sodass eine systematische Benachteiligung derjenigen, die Deutsch nicht als Erstsprache im familiären Umfeld verwenden, ausgeschlossen werden kann. Auf Itemebene zeigen sich hingegen bei fünf Items signifikante DIF-Effekte, von denen vier ebenfalls bei der Betrachtung des sozioökonomischen Hintergrundes auftreten, sodass Wechselwirkungen nicht ausgeschlossen werden können. Die Analyse der DIF-Effekte der Items zeichnet ein ähnliches Bild wie bei den Merkmalen Migrationshintergrund und des SÖHs.

Hinsichtlich des Aspekts der Testentwicklung können bei der Betrachtung der Item Infit-Werte sowie der DIF-Effekte für sechs Items spezifische Veränderungen vorge-

nommen werden, die in der Hauptstudie zu besseren psychometrischen Eigenschaften führen können. Hinsichtlich der Hauptstudie wurde Item FT75 von der Hauptstudie ausgeschlossen, da keine Teilnehmer:innen dieses gelöst hat und es aus inhaltsvalider Perspektive für das Konstrukt nicht essenziell ist, da bereits andere Items Teilbereiche rund um das Inhaltsfeld „Kaufvertrag“ abdecken. Für die Items (FT23 und FT51) mit Partial-Credit Codierung scheint das Erkennen der teilweise richtigen Antwortoptionen schwerer zu erkennen zu sein als die vollständig richtige. Für beide Items wurden in der Hauptstudie daher die Distraktoren umformuliert und sprachlich präzisiert, ohne jedoch den zugrundeliegenden Iteminhalt zu verändern. Die Items FT52, FT53 und FT62 thematisieren das Rechnen in Prozent in unterschiedlichen Konstellationen und Komplexitätsgraden. Diese mathematikbezogenen Items weisen flache ICCs bei hohen Itemschwierigkeiten auf. Item FT52 wurde in seiner Darbietungsform belassen, da dieses curricular bereits in der Jahrgangsstufe 7 thematisiert sein müsste. Item FT62 wurde bezogen auf die Modellierung und die kognitive Komplexität verringert, während Item FT53 nicht mehr in das Testinstrument der Hauptstudie aufgenommen wurde, da eine Exkludierung die Konstruktbreite nicht beeinträchtigt. Gerade bei der Betrachtung des DIF-Effekts für die familiäre Verwendung der Landessprache wurden sowohl die auffälligen als auch andere textlastige Items auf ihre sprachliche Komplexität geprüft und versucht konstruktirrelevanten kognitiven Load beispielsweise das Verwenden von nicht domänenspezifischen Fachtermini, zu verringern.

5.3 Limitationen

Aufgrund der Befunde und methodischen Restriktionen sind die Ergebnisse mit Vorsicht zu betrachten. Bezüglich der Stichprobenauswahl haben mit 17 % mehr als dreimal so viele Hauptschüler an der Erhebung teilgenommen, als diese in der Gesamtschülerpopulation vorkommen (5 %). Demnach sind andere Schülergruppen unterrepräsentiert. Darüber hinaus ist die Stichprobe im Gegensatz zur Stichprobe der Hauptstudie nicht repräsentativ, sodass keinerlei Aussagen über die Gesamtpopulation der Sekundarstufe getroffen werden können. Das Querschnittsdesign der Studie erlaubt darüber hinaus auch keine kausalen Aussagen zwischen dem Einfluss der personenbezogenen Merkmale auf die Testleistung. Allerdings reihen sich die gefundenen Ergebnisse hinsichtlich des Einflusses des sozioökonomischen Hintergrundes, des Migrationshintergrundes und der Schulform in die vorliegenden Befunde nationaler und internationaler Large-Scale-Studien ein. Der ausbleibende Gendereffekt kann ebenfalls nicht abschließend geklärt werden, da unklar ist, ob die Ursachen am authentischen Assessmentdesign, dem gewählten Verhältnis an unterschiedlichen Itemformaten oder an der Stichprobe selbst liegen. Belastbare Aussagen können hier erst mit den Erkenntnissen aus der Hauptstudie getroffen werden. Darüber hinaus wird mit dieser Studie nur die kognitive Facette wirtschaftlicher Kompetenz adressiert. Gleichwohl wurden innerhalb des Assessments Einstellungsfragen zu ausgewählten wirtschaftlichen Inhalten implementiert und auch im Begleitfragebogen Einstellungen, ökonomische Selbstwirksamkeit als

auch weitere Merkmale erhoben, die jedoch nicht Gegenstand dieses Artikels sind. Das Testinstrument ist nicht frei von signifikanten DIF-Effekten; für die Hauptstudie wurde daher entweder durch Ausschluss oder durch sprachliche und modellierungsbezogene Revision versucht, DIF-Effekte zu reduzieren. Die linksschiefe Verteilung der Personenfähigkeitsparameter könnte auch an der Stichprobenkomposition liegen, sodass bei der Itemrevision darauf geachtet werden muss, nicht vorschnell schwierigere Items zu entfernen oder in ihrer Komplexität zu verringern.

5.4 Implikationen für die Testentwicklung in der ökonomischen Bildung

In der Gesamtschau liegt mit dem Testinstrument TBA-EL ein hinsichtlich der psychometrischen Eigenschaften geeignetes Assessment zur Messung ökonomischer Kompetenz für die Jahrgangsstufe 8 vor. Die Verteilung der Personenfähigkeiten zeigt einen etwas zu schweren Test an. Dies eröffnet die Möglichkeit des Einsatzes in höheren Jahrgangsstufen, da der Test inhaltlich auf die gesamte Sekundarstufe I ausgerichtet worden ist. Hierbei ist allerdings zu berücksichtigen, dass bei der Testentwicklung die hohe Komplexität der ökonomischen Domäne in besonderer Weise zu berücksichtigen ist (OBERRAUCH, 2019). Insbesondere der notwendige Einbezug domänenverbindender Kompetenzen ist herausfordernd: So müssen beispielsweise mathematische Fähigkeiten sowie die Reflexion politischer Rahmenbedingungen und ihre Auswirkungen auf ökonomische Systeme in ihrem inhaltlichen Zuschnitt so modelliert werden, dass Testinstrumente in anderen Jahrgangsstufen ebenfalls einsetzbar sind. Positiv festzustellen ist, dass die DIF-Effekte geringer sind als bei vergleichbaren Testinstrumenten (bspw. WIKO-BW, siehe OBERRAUCH, 2019).

Die Neuerung des Testinstruments stellt neben dem computergestützten Format insbesondere das Einbetten wirtschaftlicher Problemstellungen in authentischen Handlungssituationen mit narrativem Kontext dar, der durch das Design und audiovisuelle Impulse unterstützt wird. Aufgrund des inhaltlichen Bezugs des Testinstruments auf ökonomische Konzepte, die in allen Curricula der ökonomischen Bildung der Sekundarstufe I in Deutschland verankert sind, wäre es von Interesse zu prüfen, ob das Testinstrument in anderen Bundesländern zu ähnlichen Ergebnissen bezogen auf die Testfairness als auch Testleistung der Schüler:innen führt.

Literaturverzeichnis

- ACKERMANN, N. (2019). *Wirtschaftsbürgerliche Kompetenz Deutschschweizer Gymnasiastinnen und Gymnasiasten: Kompetenzmodellierung, Testentwicklung und evidenzbasierte Validierung*. Universität Zürich.
- ACKERMANN, N. (2021). Zum Bildungsideal des „mündigen Wirtschaftsbürgers“: Kompetenzmodell für ökonomische Bildung und Domänenanalyse des gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereichs. In C. Fridrich, U. Hagedorn, R. Hedtke, P. Mitnik & G. Taf-

- ner (Hrsg.), *Wirtschaft, Gesellschaft und Politik* (S. 147–178). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-32910-5_7, Letzter Zugriff: 02.11.2023.
- ACKERMANN, N. / SIEGFRIED, C. (2019). Does a balanced test form regarding selected-response and constructed-response items overcome gender gap in test scores? An analysis of the format-gender relation in the test of economic-civic competence. *Citizenship, Social and Economics Education*, 18(3), 158–176. <https://doi.org/10.1177/2047173419892531>, Letzter Zugriff: 02.11.2023.
- ADAMS, R. J. / KHOO, S. T. (1996). *Quest-Interactive test analysis system*.
- ADAMS, R. J. / WILSON, M. / WANG, W.-C. (1997). The multidimensional random coefficient multinomial logit model. *Applied Psychological Measurement*. *Applied psychological measurement*, 21, 1–23.
- ADAMS, R. J. / WU, M. L. / WILSON, M. (2018). ACER ConQuest. In W. J. van der Linden (Hrsg.), *Handbook of Item Response Theory: Three Volume Set* (S. 495–505). CRC Press.
- AERA / APA / NCME. (2014). *Standards for Educational and Psychological Testing: National Council on Measurement in Education*.
- AFRIANA, I. / FESTIYED (2020). Meta-analysis of authentic assessment instrument development to measure learning outcomes of learners SMA/MA. *Journal of Physics: Conference Series*, 1481(1), 12052. <https://doi.org/10.1088/1742-6596/1481/1/012052>, Letzter Zugriff: 02.11.2023.
- ANDREWS, J. / HIGSON, H. (2014). Is Bologna Working? Employer and Graduate Reflections of the Quality, Value and Relevance of Business and Management Education in four European Union Countries. *Higher Education Quarterly*, 68(3), 267–287. <https://doi.org/10.1111/hequ.12054>, Letzter Zugriff: 02.11.2023.
- ARIKAN, S. / VAN DE VIJVER, F. J. R. / YAGMUR, K. (2017). PISA mathematics and reading performance differences of mainstream European and Turkish immigrant students. *Educational Assessment, Evaluation and Accountability*, 29(3), 229–246. <https://doi.org/10.1007/s11092-017-9260-6>, Letzter Zugriff: 02.11.2023.
- BAETHGE-KINSKY, V. / BAETHGE, M. / LISCHESKI, J. (2016). Bedingungen beruflicher Kompetenzentwicklung: institutionelle und individuelle Kontextfaktoren (SiKoFak). In K. Beck, M. Landenberger & F. Oser (Hrsg.), *Wirtschaft – Beruf – Ethik: v. 32. Technologiebasierte Kompetenzmessung in der beruflichen Bildung: Ergebnisse aus der BMBF-Förderinitiative ASCOT* (1. Auflage, S. 265–293). W Bertelsmann Verlag.
- BECK, K. / KRUMM, V. (1998). *Wirtschaftskundlicher Bildungs-Test (WBT)*. Hogrefe.
- BECK, K. / LANDENBERGER, M. / OSER, F. (Hrsg.). (2016). *Wirtschaft – Beruf – Ethik: v. 32. Technologiebasierte Kompetenzmessung in der beruflichen Bildung: Ergebnisse aus der BMBF-Förderinitiative ASCOT* (1. Auflage). W Bertelsmann Verlag. https://www.content-select.com/index.php?id=bib_view&ean=9783763954605, Letzter Zugriff: 02.11.2023.
- BEHR, A. / FUGGER, G. (2020). PISA Performance of Natives and Immigrants: Selection versus Efficiency. *Open Education Studies*, 2(1), 9–36. <https://doi.org/10.1515/edu-2020-0108>, Letzter Zugriff: 02.11.2023.
- BELLER, M. / GAFNI, N. (2000). Can Item Format (Multiple Choice vs. Open-Ended) Account for Gender Differences in Mathematics Achievement? *Sex Roles*, 42, 1–21. <https://doi.org/10.1023/A:1007051109754>, Letzter Zugriff: 02.11.2023.
- BENNER, P. (2012). Educating nurses: a call for radical transformation-how far have we come? *The Journal of nursing education*, 51(4), 183–184. <https://doi.org/10.3928/01484834-20120402-01>, Letzter Zugriff: 02.11.2023.
- BLÖMEKE, S. / GUSTAFSSON, J.-E. / SHAVELSON, R. J. (2015). Beyond Dichotomies. *Zeitschrift für Psychologie*, 223(1), 3–13. <https://doi.org/10.1027/2151-2604/a000194>, Letzter Zugriff: 02.11.2023.
- BLÖMEKE, S. / ZLATKIN-TROITSCHANSKAIA, O. / KUHN, C. / FEGE, J. (Hrsg.). (2013). *Modeling and Measuring Competencies in Higher Education*. SensePublishers. <https://doi.org/10.1007/978-94-6091-867-4>, Letzter Zugriff: 02.11.2023.

- BLOXHAM, S. (2015). Assessing Assessment: New Developments in Assessment Design, Feedback Practices and Marking in Higher Education. In H. Fry, S. Ketteridge & S. Marshall (Hrsg.), *A Handbook for Teaching and Learning in Higher Education* (4. Aufl., S. 107–122). Routledge.
- BOLGER, N. / KELLAGHAN, T. (1990). Method of Measurement and Gender Differences in Scholastic Achievement. *Journal of Educational Measurement*, 27(2), 165–174. <https://doi.org/10.1111/j.1745-3984.1990.tb00740.x>, Letzter Zugriff: 02.11.2023.
- BORGNA, C. / CONTINI, D. (2014). Migrant Achievement Penalties in Western Europe: Do Educational Systems Matter? *European Sociological Review*, 30(5), 670–683. <https://doi.org/10.1093/esr/jcu067>, Letzter Zugriff: 02.11.2023.
- CARE, E. / KIM, H. (2018). Assessment of Twenty-First Century Skills: The Issue of Authenticity. In E. Care, P. Griffin & M. Wilson (Hrsg.), *Assessment and Teaching of 21st Century Skills* (S. 21–40). Springer International Publishing.
- CHEN, P. / PILZ, M., GREIMEL-FUHRMANN, B. / RUMPOLD, H. (2019). Wirtschaftswissen von Jugendlichen in China und Österreich im Vergleich. Einstellungen, Selbsteinschätzungen und Kenntnisse im bilateralen Kontext. 0947–9732, 25. <https://doi.org/10.25656/01:24834>, Letzter Zugriff: 02.11.2023.
- COLTHORPE, K. / GRAY, H., AINSCOUGH, L. / ERNST, H. (2021). Drivers for authenticity: student approaches and responses to an authentic assessment task. *Assessment & Evaluation in Higher Education*, 46(7), 995–1007. <https://doi.org/10.1080/02602938.2020.1845298>, Letzter Zugriff: 02.11.2023.
- CORSTEN, H. / ROTH, S. (2012). Nachhaltigkeit als integriertes Konzept. In *Nachhaltigkeit* (S. 1–13). Gabler Verlag, Wiesbaden. https://doi.org/10.1007/978-3-8349-3746-9_1, Letzter Zugriff: 02.11.2023.
- DA DINIS COSTA, P. / ARAÚJO, L. (2012). *Differential item functioning (DIF): What functions differently for immigrant students in PISA 2009 reading items?* EUR (Luxembourg. Online): Bd. 25565. Publications Office.
- DEEMER, E. D. / OGAS, J. P. / BARR, A. C. / BOWDON, R. D. / HALL, M. C. / PAULA, S. / CAPOBIANCO, B. M. / LIM, S. (2022). Scientific Research Identity Development Need Not Wait Until College: Examining the Motivational Impact of a Pre-college Authentic Research Experience. *Research in Science Education*, 52(5), 1481–1496. <https://doi.org/10.1007/s11165-021-09994-6>, Letzter Zugriff: 02.11.2023.
- DÖRING, N. / BORTZ, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-642-41089-5>, Letzter Zugriff: 02.11.2023.
- EBERLE, F. (2015). Die Förderung ökonomischer Kompetenzen zwischen normativem Anspruch und empirischer Rationalität – am Beispiel der Schweizer Sekundarstufe II. *Empirische Pädagogik* (EP), 29(1), 10–34–34. <https://www.zora.uzh.ch/id/eprint/118532/>, Letzter Zugriff: 02.11.2023.
- EBERLE, F. / SCHUMANN, S. / KAUFMANN, E. / JÜTTLER, A. / ACKERMANN, N. (2016). Modellierung und Messung wirtschaftsbürgerlicher Kompetenz von kaufmännischen Auszubildenden in der Schweiz und in Deutschland (COBALIT). In K. Beck, M. Landenberger & F. Oser (Hrsg.), *Wirtschaft – Beruf – Ethik: v. 32. Technologiebasierte Kompetenzmessung in der beruflichen Bildung: Ergebnisse aus der BMBF-Förderinitiative ASCOT* (1. Auflage, S. 93–117). W Bertelsmann Verlag.
- EHRICH, J. F. / HOWARD, S. J. / BOKOSMATY, S. / WOODCOCK, S. (2021). An Item Response Modeling Approach to Cognitive Load Measurement. *Frontiers in Education*, 6. <https://doi.org/10.3389/feduc.2021.648324>, Letzter Zugriff: 02.11.2023.
- ELLSTRÖM, E. / ELLSTRÖM, P.-E. (2014). Learning outcomes of a work-based training programme. *European Journal of Training and Development*, 38(3), 180–197. <https://doi.org/10.1108/EJTD-09-2013-0103>, Letzter Zugriff: 02.11.2023.

- ENGARTNER, T. / FRIDRICH, C., GRAUPE, S. / HEDTKE, R. / TAFNER, G. (Hrsg.). (2018). *Sozioökonomische Bildung und Wissenschaft. Sozioökonomische Bildung und Wissenschaft*. Springer Fachmedien Wiesbaden. <https://doi.org/10.1007/978-3-658-21218-6>, Letzter Zugriff: 02.11.2023.
- ERCIKAN, K. / OLIVERI, M. E. (2016). In Search of Validity Evidence in Support of the Interpretation and Use of Assessments of Complex Constructs: Discussion of Research on Assessing 21st Century Skills. *Applied measurement in education*, 29(4), 310–318. <https://doi.org/10.1080/08957347.2016.1209210>, Letzter Zugriff: 02.11.2023.
- FÖRSTER, M. / BRÜCKNER, S. / ZLATKIN-TROITSCHANSKAIA, O. (2015). Assessing the financial knowledge of university students in Germany. *Empirical Research in Vocational Education and Training*, 7(1). <https://doi.org/10.1186/s40461-015-0017-5>
- FÖRSTER, M. / WALSTAD, W. B. / ASARTA, C. (Hrsg.). (2018). *Empirische Pädagogik: 32. Jahrgang, 3./4. Heft (2018). Financial literacy of students and young adults: Determinants and impacts*. Verlag Empirische Pädagogik.
- FORTUNATI, F. / WINTHER, E. (accepted with revisions, 2023a). *Curriculare Analysen als Baustein der Assessmentkonstruktion* [Faculty of Educational Sciences, University of Duisburg-Essen.]. Universität Duisburg-Essen.
- FORTUNATI, F. / WINTHER, E. (accepted with revisions, 2023b, März). Intensionen und Intentionen von Curricula: Domänenmodelle als Kohärenzanker instruktorischer Aktivität am Beispiel der ökonomischen Bildung. *Unterrichtswissenschaft* (Faculty of Educational Sciences, University of Duisburg-Essen.).
- FREY, A. (2012). Adaptives Testen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 275–293). Springer Berlin Heidelberg.
- GELMAN, R. / GREENO, J. G. (1989). On the nature of Competence: Principle for Understanding in a Domain. In L. B. Resnick (Hrsg.), *Knowing, Learning and Instruction. Essays in Honor of Robert Glaser* (S. 125–186). Lawrence Erlbaum Associates.
- GERHOLZ, K.-H. / RÜSCHEN, E. (2011). Und was mach' ich nun? – Der Übergang vom Bachelor zum Master aus einer empirischen und hochschuldidaktischen Perspektive. *bwp@ Spezial*.
- GÖKBUDAK, M. (2021). Politische, gesellschaftliche und ökonomische Bildung in der Sekundarstufe I in Deutschland. Analyse von Lehrplänen und Vorgaben. In C. Fridrich, U. Hagedorn, R. Hedtke, P. Mitnik & G. Tafner (Hrsg.), *Wirtschaft, Gesellschaft und Politik* (S. 337–360). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-32910-5_14, Letzter Zugriff: 02.11.2023.
- GOLDHAMMER, F. / GNIEWOSZ, G. / ZYLKA, J. (2016). ICT Engagement in Learning Environments. In S. Kuger, E. Klieme, N. Jude & D. Kaplan (Hrsg.), *Assessing Contexts of Learning* (S. 331–351). Springer International Publishing.
- GOLDHAMMER, F. / KRÖHNE, U. (2020). Computerbasiertes Assessment. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 119–142). Springer Berlin Heidelberg.
- GREENO, J. G. (1998). The situativity of knowing, learning, and research. *American Psychologist*, 53(1), 5–26. <https://doi.org/10.1037/0003-066X.53.1.5>, Letzter Zugriff: 02.11.2023.
- GREENO, J. G. / COLLINS, A. M. / RESNICK, L. B. (1996). Cognition and learning. In *Handbook of educational psychology* (S. 15–46). Macmillan.
- GRUNDMANN, D. (2017). *Bildung für nachhaltige Entwicklung in Schulen verankern*. Springer Fachmedien Wiesbaden.
- GULIKERS, J. T. M. / BASTIAENS, T. J. / KIRSCHNER, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development*, 52(3), 67–86. <https://doi.org/10.1007/BF02504676>, Letzter Zugriff: 02.11.2023.
- GULIKERS, J. T. M. / BASTIAENS, T. / KIRSCHNER, P. (2006). Authentic assessment, student and teacher perceptions: the practical value of the five-dimensional framework. *Journal of Vocatio-*

- nal Education & Training*, 58(3), 337–357. <https://doi.org/10.1080/13636820600955443>, Letzter Zugriff: 02.11.2023.
- GULIKERS, J. T. M. / BASTIAENS, T. J. / KIRSCHNER, P. A. / KESTER, L. (2006). RELATIONS BETWEEN STUDENT PERCEPTIONS OF ASSESSMENT AUTHENTICITY, STUDY APPROACHES AND LEARNING OUTCOME. *Studies in Educational Evaluation*, 32(4), 381–400. <https://doi.org/10.1016/j.stueduc.2006.10.003>, Letzter Zugriff: 02.11.2023.
- HAERAZI, H. / PRAYATI, Z. / VIKASARI, R. M. (2019). Practicing contextual teaching and learning (CTL): Approach to improve students' reading comprehension in relation to motivation. *English Review: Journal of English Education*, 8(1), 139. <https://doi.org/10.25134/erjee.v8i1.2011>, Letzter Zugriff: 02.11.2023.
- HAPP, R. / FÖRSTER, M. (2019). The relationship between migration background and knowledge and understanding of personal finance of young adults in Germany. *International Review of Economics Education*, 30, 100141. <https://doi.org/10.1016/j.iree.2018.06.003>, Letzter Zugriff: 02.11.2023.
- HAPP, R. / FÖRSTER, M. / SIEGFRIED, C. (2019). Gender-specific differences in economic competence. *Citizenship, Social and Economics Education*, 18(3), 123–127. <https://doi.org/10.1177/2047173419892530>, Letzter Zugriff: 02.11.2023.
- HAPP, R., / KATO, M. / RÜTER, I. (2021). Results from the test of economic literacy in Germany and Japan: A critical discussion on the gender effect. *Citizenship, Social and Economics Education*, 20(1), 48–68. <https://doi.org/10.1177/20471734211004117>, Letzter Zugriff: 02.11.2023.
- HARTIG, J. / KLIEME, E. (2006). Kompetenz und Kompetenzdiagnostik. In K. Schweizer (Hrsg.), *Leistung und Leistungsdiagnostik (German Edition)* (S. 127–143). Springer.
- HERING, R. VON, RIETENBERG, A. / HEINZE, A. / LINDMEIER, A. (2021). Nutzen Auszubildende bei der Bearbeitung berufsfeldbezogener Mathematikaufgaben ihr Wissen aus der Schule? Eine qualitative Untersuchung mit angehenden Industriekaufleuten. *Journal für Mathematik-Didaktik*, 42(2), 459–490. <https://doi.org/10.1007/s13138-021-00181-8>, Letzter Zugriff: 02.11.2023.
- HERING, R. VON, ZINGELMANN, H. / HEINZE, A. / LINDMEIER, A. (2020). Lerngelegenheiten mit kaufmännischem Kontext im Mathematikunterricht der allgemeinbildenden Schule – Eine Schulbuch- und Aufgabenanalyse. *Zeitschrift für Erziehungswissenschaft*, 23(1), 193–213. <https://doi.org/10.1007/s11618-019-00925-w>, Letzter Zugriff: 02.11.2023.
- HÜBNER, R. (2017). Nachhaltigkeits- und Verbraucherforschung. Ein Cross-Over-Versuch interventionsorientierter Wissenschaft. In C. Fridrich, R. Hübner, K. Kollmann, M.-B. Piorkowsky & N. Tröger (Hrsg.), *Abschied vom eindimensionalen Verbraucher* (S. 161–200). Springer Fachmedien Wiesbaden.
- JAMES, L. T. / CASIDY, R. (2018). Authentic assessment in business education: its effects on student satisfaction and promoting behaviour. *Studies in Higher Education*, 43(3), 401–415. <https://doi.org/10.1080/03075079.2016.1165659>, Letzter Zugriff: 02.11.2023.
- JERRIM, J. (2015). Why do East Asian children perform so well in PISA? An investigation of Western-born children of East Asian descent. *Oxford Review of Education*, 41(3), 310–333. <https://doi.org/10.1080/03054985.2015.1028525>, Letzter Zugriff: 02.11.2023.
- KAISER, T. / OBERRAUCH, L. / SEEGER, G. (2020). Measuring economic competence of secondary school students in Germany. *The Journal of Economic Education*, 51(3–4), 227–242. <https://doi.org/10.1080/00220485.2020.1804504>, Letzter Zugriff: 02.11.2023.
- KASTBERG, D. / CUMMINGS, L. / FERRARO, D. / PERKINS, R. C. (2021). *Technical Report and User Guide for the 2018 Program for International Student Assessment (PISA)*. National Center for Education Statistics. <https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2021011>, Letzter Zugriff: 02.11.2023.
- KIRSCH, A. (2021). *Professionalitätsentwicklung angehender Lehrkräfte*. Springer Fachmedien Wiesbaden. <https://doi.org/10.1007/978-3-658-36123-5>, Letzter Zugriff: 02.11.2023.

- KLIEME, E. / HARTIG, J. (2007). Kompetenzkonzepte in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs. In I. Gogolin, H. H. Krüger & M. Prenzel (Hrsg.), *Zeitschrift für Erziehungswissenschaften : ZfE. Sonderheft: 8, 2007. Kompetenzdiagnostik* (1. Aufl., S. 11–29). vs Verlag für Sozialwissenschaften.
- KLOTZ, V. K. (2015). *Diagnostik beruflicher Kompetenzentwicklung: Eine wirtschaftsdidaktische Modellierung für die kaufmännische Domäne*. Springer Fachmedien Wiesbaden; Imprint; Springer Gabler.
- KLOTZ, V. K. / WINTHER, E. / FESTNER, D. (2015). Modeling the Development of Vocational Competence: a Psychometric Model for Economic Domains. *Vocations and Learning*, 8(3), 247–268. <https://doi.org/10.1007/s12186-015-9139-y>, Letzter Zugriff: 02.11.2023.
- KMK. (2006). *Einheitliche Prüfungsanforderungen in der Abiturprüfung Wirtschaft*. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/1989/1989_12_01-EPA-Wirtschaft.pdf, Letzter Zugriff: 27.10.2023
- KRÖHNE, U. (2023). *Open Computer-based Assessment with the CBA ItemBuilder*. <https://cba.item-builder.de>
- KÜHLING-THEES, C. / HAPP, R. / ZLATKIN-TROITSCHANSKAIA, O. / PANT, H. A. (2020). The Impact of Entry Preconditions on Student Dropout and Subject Change in Business and Economics. In O. Zlatkin-Troitschanskaia, H. A. Pant, M. Toepper & C. Lautenbach (Hrsg.), *Student Learning in German Higher Education* (S. 351–370). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-27886-1_18, Letzter Zugriff: 02.11.2023.
- KÜHLING-THEES, C. / ZLATKIN-TROITSCHANSKAIA, O. / NAGEL, M.-T. / YAMADA, R. / SHIRAKAWA, Y. (2021). Generic skills of economics and educational science students – insights from a German-Japanese study. In E. Braun, R. Esterhazy & R. Kordts-Freudinger (Hrsg.), *Research on teaching and learning in higher education* (S. 83–98). Waxmann Verlag.
- LINACRE, J. (1994). Sample size and item calibration stability. *Rasch Mes Trans*, 7, 328. <https://ci.nii.ac.jp/naid/10031091465/>, Letzter Zugriff: 02.11.2023.
- LUSARDI, A. / MITCHELL, O. S. (2014). The Economic Importance of Financial Literacy: Theory and Evidence. *Journal of Economic Literature*, 52(1), 5–44. <https://doi.org/10.1257/jel.52.1.5>, Letzter Zugriff: 02.11.2023.
- LUSARDI, A. / TUFANO, P. (2015). Debt literacy, financial experiences, and overindebtedness. *Journal of Pension Economics and Finance*, 14(4), 332–368. <https://doi.org/10.1017/S1474747215000232>, Letzter Zugriff: 02.11.2023.
- MACHA, K. (2015). *Ökonomische Kompetenz messen theoretisches Modell und Ergebnisse der Economic Competencies Study (ECOS)* [Dissertation]. GBV Gemeinsamer Bibliotheksverbund.
- MACHA, K. / SCHUHEN, M. (2011). *Modellierung ökonomischer Kompetenz in einer Pilotstudie zu ECOS*. Zentrum für ökonomische Bildung in Siegen. <https://dspace.ub.uni-siegen.de/handle/ubsi/599>, Letzter Zugriff: 02.11.2023.
- MACHA, K. / SCHUHEN, M. (2012). Kompetenzmessung in der Domäne Ökonomie. In M. Schuhen, M. Wohlgemuth & C. Müller (Hrsg.), *Ökonomische Bildung und Wirtschaftsordnung* (S. 126–138). Lucius & Lucius.
- MACHA, K. / SCHUHEN, M. (2013). ECOS – Ein unter Gendergesichtspunkten fairer Test allgemeiner ökonomischer Kompetenzen. In T. Retzmann (Hrsg.), *Didaktik der ökonomischen Bildung. Ökonomische Allgemeinbildung in der Sek. II: Konzepte, Analysen und empirische Befunde* (S. 140–152). Wochenschau Verl.
- MARZANO, R. J. / KENDALL, J. S. (2007). *The new taxonomy of educational objectives* (2nd ed.). Corwin Press.
- MCARTHUR, J. (2023). Rethinking authentic assessment: work, well-being, and society. *Higher education*, 85(1), 85–101. <https://doi.org/10.1007/s10734-022-00822-y>, Letzter Zugriff: 02.11.2023.
- MINKLEY, N. / XU, K. M. / KRELL, M. (2021). Analyzing Relationships Between Causal and Assessment Factors of Cognitive Load: Associations Between Objective and Subjective Measures of

- Cognitive Load, Stress, Interest, and Self-Concept. *Frontiers in Education*, 6, Artikel 632907. <https://doi.org/10.3389/feduc.2021.632907>, Letzter Zugriff: 02.11.2023.
- MISLEVY, R. J. (2016). How Developments in Psychology and Technology Challenge Validity Argumentation. *Journal of Educational Measurement*, 53(3), 265–292.
- MISLEVY, R. J. (2018). *Sociocognitive Foundations of Educational Measurement*. Routledge. <https://doi.org/10.4324/9781315871691>, Letzter Zugriff: 02.11.2023.
- MISLEVY, R. J. / RICONSCENTE, M. (2005). *Evidence-centered Assessment Design: Layers, Structures, and Terminology* (PADI Technical Report 9).
- MOOSBRUGGER, H. / KELAVA, A. (Hrsg.). (2012). *Testtheorie und Fragebogenkonstruktion*. Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-642-20072-4>, Letzter Zugriff: 02.11.2023.
- MOOSBRUGGER, H. / KELAVA, A. (2020). *Testtheorie und Fragebogenkonstruktion*. Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-662-61532-4>, Letzter Zugriff: 02.11.2023.
- MSB. (2022). *Das Schulwesen in Nordrhein-Westfalen aus quantitativer Sicht: 2021/22*. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.schulministerium.nrw/system/files/media/document/file/quantita_2019.pdf, Letzter Zugriff: 27.10.2023
- NICKOLAUS, R. / RETELSDORF, J. / WINTHER, E. / KÖLLER, O. (Hrsg.). (2013). *Zeitschrift für Berufs- und Wirtschaftspädagogik: Bd. 26. Mathematisch-naturwissenschaftliche Kompetenzen in der beruflichen Erstausbildung: Stand der Forschung und Desiderata*. Franz Steiner Verlag. <http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10864922>, Letzter Zugriff: 02.11.2023.
- NICOL, D. / THOMSON, A. / BRESLIN, C. (2014). Rethinking feedback practices in higher education: a peer review perspective. *Assessment & Evaluation in Higher Education*, 39(1), 102–122. <https://doi.org/10.1080/02602938.2013.795518>, Letzter Zugriff: 02.11.2023.
- OBERRAUCH, L. (2019). Ökonomische Kompetenzen von Heranwachsenden: Entwicklung und Validierung eines Testinstruments. *Zeitschrift für ökonomische Bildung*, 8, 60–98.
- OBERRAUCH, L. / BRAHM, T. (2022). *Ökonomische Kompetenz und Geschlecht: Ausmaß und Ursachen des Gender Gaps: Working Paper*. zBW – Leibniz Information Centre for Economics. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.econstor.eu/bitstream/10419/250908/1/Oberrauch_Brahm_gender_gap_working_paper.pdf, Letzter Zugriff: 02.11.2023.
- OBERRAUCH, L. / KAISER, T. (2020). Economic competence in early secondary school: Evidence from a large-scale assessment in Germany. *International Review of Economics Education*, 35, 100172. <https://doi.org/10.1016/J.IREE.2019.100172>, Letzter Zugriff: 02.11.2023.
- OECD. (2013). *The Survey of Adult Skills*. OECD. <https://doi.org/10.1787/9789264204027-en>, Letzter Zugriff: 02.11.2023.
- OECD. (2014). *PISA 2012 Results: Students and Money (Volume VI)*. OECD. <https://doi.org/10.1787/9789264208094-en>, Letzter Zugriff: 02.11.2023.
- OECD. (2016a). *Global competency for an inclusive world*. OECD Publishing.
- OECD. (2016b). *The Survey of Adult Skills*. OECD. <https://doi.org/10.1787/9789264258075-en>, Letzter Zugriff: 02.11.2023.
- OECD. (2019a). *The Survey of Adult Skills*. OECD. <https://doi.org/10.1787/f70238c7-en>
- OECD. (2019b). *PISA 2018 Assessment and Analytical Framework*. Organization for Economic Co-operation & Development; Two Rivers Distribution [Distributor]. <https://doi.org/10.1787/b25efab8-en>, Letzter Zugriff: 02.11.2023.
- OECD. (2020). *PISA 2018 Results (Volume IV): Are Students Smart about Money?* PISA OECD Publishing. <https://doi.org/10.1787/48ebd1ba-en>.
- PAEK, I. (2002). *Investigations of differential item functioning: Comparisons among approaches, and extension to a* multidimensional context*. <https://search.proquest.com/openview/a92992ec-1fbc1eao894b9e8e3842fabd/1?pq-origsite=gscholar&cbl=18750&diss=y>, Letzter Zugriff: 02.11.2023.

- PAEK, I. / WILSON, M. (2011). Formulating the Rasch Differential Item Functioning Model Under the Marginal Maximum Likelihood Estimation Context and Its Comparison With Mantel-Haenszel Procedure in Short Test and Small Sample Conditions. *Educational and Psychological Measurement*, 71(6), 1023–1046. <https://doi.org/10.1177/0013164411400734>, Letzter Zugriff: 02.11.2023.
- PELLEGRINO, J. W. (2002). Knowing what Students know. *Issues in Science & Technology*, 19(2), 48–52.
- PELLEGRINO, J. W. (2012). The design of an assessment system focused on student achievement. A learning sciences perspective on issues of competence, growth and measurement. In S. Bernholt, K. Neumann & P. Nentwig (Hrsg.), *Making it tangible. Learning outcomes in science education* (1. Aufl., S. 79–107). Waxmann Verlag GmbH.
- PELLEGRINO, J. W. (2014). Assessment as a positive influence on 21st century teaching and learning: A systems approach to progress. *Psicología Educativa*, 20(2), 65–77. <https://doi.org/10.1016/j.pse.2014.11.002>, Letzter Zugriff: 02.11.2023.
- PELLEGRINO, J. W. / CHUDOWSKY, N. / GLASER, R. (2001). *Knowing What Students Know*. National Academies Press. <https://doi.org/10.17226/10019>
- PELLEGRINO, J. W. / DIBELLO, L. V. / GOLDMAN, S. R. (2016). A Framework for Conceptualizing and Evaluating the Validity of Instructionally Relevant Assessments. *Educational Psychologist*, 51(1), 59–81. <https://doi.org/10.1080/00461520.2016.1145550>, Letzter Zugriff: 02.11.2023.
- PELLEGRINO, J. W. / HILTON, M. L. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. The National Academies Press.
- PLUMP, C. M. (2010). Dealing with problem employees: A legal guide for employers. *Business Horizons*, 53(6), 607–618. <https://doi.org/10.1016/j.bushor.2010.07.003>, Letzter Zugriff: 02.11.2023.
- POHL, S. / GRÄFE, L. / ROSE, N. (2014). Dealing With Omitted and Not-Reached Items in Competence Tests. *Educational and Psychological Measurement*, 74(3), 423–452. <https://doi.org/10.1177/0013164413504926>, Letzter Zugriff: 02.11.2023.
- RAYMOND, J. E. / HOMER, C. S. E. / SMITH, R. / GRAY, J. E. (2013). Learning through authentic assessment: an evaluation of a new development in the undergraduate midwifery curriculum. *Nurse education in practice*, 13(5), 471–476. <https://doi.org/10.1016/j.nepr.2012.10.006>, Letzter Zugriff: 02.11.2023.
- REARDON, S. F. / KALOGRIDES, D. / FAHLE, E. M. / PODOLSKY, A. / ZÁRATE, R. C. (2018). The Relationship Between Test Item Format and Gender Achievement Gaps on Math and ELA Tests in Fourth and Eighth Grades. *Educational Researcher*, 47(5), 284–294. <https://doi.org/10.3102/0013189X18762105>, Letzter Zugriff: 02.11.2023.
- REICHERT-SCHLAX, J. / ZLATKIN-TROITSCHANSKAIA, O. / HAPP, R. / YAMAOKA, M. / ASANO, T. / ABE, S. (2022). Economic literacy among beginning higher education economics students in Japan and Germany and its socio-demographic influencing factors. *Research in Comparative and International Education*, 17(2), 281–300. <https://doi.org/10.1177/17454999211057621>, Letzter Zugriff: 02.11.2023.
- RIENER, G. / WAGNER, V. (2018). Gender differences in willingness to compete and answering multiple-choice questions—The role of age. *Economics Letters*, 164, 86–89. <https://doi.org/10.1016/j.econlet.2018.01.012>, Letzter Zugriff: 02.11.2023.
- RINDERMAN, H. / THOMPSON, J. (2016). THE COGNITIVE COMPETENCES OF IMMIGRANT AND NATIVE STUDENTS ACROSS THE WORLD: AN ANALYSIS OF GAPS, POSSIBLE CAUSES AND IMPACT. *Journal of biosocial science*, 48(1), 66–93. <https://doi.org/10.1017/S0021932014000480>, Letzter Zugriff: 02.11.2023.
- ROST, J. (2004). Psychometrische Modelle zur Überprüfung von Bildungsstandards anhand von Kompetenzmodellen. Vorab-Onlinepublikation. <https://doi.org/10.25656/01:4834>, Letzter Zugriff: 02.11.2023.

- RUMPOLD, H. / GREIMEL-FUHRMANN, B. (2016). Wirtschaftswissen in der Sekundarstufe I: Entwicklung eines Erhebungsinstruments für die Zielgruppe von Schüler/inne/n der achten Schulstufe. *Zeitschrift für ökonomische Bildung*, 5, 119–149.
- SAMBELL, K. / MCDOWELL, L. / MONTGOMERY, C. (2013). *Assessment for learning in higher education*. Routledge.
- SANGMEISTER, J. / WINTHER, E. / DEUTSCHER, V. / BLEY, S. / KREUZER, C. / WEBER, S. (2018). Designing Competence Assessment in VET for a Digital Future. In D. Ifenthaler (Hrsg.), *Digital Workplace Learning* (S. 65–92). Springer International Publishing. https://doi.org/10.1007/978-3-319-46215-8_5, Letzter Zugriff: 02.11.2023.
- SCHANK, C. / LORCH, A. (2018). Der Nachhaltigkeitsbürger in der sozioökonomischen Bildung. Überlegungen zu einem wirtschaftsethisch fundierten sozioökonomischen Bildungsideal sozioökonomischen Bildungsideal. In T. Engartner, C. Fridrich, S. Graupe, R. Hedtke & G. Taffner (Hrsg.), *Sozioökonomische Bildung und Wissenschaft. Sozioökonomische Bildung und Wissenschaft* (S. 215–242). Springer Fachmedien Wiesbaden.
- SCHULTZ, M. / YOUNG, K. K. / GUNNING, T. / HARVEY, M. L. (2022). Defining and measuring authentic assessment: a case study in the context of tertiary science. *Assessment & Evaluation in Higher Education*, 47(1), 77–94. <https://doi.org/10.1080/02602938.2021.1887811>, Letzter Zugriff: 02.11.2023.
- SCHUMANN, S. / EBERLE, F. (2011). Bedeutung und Verwendung schwierigkeitsbestimmender Aufgabenmerkmale für die Erfassung ökonomischer und beruflicher Kompetenzen. In U. Faßhauer, B. Fürstenau & E. Wuttke (Hrsg.), *Grundlagenforschung zum Dualen System und Kompetenzentwicklung in der Lehrerbildung* (S. 77–90). Verlag Barbara Budrich.
- SCHUMANN, S. / EBERLE, F. (2014). Ökonomische Kompetenzen von Lernenden am Ende der Sekundarstufe II. *Zeitschrift für Erziehungswissenschaft*, 17(S1), 103–126. <https://doi.org/10.1007/s11618-013-0459-0>, Letzter Zugriff: 02.11.2023.
- SCHUMANN, S. / KAUFMANN, E. / EBERLE, F. / JÜTTLER, A. / ACKERMANN, N. (2017). Being an economic-civic competent citizen: A technology-based assessment of commercial apprentices in Germany and Switzerland. *Empirical Research in Vocational Education and Training*, 9(1). <https://doi.org/10.1186/s40461-017-0056-1>, Letzter Zugriff: 02.11.2023.
- SEEBER, G. / KAISER, T. / OBERRAUCH, L. / EBERLE, M. (2022). *Wirtschaft als eigenes Schulfach? Empirische Evidenz zur Facheinführung in Baden-Württemberg*. WBV. <https://doi.org/10.3278/9783763972456>, Letzter Zugriff: 02.11.2023.
- SEEBER, G. / KÖRBER, L. / HENTRICH, S. / ROLFES T. / HAUSTEIN, B. (2018). *Ökonomische Kompetenzen Jugendlicher in Baden-Württemberg – WIKO BW: Eine Studie der Universität Koblenz-Landau zum Stand der ökonomischen Bildung in Baden-Württemberg unter der Leitung von Prof. Dr. Günther Seeber. Wirtschaftskompetenzstudie Baden-Württemberg WIKO BW*. Swiridoff Verlag.
- SEEBER, S. / SCHUMANN, S. / NICKOLAUS, R. (2015). Ökonomische Kompetenzen: Konzeptuelle Grundlagen und empirische Befunde. In G. Weißeno & C. Schelle (Hrsg.), *Empirische Forschung in gesellschaftswissenschaftlichen Fachdidaktiken* (S. 169–183). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-06191-3_12, Letzter Zugriff: 02.11.2023.
- SIEGFRIED, C. / ACKERMANN, N. (2020). Die Bedeutung curricularer Lerngelegenheiten im gymnasialen Bereich für die wirtschaftsbürgerliche Kompetenz Eine explorative Studie. *Zeitschrift für ökonomische Bildung*(9), 1–31.
- SIEGFRIED, C. / WUTTKE, E. (2019). Are multiple-choice items unfair? And if so, for whom? *Citizenship, Social and Economics Education*, 18(3), 198–217. <https://doi.org/10.1177/2047173419892525>, Letzter Zugriff: 02.11.2023.
- SOPER, J. C. / WALSTAD, W. B. (1987). *Test of economic literacy* (2. Aufl.). New York: Examiner's Manual. Examiner's Manual.
- SOTIRIADOU, P. / LOGAN, D. / DALY, A. / GUEST, R. (2020). The role of authentic assessment to preserve academic integrity and promote skill development and employability. *Studies in*

- Higher Education*, 45(11), 2132–2148. <https://doi.org/10.1080/03075079.2019.1582015>, Letzter Zugriff: 02.11.2023.
- TAYLOR, C. S. / LEE, Y. (2012). Gender DIF in Reading and Mathematics Tests With Mixed Item Formats. *Applied measurement in education*, 25(3), 246–280. <https://doi.org/10.1080/08957347.2012.687650>, Letzter Zugriff: 02.11.2023.
- TERESI, J. A. / RAMIREZ, M. / LAI, J.-S. / SILVER, S. (2008). Occurrences and sources of Differential Item Functioning (DIF) in patient-reported outcome measures: Description of DIF methods, and review of measures of depression, quality of life and general health. *Psychology science quarterly*, 50(4), 538.
- TRAMM, T. / SEEBER, S. (2006). Überlegungen und Analysen zur Spezifität kaufmännischer Kompetenzen 273–288. In G. Minnameier (Hrsg.), *Berufs- und wirtschaftspädagogische Grundlagenforschung: Lehr-Lern-Prozesse und Kompetenzdiagnostik; Festschrift für Klaus Beck* (S. 273–288). Lang.
- VANAKI, Z. / MEMARIAN, R. (2009). Professional ethics: beyond the clinical competency. *Journal of professional nursing : official journal of the American Association of Colleges of Nursing*, 25(5), 285–291. <https://doi.org/10.1016/j.profnurs.2009.01.009>, Letzter Zugriff: 02.11.2023.
- VILLARROEL, V. / BLOXHAM, S. / BRUNA, D. / BRUNA, C. / HERRERA-SEDA, C. (2018). Authentic assessment: creating a blueprint for course design. *Assessment & Evaluation in Higher Education*, 43(5), 840–854. <https://doi.org/10.1080/02602938.2017.1412396>, Letzter Zugriff: 02.11.2023.
- WALSTAD, W. B. / REBECK, K. (2017). The Test of Financial Literacy : Development and measurement characteristics. *The Journal of Economic Education*, 48(2), 113–122. <https://doi.org/10.1080/00220485.2017.1285739>, Letzter Zugriff: 02.11.2023.
- WALSTAD, W. B. / REBECK, K. / BUTTERS, R. B. (2013). The Test of Economic Literacy: Development and Results. *The Journal of Economic Education*, 44(3), 298–309. <https://doi.org/10.1080/00220485.2013.795462>, Letzter Zugriff: 02.11.2023.
- WALSTAD, W. B. / URBAN, C., J. / ASARTA, C. / BREITBACH, E. / BOSSHARDT, W. / HEATH, J. / O'NEILL, B. / WAGNER, J. / XIAO, J. J. (2017). Perspectives on evaluation in financial education: Landscape, issues, and studies. *The Journal of Economic Education*, 48(2), 93–112. <https://doi.org/10.1080/00220485.2017.1285738>, Letzter Zugriff: 02.11.2023.
- WEBER, B. (2023). *Was Jugendliche über Wirtschaft wissen und können sollen: Eine vergleichende Curriculumanalyse zur ökonomischen Bildung in der Sekundarstufe I*. Wochenschau Wissenschaft: Band 2. Wochenschau Verlag.
- WELSANDT, N. / ABS, H.-J. (2023). Testing Economic Literacy: An Overview of Measurement Instruments of the past 30 Years. *Journal of Social Science Education*, 22(2), 1–35 (Faculty of Educational Sciences, University of Duisburg-Essen.).
- WELSANDT, N. / FORTUNATI, F. / WINTER, E. / ABS, H.-J. (submitted, 2023). Constructing and Validating Authentic Assessments: The Case of a New Technology-Based Assessment of Economic Literacy. *Empirical Research in Vocational Education and Training*.
- WIGGINS, G. (1990). *The Case for Authentic Assessment*. <https://doi.org/10.7275/ffb1-mm19>
- WIGGINS, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. Jossey-Bass.
- WILSON, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, 46(6), 716–730. <https://doi.org/10.1002/tea.20318>
- WILSON, M. / DRANEY, K. (2004). Some Links Between Large-Scale and Classroom Assessments: The Case of the BEAR Assessment System. *Yearbook of the National Society for the Study of Education*, 103(2), 132–154. <https://doi.org/10.1111/j.1744-7984.2004.tb00051.x>, Letzter Zugriff: 02.11.2023.
- WINTER, E. (2010). *Kompetenzmessung in der beruflichen Bildung. Habilitation*. Bertelsmann.
- WINTER, E. / ACHTENHAGEN, F. (2009). Skalen und Stufen kaufmännischer Kompetenz : Skalen und Stufen kaufmännischer Kompetenz. *Zeitschrift für Berufs- und Wirtschaftspädagogik*,

- 105(4), 521–556. <https://elibrary.steiner-verlag.de/article/99.105010/zbw200904052101>, Letzter Zugriff: 02.11.2023.
- WINTHER, E. / ACHTENHAGEN, F. (2013). Kompetenzdiagnostik im kaufmännischen Bereich – ein Umsetzungsbeispiel. In R. Nickolaus, J. Retelsdorf, E. Winther & O. Köller (Hrsg.), *Zeitschrift für Berufs- und Wirtschaftspädagogik: Bd. 26. Mathematisch-naturwissenschaftliche Kompetenzen in der beruflichen Erstausbildung: Stand der Forschung und Desiderata* (S. 203–224). Franz Steiner Verlag.
- WU, X. V. / HENG, M. A. / WANG, W. (2015). Nursing students' experiences with the use of authentic assessment rubric and case approach in the clinical laboratories. *Nurse education today*, 35(4), 549–555. <https://doi.org/10.1016/j.nedt.2014.12.009>, Letzter Zugriff: 02.11.2023.
- WUTTKE, E. / SIEGFRIED, C. / APREA, C. (2020). Measuring financial literacy with a Situational Judgement Test: do some groups really perform worse or is it the measuring instrument? *Empirical Research in Vocational Education and Training*, 12(1). <https://doi.org/10.1186/s40461-020-00103-x>, Letzter Zugriff: 02.11.2023.
- ZLATKIN-TROITSCHANSKAIA, O. / JITOMIRSKI, J. / HAPP, R. / MOLEROV, D. / SCHLAX, J. / KÜHLING-THEES, C. / FÖRSTER, M. / BRÜCKNER, S. (2019). Validating a Test for Measuring Knowledge and Understanding of Economics Among University Students. *Zeitschrift für Pädagogische Psychologie*, 33(2), 119–133. <https://doi.org/10.1024/1010-0652/a000239>, Letzter Zugriff: 02.11.2023.
- ZLATKIN-TROITSCHANSKAIA, O. / PANT, H. A. / TOEPPER, M. / LAUTENBACH, C. (Hrsg.). (2020). *Student Learning in German Higher Education*. Springer Fachmedien Wiesbaden. <https://doi.org/10.1007/978-3-658-27886-1>, Letzter Zugriff: 02.11.2023.
- ZU, T., MUNSELL, J. / REBELLO, N. S. (2021). Subjective Measure of Cognitive Load Depends on Participants' Content Knowledge Level. *Frontiers in Education*, 6, Artikel 647097. <https://doi.org/10.3389/feduc.2021.647097>, Letzter Zugriff: 02.11.2023.

Anhang 1: Überblick über die Iteminhalte und Aufgabenformate des Testinstruments

Item	Nr.	Iteminhalt	Aufgabenformat	Beantwortungsformat
FT11	1	Preisberechnung Einkaufszettel, Grundrechenarten	Auswahlaufgabe	Single-Choice
FT12	2	Bedürfnisse & Bedarf	Auswahlaufgabe	Multiple-Choice
FT13	3	Wirtsch. Unterschied Bio-Produkte/ konv. Produkte	Auswahlaufgabe	Multiple-Choice
FT14	4	Knappheitskonzept	Zuordnung	Drag & Drop
FT21	5	Influencer-Marketing	Zuordnung	Drag & Drop
FT22	6	Wirkung von digitalen Marketingstrategien	Offenes Format	Kurzaufsatz
FT23	7	Nutzen digitaler Marketingstrategien aus Unt.-Sicht	Auswahlaufgabe	Multiple-Choice

Item	Nr.	Iteminhalt	Aufgabenformat	Beantwortungsformat
FT31	8	Definition Nachhaltigkeit	Offenes Format	Kurzaufsatz
FT32	9	Facetten von Nachhaltigkeit	Zuordnung	Drag & Drop
FT33	10	Fair-Trade-Produkte	Offenes Format	Ergänzung
FT34	11	Fair-Trade-Konzept	Auswahlaufgabe	Multiple-Choice
FT35	12	Informationsquellen zu Produktinformationen	Auswahlaufgabe	Multiple-Choice
FT41	13	Berechnung Jahreszinsen	Auswahlaufgabe	Single-Choice
FT42	14	Konzept Zinseszins	Auswahlaufgabe	Single-Choice
FT43	15	Berechnung unterjährige Zinsen	Auswahlaufgabe	Single-Choice
FT44	16	Zölle & Auswirkungen auf Unternehmen	Auswahlaufgabe	Multiple-Choice
FT45	17	Gewinnkonzept	Auswahlaufgabe	Single-Choice
FT46	18	Kaufkraft	Auswahlaufgabe	Multiple-Choice
FT51	19	Zentrale Begriffe Kaufvertrag	Auswahlaufgabe	Single-Choice
FT52	20	Prozentrechnen In-App-Kauf (verminderter GW)	Offenes Format	Wertberechnung
FT53	21	Prozentrechnen In-App-Kauf (Preisdifferenz)	Offenes Format	Wertberechnung
FT54	22	Gefahren des In-App-Kaufs	Offenes Format	Kurzaufsatz
FT61	23	Subtraktion Preisunterschied	Offenes Format	Wertberechnung
FT62	24	Prozentrechnen Angebotsvergleich (vermehrter GW)	Offenes Format	Wertberechnung
FT63	25	Ursachen Kostenvorteile für Online-Shopping	Auswahlaufgabe	Multiple-Choice
FT64	26	Wirkung von AGBs bei Kaufverträgen	Auswahlaufgabe	Single-Choice

Item	Nr.	Iteminhalt	Aufgabenformat	Beantwortungsformat
FT65	27	Währungsumrechnung	Auswahlaufgabe	Single-Choice
FT66	28	Institutionen des Verbraucherschutzes	Auswahlaufgabe	Multiple-Choice
FT71	29	Bezahlen mit Kleingeld	Offenes Format	Kurzaufsatz
FT72	30	Funktionen des Geldes	Auswahlaufgabe	Single-Choice
FT73	31	Kaufvertrag rechtswirksam?	Offenes Format	Kurzaufsatz
FT74	32	Bezahlen mit EC-Karte	Offenes Format	Kurzaufsatz
FT75	33	Kaufvertrag Botengang	Offenes Format	Kurzaufsatz
FT81	34	Preisbildung	Auswahlaufgabe	Single-Choice
FT82	35	Wirtschaftskreislauf	Zuordnung	Drag & Drop

FABIO FORTUNATI, M. SC.

Universität Duisburg-Essen, Lehrstuhl für berufliche Aus- und Weiterbildung, Universitätsstraße 2, 45141 Essen, fabio.fortunati@uni-due.de, ORCID: 0000-0001-8349-7718

NINA JOHANNA WELSANDT, M. A.

Universität Duisburg-Essen, Lehrstuhl Educational Research and Schooling, Universitätsstraße 2, 45141 Essen, nina.welsandt@uni-due.de, ORCID: 0000-0002-8189-1823

PROF. DR. HERMANN JOSEF ABS

Universität Duisburg-Essen, Lehrstuhl Educational Research and Schooling, Universitätsstraße 2, 45141 Essen, h.j.abs@uni-due.de, ORCID: 0000-0002-4230-4200

PROF. DR. ESTHER WINTHER

Universität Duisburg-Essen, Lehrstuhl für berufliche Aus- und Weiterbildung, Universitätsstraße 2, 45141 Essen, esther.winter@uni-due.de, ORCID: 0000-0002-5390-4041