ZEITSCHRIFT FÜR BERUFS- UND WIRTSCHAFTSPÄDAGOGIK, 113, 2017/3, 366–396

MANUEL FÖRSTER / SEBASTIAN BRÜCKNER / ROLAND HAPP / KLAUS BECK / OLGA ZLATKIN-TROITSCHANSKAIA

Strukturanalyse eines kognitiven Messinstruments im Multiple Choice-Format

Das Beispiel des Test of Economic Literacy (TEL4-G)

Structural Analysis of a Cognitive Multiple Choice Measuring Instrument
Exemplified by the Test of Economic Literacy (TEL4-G)

KURZFASSUNG: Trotz beachtlicher Fortschritte in der Entwicklung computerbasierter, simulativ und adaptiv angelegter Messinstrumente werden in Bildungsforschung und Bildungspraxis immer noch vorwiegend herkömmliche Leistungstests eingesetzt. Ihre interne Struktur weist bei genauerem Hinsehen Facetten auf, die nicht in den Blick geraten, wenn, wie das nicht selten der Fall ist, lediglich die in Pilotierungen standardmäßig ermittelten Testkennwerte zur Beurteilung herangezogen werden. Am Beispiel der deutschsprachigen Version des international eingesetzten „Test of Economic Literacy, 4th Ed." wird diskutiert, welche weiteren Merkmale für eine adäquate Interpretation der Test(kenn)werte in Betracht gezogen werden sollten, welche Bedeutung ihnen zukommt und welche Weiterentwicklungen unter dem Validitätsaspekt erforderlich erscheinen. Wie die Daten (nicht nur) unserer Validierungsstudie zeigen, erbringen die in der Testentwicklung international verbreitet angewandten kognitiven Merkmale nicht uneingeschränkt die ihnen theoretisch unterstellte Erklärungsleistung für die Aufgabenschwierigkeit. Weitere, auch nicht-kognitive Bedingungen und Merkmale des Testlösungsverhaltens bedürfen einer genaueren Untersuchung und differenzierteren Kontrolle.
Schlagworte: Itemschwierigkeit; Lösungsstrategien; Sprachstruktur; Testadaption; Validität

ABSTRACT: Despite considerable progress in the development of computer-based, simulative, and adaptive assessment tools, traditional paper-pencil performance tests are predominantly used in education research and practice. On closer examination, however, the internal structure of these tests reveals aspects that remain unnoticed if only standard testing parameters, usually determined in pilot studies, are considered for the assessment. We use the German version of the internationally administered „Test of Economic Literacy, 4th Ed." as an example to discuss further features that should be considered for an adequate interpretation of testing parameters and scores, their importance, and possible enhancements needed to ensure the validity of interpretations. As (not only) the results of our validity study indicate, the cognitive features widely applied in international test development do not have the theoretically alleged power to fully determine task difficulty. From such observations we conclude that additional features,

including non-cognitive conditions and characteristics of test solving strategies, require closer investigation and more elaborated control.
Keywords: item difficulty; solution strategies; linguistic structure; test adaption; validity

1. Einleitung[1]

Obwohl die neueren Entwicklungen im Bereich der Messung von Lehr-Lernergebnissen im Bildungsbereich einen unübersehbaren Trend zu technologiebasierten Verfahren aufweisen (vgl. z. B. das ASCOT- oder das KoKoHs-Programm[2]), werden weiterhin herkömmliche Tests (häufig im MC-Format) zur Erfassung von kognitiven Leistungsfähigkeiten von Lernenden eingesetzt. Ihr Hauptvorteil liegt in ihrer vergleichsweise etwas weniger aufwendigen Entwicklung und der relativ hohen Praktikabilität ihrer Anwendung, ohne dass auf wesentliche Qualitätsmerkmale der Messergebnisse verzichtet werden müsste. Zwar weisen die modernen diagnostischen Verfahren[3] im Hinblick auf den Aspekt der ökologischen Validität insbesondere zur Erfassung von handlungsnahen Fähigkeiten deutliche Stärken und Potentiale auf. In der (Aus-)Bildungspraxis stellen sich jedoch oftmals spezielle Fragen, etwa nach dem erreichten Wissensstand der Lernenden oder danach, ob bei der Testbearbeitung bestimmte Denkprozesse vollzogen werden. Für deren Beantwortung bieten MC-Tests eine brauchbare Alternative zu komplexeren Testverfahren. In der akademischen ebenso wie in der nicht-akademischen Bildungspraxis stellt das MC-Format sogar eines der am häufigsten eingesetzten Messverfahren dar (KUHN, ZLATKIN-TROITSCHANSKAIA, PANT & HANNOVER, 2016).[4] Mit seinen vorformulierten Lösungsmöglichkeiten ist die Testauswertung besonders einfach und objektiv, wobei es das Testformat gleichwohl erlaubt, die Schwierigkeit der Aufgabenstellungen bis hin zu hoch anspruchsvollen Leistungsanforderungen breit zu variieren (z. B. BUCKLES & SIEGFRIED, 2006).

Bei der Auswertung und Interpretation von MC-Tests werden i. d. R. die empirisch ermittelten Testkennwerte („Rohwerte") herangezogen. Einige Studien zur Analyse ihrer internen Struktur (z. B. BRÜCKNER & PELLEGRINO, 2016), die im Kontext eines umfassenden Validierungsansatzes (KANE, 2013), wie er auch von den „Standards for Educational and Psychological Testing" empfohlen wird (AERA, NCME & AEA, 2014), durchgeführt wurden, machen jedoch darauf aufmerksam, dass bei solchen Tests nicht nur konstruktrelevante, sondern auch nicht-konstruktrelevante Aspekte eine Rolle spielen können (vgl. FÖRSTER et al., 2015b). Unter dem Validierungsaspekt stellt sich daher die kritische Frage, welche dieser Aspekte für eine adäquate theoriegeleitete Interpretation der Test(kenn)werte von Bedeutung sind. So macht NICKOLAUS (2016) in seiner

1 Wir danken den anonymen Reviewern für hilfreiche Überarbeitungshinweise.
2 Vgl. zu ASCOT: BECK, LANDENBERGER & OSER (Hrsg.) (2016), zu KoKoHs: z. B. ZLATKIN-TROITSCHANSKAIA, PANT, KUHN, TOEPPER & LAUTENBACH (2016a).
3 Vgl. dazu z. B. FREY & HARTIG (2013).
4 Zur Assessementpraxis s. ZLATKIN-TROITSCHANSKAIA, PANT, KUHN, TOEPPER & LAUTENBACH (2016b).

Überblicksanalyse darauf aufmerksam (vgl. auch ABELE, 2016, S. 42), dass die Wahrscheinlichkeit einer Aufgabenlösung nicht zwingend von „objektiven" und weitgehend inhaltsgleichgültigen Merkmalen (z. B. basierend auf den gängigen und in der Praxis oft angewandten Lehr-Lern-Taxonomien) abhängig ist, obwohl das Argument für deren Bedeutsamkeit eine hohe Plausibilität beanspruchen kann (vgl. NICKOLAUS, 2016, S. 167–168). Wie BRÜCKNER & PELLEGRINO (2016) zeigen, muss man zwischen den objektiven Merkmalen eines zu lösenden Problems und dem psychischen Prozess seiner Lösung unterscheiden, weil dieser letztere bei ein und demselben Problem höchst unterschiedlich und doch erfolgreich verlaufen könne (s. auch ABELE, 2016, S. 41–42).

Anhand einer Analyse der internen Struktur der deutschsprachigen Version des international verbreiteten „Test of Economic Literacy, 4th Ed." (WALSTAD, REBECK & BUTTERS, 2013a)[5] („TEL4-G") wird im Folgenden gezeigt, inwieweit sich die in der Literatur vorfindbaren theoretischen, primär auf kognitive Leistungen fokussierten Merkmale unter dem Validitätsaspekt für eine Erklärung der Testwerte eignen. Weiterhin wird erörtert, wo Weiterentwicklungen in der für zulängliche Ergebnisinterpretationen erforderlichen Charakterisierung von solchen Messinstrumenten notwendig erscheinen. Zunächst wird der TEL4-G in seinen wichtigen Merkmalen vorgestellt (Kap. 2), um danach (Kap. 3) seine Aufgaben mehreren gängigen qualitativen Analyseverfahren zu unterziehen und an Testergebnissen aus unserer Validierungsstudie zu prüfen, welche Erklärungsleistungen von ihnen erwartet werden können. Unter Heranziehung weiterer Merkmale der internen Teststruktur werden Facetten freigelegt, die bei der bloßen Betrachtung von Rohwerten (die Anzahl der richtig gelösten Aufgaben) verborgen bleiben, und die einer adäquaten theoretischen Einordnung bedürfen. Dabei zeigt sich, dass auch nicht-kognitive Bedingungen und Merkmale des Testlösungsverhaltens einer differenzierten Analyse bedürfen (Kap. 4).

2. Das Testinstrument und seine Eigenschaften

2.1 Entwicklung und Aufbau des TEL4-G

Der TEL4 zielt auf die Erfassung des ökonomischen Wissens und des darauf operierenden Denkens bei jungen Erwachsenen (WALSTAD, REBECK & BUTTERS, 2013b, S. 308). Von seinen Entwicklern wird ihm ein breites Anwendungsfeld zugewiesen, das von verschiedenen Formen der didaktischen Nutzung in Unterricht und Lehre bis zum Einsatz in der Lehr-Lern-Forschung und in internationalen Vergleichsstudien reicht (WALSTAD et al., 2013a, S. 8–12, 31–32; YAMAOKA et al., 2010).[6] Mögliche Interpretationen seiner Ergebnisse sind im Manual erörtert, das die Anforderungen der international weit verbreiteten Validitätsstandards weitgehend erfüllt (vgl. AERA, APA & NCME, 2014).

5 Es handelt sich um eine Folgeversion der zweiten Ausgabe, für die eine deutschsprachige Version als Wirtschaftskundlicher Bildungstest (WBT) erschienen ist (BECK, KRUMM & DUBS, 1998).

6 Eine individualdiagnostische Funktion wird dem TEL4 nicht explizit zugeschrieben.

Eine Anpassung der Vorgängerversion (TEL3) war notwendig, nachdem 2010 die „Voluntary National Content Standards in Economics", also seine curricularen Grundlagen, aktualisiert worden waren. Sie umfassen aktuell 20 inhaltlich spezifizierte Standards, die in Kooperation mit renommierten Hochschulprofessoren entwickelt worden und in den USA inzwischen in 22 Bundesstaaten curricular fest etabliert sind (zur curricularen Verankerung in Deutschland s. Kap. 2.3). Der Test wurde für den US-amerikanischen Bereich genormt und standardisiert. Die Testentwickler sind der Auffassung, dass ihr Instrument eine reliable und valide Erfassung des ökonomischen Wissens und Denkens bei Lernenden in der Abschlussphase der High School erlaubt. Ihnen entsprechen im deutschen Bildungssystem jene Gruppen, die sich am Ende der Sekundarstufe oder am Beginn der akademischen oder nicht-akademischen Bildung befinden.

Zum TEL4 liegen zwei Parallelversionen (A & B) mit je 45 Aufgaben im Multiple Choice-Format vor (s. die Beispielitems in Kap. 2.4). Die beiden Parallelversionen sind über 10 (identische) Ankeritems miteinander verbunden. Die beiden Versionen umfassen also insgesamt 80 unterschiedliche Aufgaben. Bei den übrigen 35 Aufgaben der beiden Versionen wurde darauf geachtet, dass sie inhaltlich und hinsichtlich der (kognitiven) Anforderungsstruktur weitgehend parallel aufgebaut sind[7] (WALSTAD et al., 2013b, S. 300). Die Bearbeitungsdauer geben die Testentwickler mit 40 Minuten an.

Analog zu den Vorgängerversionen (zu TEL2 vgl. SOPER & WALSTAD, 1987; WBT: BECK et al., 1998) lassen sich die Items den vier Inhaltsbereichen „Grundlagen" (12 Items), „Mikroökonomie" (11 Items), „Makroökonomie" (15 Items) und „Internationale Beziehungen" (7 Items) zuordnen (s. Tab. 1).

Tab. 1: Aufbau des TEL4

Item-Nr.	Standard	Inhaltsbereich
1, 2, 3	1 Scarcity, choice, productive resources	Grundlagen
4	2 Decision-making, marginal analysis	Grundlagen
5*, 6	3 Economic systems and allocation mechanisms	Grundlagen
7*, 8	4 Economic incentives – prices, wages, profits etc.	Grundlagen
9	5 Voluntary exchange and trade	Grundlagen
10	5 Voluntary exchange and trade	Internationale Beziehungen
11, 12	6 Specialization and comparative advantage	Internationale Beziehungen
13	7 Markets and prices	Mikroökonomie
14	7 Markets and prices	Internationale Beziehungen
15, 16, 17	8 Supply and demand	Mikroökonomie
18*, 19, 20	9 Competition	Mikroökonomie

7 Bei der genauen Analyse konnten allerdings in einigen Fällen geringe, jedoch z. T. kognitiv bedeutsame, Differenzen festgestellt werden, die aus Vergleichbarkeitsgründen in die deutsche Version übernommen, bei den Aufgabenanalysen jedoch berücksichtigt wurden (s. Kap. 3.3).

Item-Nr.	Standard	Inhaltsbereich
21, 22	10 Economic institutions	Grundlagen
23, 24*, 25	11 Money and inflation	Makroökonomie
26, 27	12 Interest rates	Makroökonomie
28, 29*	13 Labor markets and income	Mikroökonomie
30	14 Entrepreneurship	Grundlagen
31, 32	15 Physical and human capital investment	Makroökonomie
33, 34*	16 Economic role of government	Mikroökonomie
35	17 Government failure, special interest groups	Internationale Beziehungen
36*	18 Output, income, employment, and the price level	Makroökonomie
37, 38	18 Output, income, employment, and the price level	Internationale Beziehungen
39, 40	18 Output, income, employment, and the price level	Makroökonomie
41, 42	19 Unemployment and inflation	Makroökonomie
43*, 44, 45	20 Fiscal and monetary policy	Makroökonomie
* Ankeritem		

2.2 Übersetzungs- und Adaptionsprozess

Gegenwärtig ist in der Übersetzungswissenschaft das „best-practice"-Verfahren zur linguistischen Adaptation von Testinstrumenten nicht unumstritten (z. B. ARFFMAN, 2013, S. 2). Während die herkömmliche Vorgehensweise zunächst Wort für Wort und Satz für Satz arbeitet, orientieren sich neuere Verfahren an der Perspektive funktionaler Äquivalenz zwischen den Testversionen (REISS & VERMEER, 2014), bei der die linguistischen und kulturellen Eigenschaften der Zielsprache bereits in die erste Übersetzung integriert werden.

In der Übersetzung des TEL4 wurde diesem Aspekt der funktionalen Äquivalenz durch die Orientierung am translationswissenschaftlichen TRAPD-Modell (Translation, Review, Adjudication, Pretesting and Documentation) Rechnung getragen (BRAY, ADAMSON & MASON, 2007; HARKNESS, 2008): Die deutschsprachige Version des TEL4 wurde in einem ersten Schritt durch zwei unabhängig arbeitende professionelle Fachübersetzer erstellt und in einer von ihnen kommentierten Rohfassung zusammengeführt. Diese wurde mit Experten aus den Wirtschafts- und den Translationswissenschaften unter Bezugnahme auf die „Test Adaption Guidelines" (insbes. D1-D3; HAMBLETON, 2001; ITC, 2005) diskutiert.[8] Einige wenige Zweifelsfälle konnten mit den Entwicklern des TEL4 geklärt werden.

8 Dabei folgten wir insbesondere auch einer der Empfehlungen zur Testadaption von SOLANO-FLORES, BACKHOFF & CONTRERAS-NINO (2009, S. 88).

Für die Testadaptation erleichternd wirkte sich der Umstand aus, dass der (fach) inhaltliche Gegenstand des TEL4, nämlich Grundlagenbereiche der neueren Nationalökonomie, aus einem anglo-amerikanisch dominierten Entwicklungsprozess hervorgegangen ist und heute auf einem international geteilten Grundverständnis beruht, das sich teilweise dadurch einstellte, dass auch außerhalb der USA die dort entwickelten Lehrbücher eingesetzt werden (OECD, 2013). Dies gilt insbesondere für Deutschland. Inhaltliche Anpassungen waren daher kaum erforderlich. Alle Aufgaben des TEL4 konnten sprachlich und kulturell adaptiert werden.[9] Aus mehreren Feedbackrunden mit Experten ging schließlich eine finale Version der beiden Testversionen als TEL4-G, Form A und B, hervor. Sie soll, wie die Vorgängerversionen, auch internationale Vergleiche ermöglichen (BECK, 1991; BECK & KRUMM, 1992).

2.3 Curriculare Validierung

Obwohl seit Jahrzehnten immer wieder die Forderung erhoben worden ist, wirtschaftskundliche Inhalte zum Gegenstand des schulischen Obligatoriums im allgemeinbildenden Bereich zu machen, schreitet dieser Prozess nur langsam voran (s. den Überblick bei KUTSCHA, 1975; BECK & KRUMM, 1992; DEGÖB, 2004). So hat z. B. Baden-Württemberg erst 2016 das Fach „Wirtschaft" an allgemeinbildenden Schulen eingeführt (MINISTERIUM FÜR KULTUS, JUGEND UND SPORT BADEN-WÜRTTEMBERG, 2016). Die deutschen Bundesländer verfahren in dieser Angelegenheit ganz unterschiedlich. So gibt es derzeit in Sachsen-Anhalt und Rheinland-Pfalz im allgemeinbildenden Bereich kein Unterrichtsfach, in dessen Bezeichnung der Begriff „Wirtschaft" erscheint. Nur wenige Bundesländer führen Ökonomie als ein eigenständiges Unterrichtsfach (u. a. Baden-Württemberg, Berlin, Brandenburg). Meist finden sich einige wirtschaftsbezogene Unterrichtsinhalte verstreut in Fächern wie Erdkunde, Sozial- und Gemeinschaftskunde oder Geschichte, teilweise auch in Mathematik (Zinsrechnung) oder Religion/Ethik (Wirtschaftsethik), wo sie jedoch in aller Regel ohne Einbettung in eine Fachsystematik behandelt werden (BRÜCKNER, FÖRSTER, ZLATKIN-TROITSCHANSKAIA & WALSTAD, 2015a).

Anders verhält es sich im berufsbildenden Schulwesen, wo in Abhängigkeit von den institutionellen Gegebenheiten ökonomische Inhalte auf unterschiedlichen Niveaus systematisch in einem oder in verschiedenen Teilfächern (z. B. Betriebs-, Volkswirtschaftslehre, Rechnungswesen) vermittelt werden. Die Spanne reicht hier vom berufsbezogenen Unterricht im Dualen System oder in der vollschulischen Berufsausbildung bis zur propädeutisch konzeptualisierten Wirtschaftslehre an Fachgymnasien.

9 Wie bspw. die Adaption eines mexikanischen Testinstruments zur Erfassung des betriebswirtschaftlichen Wissens und Denkens im Projekt WiWiKom verdeutlicht (ZLATKIN-TROITSCHANSKAIA, FÖRSTER, BRÜCKNER & HAPP, 2014), können in der Testpraxis manche Aufgaben aufgrund kultureller Spezifika in Fachinhalten nicht äquivalent adaptiert werden.

Kenntnisse im breiten Feld der Ökonomie werden in unterschiedlichem Ausmaß auch in den betrieblichen Teilen einer formellen Berufsausbildung erworben, wo nicht nur in den kaufmännischen, sondern – in Abhängigkeit vom inhaltlichen Zuschnitt – auch im gewerblich-technischen und im gesundheitlich-pflegerischen Bereich ökonomiespezifische Inhalte eine bedeutsame Rolle spielen. Wirtschaftsbezogene Fragestellungen durchziehen im Übrigen von früh auf den Alltag aller Menschen ebenso wie die öffentliche Debatte, wo sie zeit- und teilweise sogar den politischen Diskurs dominieren, wie etwa im Kontext der aktuellen europäischen Problemlagen (Stichwort „Brexit") in einer sich immer weiter globalisierenden Welt. So rücken in den variierenden Sozialisationsumwelten, in denen sich Jugendliche und Heranwachsende bewegen, Themen der Ökonomie mit unterschiedlicher Intensität und unterschiedlicher Reichweite ins Blickfeld, die im Nahbereich von Fragen der Bestreitung des Lebensunterhalts bis zur Altersvorsorge, im ferneren gesamtwirtschaftlichen Bereich von Problemen der Staatsverschuldung bis zu internationalen Handelsvereinbarungen reichen.

Obwohl die fächerbezogenen Erfahrungen bei den einzelnen Individuen breit streuen, kann von einem adaptierten deutschsprachigen Messinstrument zur „economic literacy"[10] erwartet werden, dass es die Abbildung jener Inhalte erlaubt, die gemäß der Verankerung in Lehrplänen der Sekundarstufe II, soweit sie dort auftreten, erworben sein sollten. Am Beispiel des Bundeslands Rheinland-Pfalz wurde geprüft, in welchen Bildungsinstitutionen welche ökonomiebezogenen Inhalte mit welchem fachsystematischen Anspruch gelehrt werden und ob sie die Inhaltsstandards enthalten (vgl. CEE, 2010), wie sie im TEL4 repräsentiert sind (WALSTAD et al., 2013a).[11] Dazu wurden die Lehrpläne zu den jeweiligen Schulformen (Ausbildungsberufe, Berufsoberschule I/II, Berufliches Gymnasium und Allgemeinbildendes Gymnasium) analysiert (s. Tabelle 2). Da diese jedoch an einigen Stellen wenig detailliert ausformuliert sind, wurden in die Analyse auch relevante Lehrbücher mit einbezogen.[12]

10 Zur Konzeptdefinition s. BECK (1991), BECK & KRUMM (1992), BECK, KRUMM & DUBS (1998), WALSTAD et al. (2013a), FÖRSTER, ZLATKIN-TROITSCHANSKAIA & HAPP (2015).

11 Als exemplarisch für die große Zahl der Ausbildungsberufe wurden in die Analyse die Curricula für Bankkaufleute und Industriekaufleute einbezogen.

12 Für den Ausbildungsberuf Industriekaufmann/-frau: HARTMANN (2015); für den Ausbildungsberuf Bankkaufmann/-frau: MÖHLMEIER, SKORZENSKI, WIERICHS & WURM, (2015). Parallel hierzu wurden Lehrbücher für die berufsbildenden (Berufsoberschule I/II; Berufliches Gymnasium) und allgemeinbildenden Schulformen analysiert: LORZ & SIEBERT (2007); KRUGMAN, OBSTFELD & MELITZ (2011); PINDYCK & RUBINFELD (2014); PINDYCK & RUBINFELD (2009); MANKIW & TAYLOR (2012).

Tab. 2: Ökonomie-Standards in rheinland-pfälzischen Lehrplänen und Lehrbüchern

Standard des CEE (2010)	**Bankkaufmann/-frau**	**Industriekaufmann/-frau**	**Berufsoberschule I/II**	**Berufliches Gymnasium**	**Allgemeinbildendes Gymnasium**
1 Scarcity, choice, productive resources	+	+	+	+++	+
2 Decision-making, marginal analysis	+	++	++	+++	+
3 Economic systems and allocation mechanisms	++	+	++	+++	++
4 Economic incentives – prices, wages, profits etc.	+	++	++	++	+
5 Voluntary exchange and trade	+	o	o	+	o
6 Specialization and comparative advantage	+	+	+	+++	+++
7 Markets and prices	+++	+++	+	+++	+
8 Supply and demand	+++	+++	o	+++	o
9 Competition	++	+++	o	+++	o
10 Economic institutions	++	++	+	+	++
11 Money and inflation	+++	+++	+++	+++	+
12 Interest rates	+++	+	++	++	+
13 Labor markets and income	++	++	+	+++	++
14 Entrepreneurship	++	+++	o	o	o
15 Physical and human capital investment	++	+++	++	+++	++
16 Economic role of government	+	++	+	++	o
17 Government failure, special interest groups	+	++	+	++	+
18 Output, income, employment, and the price level	+++	+++	+++	+++	++
19 Unemployment and inflation	+++	++	+	++	+
20 Fiscal and monetary policy	+++	+++	+++	+++	+++
Gesamt	**40**	**42**	**27**	**48**	**24**

* o: nicht gelehrt; +: ansatzweise gelehrt; ++: großenteils gelehrt; +++: vollumfänglich gelehrt
Zu Zeile „Gesamt“: Vergibt man für die Pluszeichen Punkte, so erhält man eine grobe quantitative Orientierung zum curricularen Gewicht der Standards in den verschiedenen Bildungsgängen.

In der Gesamtschau der Ergebnisse ist die stärkste Verortung der Inhaltsstandards im Curriculum des beruflichen Gymnasiums zu verzeichnen. Mit einigen Abstrichen – insbesondere in der Tiefe der Vermittlung – finden die Standards auch in den beiden untersuchten Ausbildungsberufen Berücksichtigung. Den geringsten fachsystematischen

Bezug weisen erwartungsgemäß die Curricula für das allgemeinbildende Gymnasium auf.[13]

2.4 Itembeispiele

Die folgenden Itembeispiele[14] (Abb. 1) aus Form A stehen für die vier Inhaltsbereiche: „Grundlagen" (Nr. A7), „Mikroökonomie" (Nr. A19), „Makroökonomie" (Nr. A43) und „Internationale Beziehungen" (Nr. A14). An diesen Aufgaben wird weiter unten die Analyse zu den kognitiven Anforderungen des TEL4-G veranschaulicht.

A7. Profits are equal to total	**A19. A newspaper reports, „COFFEE GROWERS' MONOPOLY BROKEN INTO SEVERAL COMPETING FIRMS." If this is true, we would expect the coffee-growing industry to**
□ A) revenue minus total cost. □ B) assets minus total liabilities. □ C) sales minus wages and salaries. □ D) sales minus taxes and depreciation.	□ A) decrease output and decrease prices. □ B) increase output and increase prices. □ C) decrease output and increase prices. □ D) increase output and decrease prices.
A14. The exchange rate between the U. S. dollar and the euro changes from $1 = 1.50 euros to $1 = 1.25 euros. Germany uses the euro as its currency. This change means that	**A43. One reason the federal government might reduce taxes is to:**
□ A) U. S. goods will be more expensive for Germans. □ B) German goods will be more expensive for Americans. □ C) there will be an increase in U. S. imports from Germany. □ D) there will be a decrease in German imports from the U. S.	□ A) slow the rate of inflation. □ B) slow a rapid rise in interest rates. □ C) decrease business spending on plant and equipment. □ D) increase consumer spending and stimulate the economy.

Abb. 1: Beispielitems aus dem TEL4, Form A

13 Eine detailliertere Darstellung zu diesen Curriculumanalysen findet sich bei HONTHEIM (2016).

14 Aus Gründen der Testsicherheit und der noch laufenden Validierung der deutschen Version erfolgt hier die Darstellung der strukturgleichen englischen Originalaufgaben.

3. Kognitive Anforderungen an die Testbearbeitung

3.1 Ansätze zur Aufgabenanalyse

Unter dem Aspekt der Validitätsbeurteilung eines Tests sowie der auf den Testergebnissen aufbauenden Interpretationen sind neben den inhaltlichen bzw. curricularen Dimensionen die damit einhergehenden kognitiven Anforderungen der Aufgabenstellungen zentral. Ein konstitutives Element aller Curricula ist – neben der sich i. d. R. schrittweise ausweitenden inhaltlichen Breite des Lehrstoffs – das Konzept der zunehmenden Komplexität bzw. des steigenden Anspruchsniveaus.[15] Zwar konkretisiert es sich in jeder Domäne[16] auf spezifische Weise, entfaltet sich jedoch zugleich entlang einer Struktur bzw. Systematik, die von den Dimensionen inhaltsunabhängiger kognitiver Operationen bzw. Leistungen aufgespannt wird (wie Wissensverfügbarkeit, intellektuelle Operationen). Pädagogische Psychologie, Lehr-Lern-Forschung und Didaktik haben eine ganze Reihe solcher Dimensionen aus dem Lehr-Lern-Geschehen herauspräpariert, die i. d. R. hierarchisch aufgebaut sind und so die Zuordnung von erforderlichen Aufgabenlösungsleistungen zu Schwierigkeitsstufen oder -graden erlauben.

Im Rahmen der Validierungsstudie wurden die Aufgaben des TEL4-G den verschiedenen nachfolgend dargestellten Kategorisierungen bzw. Dimensionen zugeordnet, die es ermöglichen (sollen), sie nach damit korrespondierenden schwierigkeitsbestimmenden Merkmalen zu ordnen und damit auch Annahmen über die zu erwartenden Lösungsleistungen zu treffen. So kann geprüft werden, welche dieser Kategorisierungen die empirisch gewonnenen Daten am besten vorhersagen, sowie welchen Beitrag sie einzeln und gemeinsam zur Aufklärung der Testlösungsvarianz leisten. Im Kontext der Studie übernehmen wir dazu zunächst das theoretische Rationale, das die Autoren des TEL4 herangezogen haben, nämlich die BLOOM'sche Taxonomie (s. u. (1)) und ergänzen es dann durch die am weitesten verbreiteten, mit ihr verwandten Ansätze ((2), (3) und (4)) sowie schließlich durch die üblichen formalen Verfahren der Aufgabenanalyse ((5), (6) und (7)), um abweichende Annahmen über die Aufgabenschwierigkeiten prüfen zu können.

(1) Die Autoren des TEL4 kategorisieren die (englischsprachigen) Testaufgaben in loser Anlehnung an die BLOOM'sche Taxonomie. Sie ordnen sie drei hierarchischen Stufen zu: „I Knowledge", „II Comprehension" und „III Application" (WALSTAD et al., 2013a, S. 7). „Knowledge" korrespondiert mit dem Erinnern und Wiederabrufen ökonomisch bedeutsamer Konzepte (grundlegende Begriffe, Definitionen, Kategorien) und

15 Wir verzichten auf eine eingehende Differenzierung zwischen „Komplexität" und „Anspruchsniveau" und die Erörterung der Frage, ob sie ggf. einem zweidimensionalen Konzept zugeordnet werden können, oder ob „Komplexität" die Perspektive des höchstmöglich entfalteten menschlichen Intellekts abbildet und insofern auch als Maß für das Anspruchsniveau eine eindimensionale Sicht begründen könnte (vgl. MINNAMEIER, 2000, S. 168–175).

16 Zur Diskussion des Begriffs „Domäne" s. SEIFRIED & ZIEGLER (2009), WITT (2009), BRÜCKNER (2017). Die Definition und Abgrenzung des Domänenbegriffs stellt einen ersten fundamentalen Schritt bei der Entwicklung jedes Testverfahrens dar (MISLEVY & HAERTEL, 2006).

Zusammenhangsaussagen (Korrelationsbeziehungen, Gesetzmäßigkeiten, Modelle), und zwar so, wie sie zuvor gelehrt wurden. Auf Stufe II „Comprehension" kommt das Paraphrasieren des Wissens und die Veranschaulichung mit Beispielen hinzu. „Application" (Stufe III) kennzeichnet die Fähigkeit, das Wissen samt seinen inneren Bezügen auf subjektiv neue, reale Sachverhalte zu projizieren bzw. diese dem bestehenden Wissenskorpus zuzuordnen, sie u. U. auch zu analysieren und zu evaluieren. Diese Zuordnung der Testaufgaben erfolgte unter der hauptsächlich pragmatischen Perspektive, den Lehrenden für den Umgang mit den Testresultaten ein eher unkompliziertes Rationale an die Hand zu geben (DAVIS, 2001).

(2) In analoger Weise, jedoch in konsequenterer Orientierung an BLOOMS Taxonomie (BLOOM, 1956), die ja sechs Stufen mit zusätzlichen Unterstufen aufweist (Wissen, Verstehen, Anwenden, Analyse, Synthese, Evaluation), wurden die deutschsprachigen Aufgaben des TEL4-G kategorisiert. Dabei übersteigt auch diese Zuordnung wie diejenige der TEL4-Autoren nicht die dritte Stufe der BLOOM-Taxonomie[17]; der Einschluss der Unterstufen erlaubt jedoch eine differenziertere Anordnung (zu abweichenden Hauptstufenzuordnungen s. weiter unten).

(3) ANDERSON und KRATHWOHL haben 2001 eine revidierte Fassung der BLOOM-Taxonomie vorgelegt, die sich von jener wesentlich dadurch unterscheidet, dass sie zweidimensional angelegt ist, indem sie zwischen Wissenstypen (Fakten-, konzeptuelles, prozedurales, metakognitives Wissen) und kognitiven Prozessen (Erinnern, Verstehen, Anwenden, Analysieren, Evaluieren, Kreieren), ebenfalls jeweils mit Unterkategorien, unterscheidet. Diese Strukturierung erlaubt eine differenziertere Aufgabenzuordnung, die zwar keine durchweg eindeutige Hierarchisierung nach Schwierigkeitsstufen ermöglicht, jedoch trotzdem eine unter einem schwierigkeitsbestimmenden Aspekt auswertbare Zuordnung erzeugt.[18]

Problematisch an den Zuordnungen (1), (2) und (3) ist, dass sie ursprünglich als Taxonomien für die Kategorisierung von Lehrzielen entwickelt worden sind und nicht zur Analyse der internen Struktur von pädagogisch-psychologischen Tests (s. auch Ebel & Frisbie, 1991, zit nach Walstad et al., 2013a, S. 7, col. 2). Lehrziele sind jedoch u. a. dadurch charakterisiert, dass sie stets auch kognitive Anforderungen beschreiben (im Prototyp: „Die Adressaten sollen … können") und insofern eins-zu-eins den Stufen einer kognitiven Taxonomie zuordenbar sind (ein Verfahren, das häufig über die vorgängige Zuordnung von denkprozessbeschreibenden Verben zu Taxonomiestufen auf eine nur scheinbar eindeutige Weise durchführbar ist). Aufgaben in Leistungstests stellen Anforderungen, die dort nicht explizit benannt werden und die allererst in der rekonstruktiven Analyse eines „idealen" Lösungsprozesses identifiziert werden können (s. Brückner, 2017; Brückner & Pellegrino, 2016). Damit kommen jedoch kaum

17 Laut den TEL4-Autoren wären einige Aufgaben nach BLOOM eigentlich den Stufen 4 (analysis) oder 6 (evaluation) zuzuordnen. Sie sind allerdings der Ansicht, dass die Anwendungsfragstellung in diesen Aufgaben überwiege und dass durch den Einbezug dieser beiden Kategorien in die Stufe 3 (application) das wichtigere Ziel der einfachen Handhabbarkeit für Lehrende erreicht werde (WALSTAD et al., 2013a, S. 8, col. 1).

18 Es handelt sich dabei um eine sog. Halbordnung. In einer zweidimensionalen Tabellendarstellung sind jeweils die Zellen a2 und b1, b3 und c2, c4 und d3 usw. nicht auf eine Schwieriger-leichter-Relation abzubilden, also als gleich oder als kognitiv äquivalent anzusehen.

kontrollierbare Unwägbarkeiten und auch Perspektivenpräferenzen ins Spiel, die zwar kein intersubjektiv eindeutiges Zuordnungsergebnis zu Taxonomiestufen erlauben, jedoch die Formulierung tentativer Hypothesen über die zu erwartenden Lösungsleistungen der Testpersonen ermöglichen, die dann z. B. im Rahmen eines sozio-kognitiven Ansatzes geprüft werden können (s. Mislevy, 2016; Brückner & Pellegrino, 2016).

(4) Eine bislang im deutschsprachigen Raum wenig beachtete Stufenordnung beruht auf der ganzheitlichen Modellierung des Verstehens von Sachverhalten mittels sog. Schwellenkonzepte (s. BRÜCKNER, 2017). In Anlehnung an MEYER und LAND (2006) repräsentieren sie jene Auswahl aus allen Konzepten einer Domäne, deren Verstehen das Denken in diesem Feld substantiell und nachhaltig auf eine höhere Stufe hebt und zugleich den Zugang zu weiteren gleichartigen Konzepten erschließt. Der Verständniszuwachs wird als das Übersteigen einer konzeptuellen Schwelle bezeichnet. In Anwendung auf ökonomische Sachverhalte werden drei Schwellen konzipiert („basic, discipline und modelling") (DAVIES & MANGAN, 2007, S. 711). Die erste („basic") bezeichnet den Übergang von einem erfahrungsgesättigten Alltagsverständnis zu einem ersten domänenspezifischen Konzeptverständnis (hier z. B. von Knappheit). Die zweite Schwelle („discipline") führt zur Entwicklung eines elaborierten, theoriegestützten Verständnisses der zentralen Zusammenhänge in der Domäne (z. B. das Verständnis des komparativen Kostenvorteils) und die dritte („modelling") zu einer umfassenden Expertise, die es erlaubt, domänenspezifische Denkweisen und Modelle der Disziplin zu nutzen (z. B. das Verständnis von komparativer Statik und intertemporalen Entscheidungen).

(5) In einem nächsten Analyseschritt werden die Anzahl der Fachtermini ausgezählt, die im Aufgabenstamm, in der richtigen Lösung und in den Distraktoren auftreten. Sie können ebenfalls schwierigkeitsbeeinflussende Wirkung erzeugen (s. hierzu DRAXLER, 2005). Zur Identifikation dieser Termini wurden Studienanfänger befragt, um deren Perspektive auf die Aufgabentexte einzunehmen und das von ihnen mitgebrachte Begriffswissen zugrunde zu legen.

(6) Für eine weitere Kategorisierung von Testaufgaben wurde – im Sinne einer Vorwärtsstrategie (s. dazu unten 4.1) – zu jedem Item die Zahl, Art und Abfolge der einzelnen Lösungsschritte ermittelt, die jeweils für die Bestimmung der richtigen Antwort erforderlich sind (s. hierzu z. B. NICKOLAUS, 2016). Dabei werden acht sprachanalytisch trennbare und in ihrer Komplexität ansteigende Aussagevarianten unterschieden, die für eine spätere Beurteilung der Schwierigkeit der Aufgaben nach ihrem kognitiven Anspruch eingeschätzt und dementsprechend gewichtet werden (vgl. Tab. 3). Jede zur Problemlösung erforderliche Aussage wird dabei als ein Lösungsschritt betrachtet.

Tab. 3: Aussagevarianten als Lösungsschritte

Aussageart	**Gewicht**[19]	**Beispiel**
Deskription	1	Der Wechselkurs der Währung hat sich wie angegeben verändert.
Norm		Die Umweltbelastung soll vermindert werden!
Definition	2	Gewinn ist definiert als Erlös minus Kosten.
Operationalisierung	3	Marktwirtschaft bedeutet auch, dass Käufer und Verkäufer über die Allokation von Ressourcen bestimmen.
Subsumption		Der Anstieg von Zinsen im Zeitablauf ist ein Merkmal von Inflation.
Äquivalenz		Eine Verringerung des Dollarpreises für Euros ist gleichbedeutend mit einer Erhöhung des Europreises für Dollars.
Gesetz(mäßigkeit)	4	Wenn die Nachfrage bei gleichbleibendem Angebot steigt, dann bewegt sich (ceteris paribus) der Preis auf ein höheres Gleichgewichtsniveau.
Schlussfolgerung	5	(a) *Gesetz*: Wenn die Nachfrage steigt, dann steigt (ceteris paribus) der Preis. (b) *Deskription*: Die Löhne als Preis für Arbeit sind gestiegen. (c) *Schluss*: Die Nachfrage nach dem hergestellten Produkt war gestiegen.

(7) In einem relativ häufig angewandten Zugriff werden die Aufgaben schließlich danach differenziert, ob sie numerische Daten oder nur Text enthalten. Dem liegt die aus bereits vorliegenden Studien hervorgegangene Annahme zugrunde, dass der geforderte Umgang mit numerischen Daten eine schwierigkeitssteigernde Wirkung erzeugt (vgl. BRÜCKNER et al., 2015b; DAMMANN, BEHRENDT, ŞTEFÁNICĂ & NICKOLAUS, 2016; FÖRSTER, BRÜCKNER & ZLATKIN-TROITSCHANSKAIA, 2015a; PETSCH, NORWIG & NICKOLAUS, 2015; SEEBER, 2008). Diese Unterscheidung wird in der nachfolgenden Analyse allerdings lediglich dichotom vorgenommen (numerische Daten enthalten vs. nicht enthalten). Für eine genauere Schwierigkeitsbestimmung wäre ein mathemati-

19 Die Gewichtung beruht nicht auf einer psychologisch begründeten Analyse, sondern auf sprachanalytischen und sprachlogischen Kriterien (vgl. z. B. OPP, 1972, S. 138–143; PRIM & TILMANN, 1997). Sie indikatorisieren von Stufe zu Stufe eine steigende Komplexität der für eine Lösung erforderlichen, vom Probanden zu generierenden (Abfolge von) Aussage(n). So sind Deskription und Norm als lösungsrelevante einfache Aussagen aus den vorgegebenen Aufgabentexten zu extrahieren; Definitionen bringen begriffliche Relationen zum Ausdruck und bleiben auf ein und derselben Sprachebene, während Operationalisierung und Subsumption sprachebenenübersteigend und Äquivalenz hier sprachebenenvergleichend angelegt sind (Theoriesprache und Beobachtungssprache i. S. v. Carnap; s. STEGMÜLLER, 1984, S. 93–96); Gesetze enthalten mindestens zwei Deskriptionen (je eine in der Wenn- und der Dann-Komponente), die in eine kausale Relation zueinander gebracht werden (ebd. S. 76–78) und Schlussfolgerungen kombinieren mehrere Aussagetypen in einer logisch regelgeleiteten (deduktiven, induktiven oder abduktiven) Weise (ebd. S. 86–90; MINNAMEIER, 2005, S. 95–114).

sches Modell heranzuziehen, das angesichts der im TEL4 gestellten Aufgaben mit numerischen Daten eine weiterführende Differenzierung ermöglicht.

Insgesamt divergieren die Einstufungen bzw. Einordnungen in die unterschiedlichen Kategorien deutlich. Daher wird im nächsten Analyseschritt anhand der empirisch ermittelten Itemschwierigkeiten geprüft, ob und welche Beziehungen zwischen den einzelnen Dimensionen (Spalten in Tab. 4) oder aller Dimensionen zusammengenommen auf der einen Seite und den empirischen Befunden zu den Lösungsleistungen der Testpersonen auf der anderen Seite ermittelt werden können (Kap. 3.3).

In Abbildung 2 sind die wichtigsten, teilweise aggregierten Aspekte dargestellt, die zu Beginn dieses Abschnitts als Merkmale zur Aufgabenbeschreibung herangezogen worden sind. Über jeder Itemnummer sind die Ausprägungen des jeweiligen Merkmals abgetragen. Diese Darstellung der Einzelbefunde vermittelt (lediglich) einen Gesamteindruck von der Heterogenität der Zuordnungen, die in Abhängigkeit vom jeweiligen Kriterium vorzunehmen sind. Als gepunktete Linie ist jeweils der Profilzug der „klassisch" ermittelten Itemschwierigkeiten (Lösungshäufigkeiten) eingetragen (Werte von 0.0 bis 1.0; rechte Skala; vgl. dazu unten Abschnitt 3.3), dessen Verlauf mit den Veränderungen in den Zuordnungen zu den Analysekategorien offenkundig nur wenige Gemeinsamkeiten aufweist.

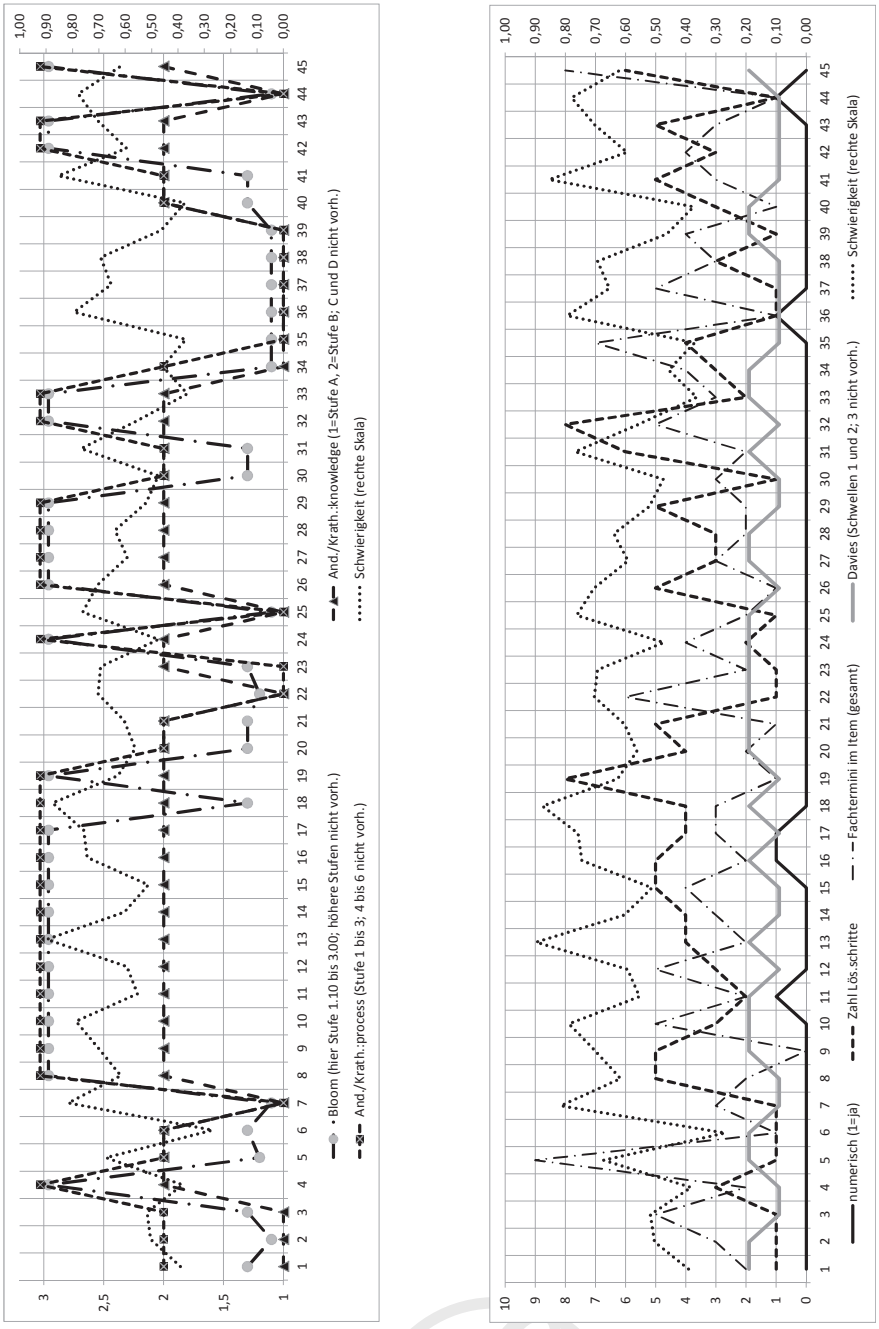

Abb. 2: Itemcharakteristika des TEL4-G im Vergleich (Form A)

Wendet man diese Zuordnungen auf die vier Beispielitems an, so ergibt sich für sie die in Tabelle 4 dargestellte Charakterisierung.

Tab. 4: Aufgabenanalyse nach schwierigkeitsbestimmenden Merkmalen

Nr.	Bereich	Standard	numerisch	kognitives Niveau				Zahl der Fachtermini	
				TEL4	*Bloom*[20]	*Anderson & Krathwohl*	*Davies & Mangan*	*Itemstamm mit Attraktor*	*Distraktor*
A7	Grundlagen	ökonomische Anreize	Nein	I	1.11	Aa/1.2	2	0	3
A14	Internat. Beziehungen	Märkte und Preise	Ja	III	3.00	Bb/3.2	1	1	2
A19	Mikro-ökonomie	Wettbewerb	nein	III	3.00	Bb/3.2	2	1	0
A43	Makro-ökonomie	Fiskal- und Geldpolitik	nein	III	3.00	Bb/3.1	1	1	2

Nr.	Anzahl lösungsrelevanter Aussagen					Höchste kognitive Operation	Anzahl Denkschritte	Gewicht
	Deskriptionen	*Definitionen*	*Op/Sub/Äquiv.*	*Gesetze*	*Schlüsse*			
A7	0	1	0	0	0	Definition	1	2
A14	1	0	1 (Operat.) 2 (Äquivalenz)	0	0	Äquivalenz	4	7
A19	1	1	0	2	4	Schluss	8	31
A43	0	1	1 (Äquivalenz)	2	0	Gesetz	4	16

3.2 Zusammenhang zwischen Aufgabenanalyse und Testlösungsleistung

Erhebungsdesign und Stichprobe

Zu Beginn des SS 2014 fand eine Paper-Pencil-Befragung bei Studieneinsteigern[21] in ein wirtschaftswissenschaftliches Hochschulstudium statt. Sie wurde in einführende Informationsveranstaltungen integriert, an denen die Studieneinsteiger der jeweiligen Hochschule noch vor Studienbeginn teilnahmen. So konnte sichergestellt werden, dass Personen erfasst wurden, die noch gar keine einschlägigen hochschulischen Lehrveranstaltungen besucht haben. Die Erhebung wurde an vier Fachhochschulen und sechs Universitäten durchgeführt.[22] Die Befragungsdauer lag bei 45 Minuten. Die Stichprobe

20 Hier werden die Stufen „1.10 Wissen von konkreten Einzelheiten" mit der Unterstufe 1.11 „Kenntnis von Begriffen", „1.20 Wissen von Klassifikationen und Kategorien", „1.30 Wissen von Verallgemeinerungen und Abstraktionen" und „3.00 Anwendung" unterschieden. Die übrigen Stufen werden von den TEL4-G-Items nicht besetzt.

21 Das in diesem Beitrag verwendete Genus ist geschlechtsunspezifisch. Die ausschließliche Verwendung der maskulinen Form soll die Lesbarkeit erhöhen, schließt jedoch stets die feminine Form mit ein.

22 Die teilnehmenden Hochschulen und Studierenden, denen wir zu Dank verpflichtet sind, stammen aus den Bundesländern Baden-Württemberg, Hessen, Niedersachsen, Nordrhein-Westfalen und Rheinland-Pfalz.

umfasst nach der Bereinigung 1.395 Datensätze, 700 in der Version A, 695 in der Version B (die 10 Ankeritems wurden von allen 1.395 Teilnehmenden beantwortet). Von den Befragten sind knapp die Hälfte weiblich (48,2 %) und 270 (19,4 %) verfügen über eine kaufmännische Berufsausbildung, 294 haben ihre Hochschulzugangsberechtigung (HZB) an einem Wirtschaftsgymnasium erworben. Von diesen 270 bzw. 294 Studieneinsteigern haben 112 (8 %) sowohl das Wirtschaftsgymnasium als auch eine kaufmännische Berufsausbildung abgeschlossen. Der Durchschnitt der Note der HZB liegt bei 2,39 (Standardabweichung 0,55).

Test-Reliabilität und Item-Trennschärfe

Beim TEL4-G handelt es sich um einen Test, der ökonomische Grundprinzipien erfasst, die in den Curricula der einzelnen Schulformen und Bundesländer sehr unterschiedlich repräsentiert sind (vgl. Kap. 2.3). Anders als in Tests zu elaborierten Fachcurricula (wie z. B. in den Schulfächern Mathematik und Naturwissenschaften) gibt es kein allgemein geteiltes Grundverständnis darüber, welche Inhalte genau in der sekundären Bildung vermittelt werden sollen. So dürften die Lösungsquoten jener Aufgaben, die sich auf Inhalte beziehen, welche nur in wenigen Schulformen oder Bundesländern vermittelt werden, nicht zwingend mit solchen Aufgaben korrelieren, die flächendeckend unterrichtet werden. Diese Aufgaben haben folglich auch geringe Trennschärfen. Würde man sie jedoch entfernen, hätte dies zur Folge, dass das Testinstrument an inhaltlicher Breite verliert. Die Inhaltsgebiete des Tests wurden aber gerade unter dem Aspekt ausgewählt, welche ökonomischen Inhaltsstandards wichtig sind, um grundlegende ökonomische Sachverhalte verstehen zu können.

Die Reliabilität des TEL4-G kann bei einem CRONBACHS Alpha von 0,84 (Version A) bzw. 0,87 (Version B) als sehr gut beurteilt werden. Die korrigierten Trennschärfen der beiden Testversionen fallen folgendermaßen aus: 0 < r <0,1: Version A kein Item/ Version B 2 Items; 0.1 < r <0.2: A 8/B 3 Items; 0.2 < r <0.3: A 17/B 8 Items; r > 0.3: A 20/B 32 Items. Aufgaben mit Trennschärfen größer 0,2 können als zufriedenstellend angesehen werden. Bei den Aufgaben mit darunter liegenden Trennschärfen war zu überlegen, ob sie aus den genannten Gründen aus dem Instrument entfernt werden sollten. Eine nochmalige inhaltliche Prüfung ergab jedoch, dass damit die Gefahr der Eliminierung eines bedeutsamen Indikators bestanden hätte (vgl. FASSOTT & EGGERT, 2005) und somit die Konstruktrepräsentativität (Messick, 1989) beeinträchtigt worden wäre. Ohnehin wird häufig empfohlen (z. B. KELAVA & MOOSBRUGGER, 2012), selbst bei reflektiven Messmodellen nur dann Items zu entfernen, wenn ihre Trennschärfen Werte nahe 0 oder im negativen Bereich aufweisen.

Die standardisierten Faktorladungen des CFA-Modells[23] (Minimum A: 0,170/B: 0,022; Maximum A: 0,696/B: 0,759; Durchschnitt A: 0,428/B: 0,476) bzw. die Trenn-

23 Als adäquate latente Modellierung des ökonomischen Wissens und Denkens bietet sich ein Item-Response-Modell (IRT-Modell) oder ein konfirmatorisches Faktormodell (CFA-Modell) an, wobei die bei-

schärfen des analogen 2PL-IRT-Modells (Minimum A: 0,173/B: 0,022; Maximum A: 0,970/B: 1,176; Durchschnitt A: 0,499/B: 0,572) weisen für dichotome Items eine gute Diskriminierung auf. Die sehr guten globalen Fitwerte (s. FN 23) weisen auf die Angemessenheit einer eindimensionalen Modellierung des ökonomischen Wissens und Denkens innerhalb des TEL4-G hin.

Erklärung der Itemschwierigkeiten

Im nächsten Analyseschritt wird geprüft, ob und inwieweit die oben beschriebenen kognitiven schwierigkeitsbestimmenden Merkmale die ermittelten empirischen Itemschwierigkeiten erklären können (vgl. HARTIG, 2007; HARTIG, FREY, NOLD & KLIEME, 2012). Als Indikatoren für die Schwierigkeit können sowohl die klassischen Lösungshäufigkeiten als auch die IRT-Schwierigkeiten herangezogen werden.[24] Beide ergeben in etwa das gleiche Bild, weshalb im Folgenden die Darstellung auf die leichter zu interpretierenden Werte der klassisch errechneten Schwierigkeitsindizes beschränkt wird.

Die mittleren 50 % der Lösungshäufigkeiten liegen zwischen 0,5 und 0,75. Es gibt sieben leichte Aufgaben (Lösungshäufigkeit über 0,8) und keine zu schwierigen Aufgaben (Lösungshäufigkeit unter 0,2).[25] Der Test eignet sich somit für die anvisierte Stichprobe. Die Häufigkeitsverteilung des Gesamtscores der Testversion A (vgl. Abb. 3) verdeutlicht, dass im Mittel mehr als die Hälfte der Fragen beantwortet wird, es jedoch keine Deckeneffekte im Test gibt, da kein Proband alle Aufgaben richtig gelöst hat (Analoges gilt für Version B).

den Modellierungen sich stark ähneln, da sich die Parameter des CFA-Modells in die eines IRT-Modells überführen lassen und vice versa (FINCH, 2005; MUTHÉN, KAO & BURSTEIN, 1991). Somit entspricht das klassische 2PL-IRT-Modell einem klassischen CFA-Modell mit frei geschätzten Faktorladungen während das klassische Raschmodell einem CFA-Modell mit gleichen Faktorladungen für alle Items gleichkommt. Beim TEL4-G zeigt sich, dass die Annahme gleicher Faktorladungen, wie es das klassische Raschmodell bzw. ein restringiertes CFA-Modell, bei dem alle Faktorladungen gleich sind, vorsieht, nicht haltbar ist. Die Fitwerte für das 2PL-Modell (Chi^2=1127,013; df=945; RMSEA = 0,017; CFI = 0,965; WRMR = 0,984) sind deutlich besser als die des 1PL-Modells (Chi^2=2086,166; df=989; RMSEA = 0,400; CFI = 0,787; WRMR = 1,687), was auch durch den angepassten Chi^2-Differenzenttest deutlich wird (Chi^2-Diff= 385,848; df=44; p=0,000). Für die Version B werden vergleichbare Werte zugunsten des 2PL-Modells erreicht. Da es sich um ein dichotomes Datenniveau handelt, wurde der angepasste Chi^2-Differenzentest für den WLSMV-Schätzer der Mplus-Version 7.3 verwendet (vgl. MUTHÉN & MUTHÉN, 1998–2012, S. 625).

24 Eine alternative Modellierung wäre die Nutzung eines Linear-Logistischen Testmodells (LLTM), bei dem bereits während der Rasch-Modellierung die Aufgabenschwierigkeit als Linearkombination der Aufgabenmerkmale geschätzt wird (HARTIG et al., 2012). Allerdings wird bei diesem deterministischen Modell angenommen, dass die Aufgabenmerkmale nahezu perfekt die Itemschwierigkeiten erklären können, was hier gemäß den oben dargestellten Analysen nicht der Fall ist.

25 Zu analogen Ergebnissen führen auch die IRT-Schwierigkeiten, für deren Berechnung der latente Wissenscore auf einen Mittelwert von 0 und eine Standardabweichung bzw. Varianz von 1 normiert wurde: Die Schwierigkeiten liegen meist knapp unter 0, was bedeutet, dass die meisten Aufgaben für diese Stichprobe tendenziell etwas leicht ausfallen.

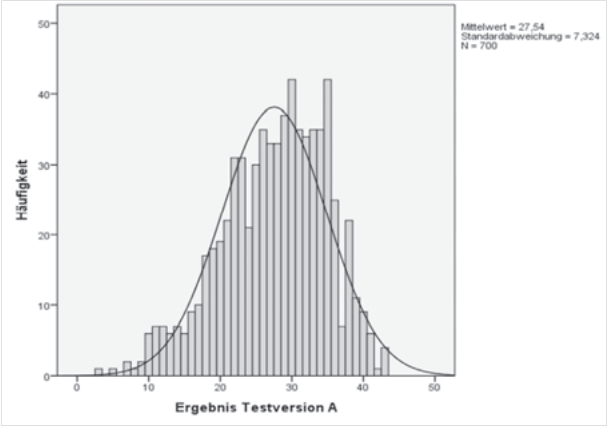

Abb. 3: Balkendiagramm des Wissenscores (Rohwerte) der Version A

Tabelle 5 verdeutlicht, dass die kognitive Prozesskomponente nach Anderson & Krathwohl erwartungskonform hoch mit der Wissenskomponente (r=0,68) und sehr hoch mit den BLOOM'schen Stufen (r=0,92) korreliert. Gleichzeitig zeigen die Daten, dass die Schwellenkonzept-Taxonomiestufen und die Anzahl der Fachtermini nahezu unabhängig voneinander und von den anderen Merkmalen variieren. Die Anzahl der kognitiven Denkoperationen, die man bei einem mustergültigen Lösungsprozess durchführen müsste, korreliert moderat mit den Stufen der Taxonomien von ANDERSON & KRATHWOHL und BLOOM (r jeweils um 0,5). Besonders auffällig ist, dass die einzelnen Kriterien durchweg nicht substantiell mit den empirisch ermittelten Itemschwierigkeiten zusammenhängen.

Tab. 5: Korrelationen zwischen den Kriterien

	Klass. Schw.-keit	IRT Schw.-keit	Bloom Tax.	Anders./ Krathw. Wissens-komp.	Anders./ Krathw. Prozess-komp.	Davies & Mangan Schwellen	Numeracy Komponente	Anzahl Denk-schritte
IRT Schwierigkeit	-,97**							
Bloom-Taxonomie	-,07	,06						
Anderson & Krathw. Wissenskomp.	-,08	,08	,72**					
Anderson & Krathw. Prozesskomp.	-,04	,04	,92**	,68**				
Davies & Mangan Schwellenkonzept	-,15	,14	,16	,12	,16			
Numeracy Komponente	-,08	,09	,25*	,14	,29*	-,13		
Anzahl Denkschritte	,08	-,08	,55**	,52**	,57**	,09	,24*	
Fachtermini Gesamt	,00	,00	-,03	-,08	-,02	,06	-,09	-,14
Signifikanz: * 0,01 ≤ p ≤ 0,05; ** p ≤ 0,01								

4. Diskussion der Befunde

4.1 Itemschwierigkeit und Aufgabenlösungsstrategien

Die Analysen zu den kognitiven Merkmalen haben ergeben, dass weder die denk-psychologischen noch die sprachanalytischen noch auch die formalen Kriterien einen substanziellen Beitrag zur Aufklärung der Varianz der Itemschwierigkeiten leisten.[26] Wie eingangs erwähnt, fügen sich damit unsere Befunde in eine Reihe vergleichbarer Resultate ein[27], die mit Instrumenten zur Erfassung fachbezogener kognitiver Leistungen gewonnen worden sind (z. B. NICKOLAUS, 2016, 169–170; ABELE, 2016, S. 43; Abele et al. 2016, S. 179, 201, FN19)[28]. Als eine Besonderheit dürfte im vorliegenden Fall eine Rolle spielen, dass die Inhalte des Tests, wie oben bereits erwähnt, wenn überhaupt, nicht auf Basis eines einheitlichen Curriculums unterrichtet werden, so dass davon auszugehen ist, dass die Probanden zum Testzeitpunkt über sehr unterschiedliche inhaltsbezogene kognitive Strukturen verfügen.[29] Unter diesen Umständen wird man damit rechnen müssen, dass die Lösungsstrategien zwischen den Probanden differieren.

Neben den aus der Expertiseforschung bekannten allgemeinen (1a) vorwärts- oder (1b) rückwärtsgerichteten, konstruktrelevanten und insoweit vorhersehbaren Lösungsstrategien, deren Anwendung in Abhängigkeit vom Expertisegrad sowie von Aufgabenmerkmalen interindividuell variiert (ABELE, 2016; ERICSSON, 1996), können weitere eher konstruktirrelevante Lösungsprozesse auftreten, nämlich (2) Heuristiken, (3) auf *testwiseness* beruhende Lösungsstrategien, (4) Intuition und (5) Zufallsraten. Diese

26 Es wäre zwar denkbar, dass Interaktionseffekte verschiedener kognitiver Merkmale einen Mehrwert bei der Varianzaufklärung leisten. Bspw. könnte ein komplizierter Lösungsprozess mit vielen Denkschritten dann deutlich schwieriger sein, wenn er ein Schwellenkonzept (DAVIES & MANGAN, 2007) enthält. Allerdings stößt die Prüfung solcher Interaktionen im Rahmen dieser Studie an ihre methodischen Grenzen und könnte allenfalls durch die Spezifizierung sog. „item models" und „evidence models" (Mislevy & Haertel, 2006) ermöglicht werden – eine Vorgehensweise, die im Rahmen dieser Darstellung nicht entfaltet werden kann.

27 In ihrer Studie zur fachspezifischen Problemlösefähigkeit von Kfz-Mechatronikern fanden NICKOLAUS, ABELE, GSCHWENDTNER, NITSCHKE und GREIFF (2012, S. 253) allerdings durchaus beachtliche signifikante Korrelationen zwischen den „üblichen" schwierigkeitsbestimmenden Merkmalen und der „Test-Schwierigkeit" von fehlerdiagnostischen Aufgaben (u. a. „Art der Informationsbeschaffung .75, Modellierungsumfang .77, Aufgabenkomplexität .90). Dies ist umso bemerkenswerter, als diese Merkmale einen eher „allgemeinen", domänen-unspezifischen Charakter zu haben scheinen. Ihre Operationalisierung wird in dieser Quelle nur teilweise genauer angegeben. Es wäre hier noch zu prüfen, ob die Größenordnung dieser Werte auch jenseits des Designs dieser Studie (N = 16) erhalten bleibt. – Auch in NICKOLAUS, GSCHWENDTNER und ABELE (2009, S. 10–12) sowie in dort angegebenen weiteren Vorstudien dieser Forschergruppe werden z. T. enge Zusammenhänge mit verschiedenen schwierigkeitsbestimmenden Merkmalen berichtet, darunter insbesondere die BLOOM'sche Taxonomie, wobei, wie oben bereits angemerkt, die Zuordnung von Testaufgaben zu den Taxonomiestufen ihre eigenen Reliabilitätsprobleme aufweist und in den Daten der – hier zitierten – Hauptstudie offenbar keine „hinreichende Varianz" der „vermuteten Schwierigkeitsparameter" realisiert werden konnte (S. 15).

28 So ergibt auch eine nachträgliche Analyse der Daten aus der Vorgängerstudie mit dem WBT/TEL2 (vgl. BECK et al., 1998) eine Korrelation zwischen („klassischer") Itemschwierigkeit und Taxonomiestufe nach BLOOM von r = -.20 für Form A und r = -.28 für Form B.

29 Eine ausreichend hohe Trennschärfe bei *allen* Items zu erreichen, ohne die inhaltliche Breite des Testinstruments aufzugeben, erweist sich daher als kaum möglich. Das gleiche gilt für den Versuch, eine klare, mehrdimensionale Faktorstruktur im Testverfahren zu etablieren.

sind problematisch, da sie den vorgesehenen und erwarteten Lösungsprozess auf unvorhersehbare Weise verändern, indem sie etwa fehlendes Wissen oder erforderliche Denkprozesse durch abseitige Überlegungen substituieren.

(1a) Bei den *vorwärtsgerichteten*, konstruktrelevanten Lösungsstrategien, wie sie von Experten angewandt werden (s. u. a. BRÜCKNER & PELLEGRINO, 2016), durchsuchen die Probanden nach Aufnahme, Verarbeitung und Evaluation der im Aufgabenstamm gegebenen Informationen und einer daraus resultierenden Zielformulierung den Lösungsraum nach geeigneten Lösungen und gleichen diese im Anschluss mit den gegebenen Informationen in den Antwortoptionen ab. Trifft ein Lösungsversuch dabei allerdings auf einen unbekannten Zentralbegriff, kann allein dieser Umstand zum Abbruch dieser Vorgehensweise (und möglicherweise zur Nutzung der Rückwärtsstrategie) führen. Das bedeutet, dass die Lösung einer Aufgabe, die unter einem taxonomischen Aspekt insgesamt als eher anspruchsvoll einzustufen ist, an einem Hindernis scheitert, das taxonomisch auf einer niedrigen Stufe („Wissen") einzuordnen ist. In diesem Fall wäre es gar nicht die „objektive" (taxonomische) Schwierigkeit der Aufgabe als Ganzer, an der ihre Lösung scheitert, sondern ein kognitiv eher anspruchsloser erster Teilschritt. Diese Unterscheidung wird jedoch im Schwierigkeitsindex nicht sichtbar, der nicht abbilden kann, dass bzw. ob ein Proband auch bei Kenntnis des Zentralbegriffs nicht zu einer richtigen Lösung gelangt wäre, da der Schwierigkeitsindex ja immer das Resultat, nicht jedoch den Prozess der Lösung zu erfassen vermag.

(1b) Anders liegen die Dinge bei den *Rückwärtsstrategien*, die von den vorgegebenen Lösungsmöglichkeiten ausgehen und zu denen eher Novizen greifen. Sie entsprechen allerdings nicht den Erwartungen der Testentwickler, weil sie verkürzte oder inkohärente Wege gehen. So nehmen bspw. Probanden, die die korrekte Lösung nicht *lege artis* herleiten können, zunächst eine der angebotenen Möglichkeiten als richtig an und versuchen, den Lösungsweg zu rekonstruieren. Hierbei schreiten sie rückwärts hin zur Problemformulierung, um zu prüfen, ob sie dort „ankommen" oder ob sie dabei an einer „objektiven" Hürde scheitern; anderenfalls wählen sie diese Lösung. Dieses Prozedere kann prinzipiell für alle (vier) angebotenen Lösungen abgearbeitet und abschließend nach dem Kriterium der „kleinsten Hürde" entschieden werden (s. dazu weiter unten)[30]. Sowohl vorwärts- als auch rückwärtsgerichtete Lösungsstrategien stellen Aggregate des Lösungsverhaltens dar, die ihrerseits weitere, granular feinere Vorgehensweisen und Prozesse umfassen (s. ausführlicher Brückner & Pellegrino, 2017).[31]

30 In der Logik des Wissensaufbaus und der Wissensnutzung kommen hier, wie MINNAMEIER im Anschluss an Peirce präzisiert, bevorzugt theorematische Varianten abduktiver, deduktiver und induktiver Schlüsse sowie die Nutzung von Analogien zum Einsatz (vgl. dazu die differenzierte Darstellung bei MINNAMEIER (2005, S. 197–218)).

31 Für den Bereich von Problemlösungen (dargestellt an physikalisch-technischen Beispielen) schlagen KORELYAKOV und LANDA (1982) eine differenziertere, sechsdimensionale Abstufungsmöglichkeit der Qualität der Denkprozesse vor (nach Theoretizität, Vollständigkeit, Korrektheit, Allgemeinheit, Tiefe und Stringenz), die allerdings lediglich binär kodiert wird (gegeben vs. nicht gegeben), dabei jedoch schon 36 plausible Varianten zu unterscheiden erlaubt (deren Zahl sich exponentiell erhöht, wenn man in den sechs Kriterien Abstufungen zulässt, die allerdings, pädagogisch und insbesondere diagnostisch gesehen, durchaus bedeutsam sind.

(2) *Heuristiken* sind Strategien, die trotz unvollständigen Wissens[32] bzw. trotz unterkomplexen Denkens und trotz restringierter Bearbeitungszeit zur Lösung eines Problems führen können (vgl. z. B. GIGERENZER, 2008). Das Verfahren besteht im Grundsatz in der Anwendung einer Schrittfolge, auch wenn es häufig unreflektiert eingesetzt wird. Bspw. bewertet man bestimmte (ökonomische) Konzepte und Prinzipien im Hinblick auf die zu wählende Lösung unter dem Aspekt ihrer Plausibilität, wobei für diese Einschätzung unterschiedliche Kriterien angelegt werden können (z. B. die begriffliche Nähe zur Problemformulierung, die interne Konsistenz der Lösungsaussage, ihre subjektive Bekanntheit oder ihre sprachliche Fassung), die das Lösungsverhalten verkürzen und einer detaillierten Exploration entgegenstehen.

(3) Die auf „*Testwiseness*" (MILLMAN, BISHOP & EBEL, 1965) beruhenden Lösungsstrategien sind auf Erfahrungen mit einem gegebenen Test- oder Aufgabentyp zurückzuführen (u. U. in einer gegebenen Domäne). „Testwiseness" besteht darin, dass Testanden Vermutungen darüber entwickelt haben, an welchen äußeren oder auch inhaltlichen, aber in der Sache irrelevanten Merkmalen richtige Lösungen zu erkennen sind. Diese Vermutungen können zu überzufällig häufigen richtigen Aufgabenlösungen führen, wenngleich sie auf Überlegungen beruhen, die nicht auf Leistungen rekurrieren, welche durch den Test erfasst werden sollten.

(4) Eine weitere Strategie besteht darin, einer *Intuition* zu folgen (HARTEIS & BILLETT, 2013). Sie hebt spontan eine der angebotenen Möglichkeiten vor den anderen heraus, ohne dass eine kontrollierende verstandesmäßige Kognition eingeschaltet wäre und ohne dass über den Verweis auf die „Eingebung" hinaus nachträglich Gründe für das Auswahlergebnis angegeben werden können. Im hier vertretenen Verständnis unterscheidet sie sich von der „Expertenintuition" vor allem darin, dass diese vom Experten selbst rational rekonstruiert werden kann; beide sind jedoch von einem Richtigkeitsgefühl begleitet. Auch wenn die damit ins Spiel kommenden Abgrenzungsfragen noch einer weiteren Klärung bedürfen, soll hier das Phänomen „Intuition" als eigenständig berücksichtigungswürdig festgehalten und nicht mit Heuristiken oder Testwiseness gleichgesetzt werden, die ihrer je eigenen subjektiven „Logik" folgen.

(5) Schließlich ist mit einem reinen Zufallsraten zu rechnen, das wegen fehlender anderer Strategien oder auch wegen fehlender Bearbeitungsmotivation vor allem bei geschlossenen Aufgabenformaten zum Einsatz kommen kann (s. BRÜCKNER, 2017). Es handelt sich dabei um die Anwendung einer bewusst und absichtsvoll durchgeführten Ratehandlung, deren Ergebnis im Unterschied zu allen anderen Strategien mit keiner Richtigkeitsvermutung verbunden ist, sondern auf einer form- und inhaltsunabhängigen Auswahl beruht, wie sie näherungsweise in einem statistischen Zufallsexperiment modelliert wird.

32 Ob unter diesen Begriff auch das vorhandene sog. implizite (also nicht sagbare) Wissen zu rechnen ist, wie es im tacit knowledge-Ansatz verstanden wird (NEUWEG, 2015), sei hier dahingestellt, obwohl zwischen ihm und dem nicht vorhandenen Wissen zweifellos eine wesentliche qualitative Differenz besteht (vgl. auch NICKOLAUS, ABELE & SCHMIDT, 2014).

Intuition und Zufallsraten fallen in die Gruppe der spontanen Lösungshandlungen, die sich durch ihre schnelle Generierung von der Gruppe der zuvor angeführten (mehr oder weniger) reflexionsgeleiteten Lösungshandlungen unterscheiden. Es liegt auf der Hand, dass mehrere der genannten Strategien während eines Aufgabenlösungsprozesses in wechselseitige Abhängigkeit zueinander treten, kombiniert oder kompensatorisch verwendet werden können. So ergab eine aktuelle Studie, dass die Anwendung von Heuristiken in einem negativen Zusammenhang mit vorwärtsgerichteten Lösungsstrategien stehen kann (BRÜCKNER & PELLEGRINO, 2017) und dass manche Strategien sowohl kompensatorisch als auch zusätzlich zur Absicherung einer Lösung Anwendung finden (BRÜCKNER, 2017). Hierzu besteht jedoch noch erheblicher Forschungsbedarf.

Als Verfahren zur Erschließung der empirisch ablaufenden latenten Lösungsprozesse haben sich – nicht zuletzt durch Arbeiten in der Expertiseforschung – kognitive Interviews etabliert (LEIGHTON, 2004). Je nach Untersuchungsfokus können in diesen Verfahren verschiedenartige Techniken zur Erfassung des Lösungsverhaltens während der Aufgabenbearbeitung eingesetzt werden (z. B. die Methode des lauten Denkens, ggf. in Kombination mit Eye-Tracking) und einen Schluss auf die latenten Lösungsprozesse sowie, damit verbunden, auf schwierigkeitserzeugende Merkmale zulassen. Vielversprechende systematische Erfahrungen liegen uns dazu aus Studien zu einem weiteren ökonomiebezogenen Test, dem TUCE, vor (s. BRÜCKNER & PELLEGRINO, 2016; BRÜCKNER, 2017).

Taxonomische Einordnung und vorwärtsgerichtete Modellierungen des Lösungsverhaltens, wie wir sie oben dargestellt haben und wie sie häufig als wesentliche Argumente in der Testkonstruktion und -beurteilung herangezogen werden, bieten – nach allem – für die Beschreibung schwierigkeitsbestimmender Merkmale bestenfalls eine grobe Orientierung. Und über die bereits angesprochenen weiteren schwierigkeitsbeeinflussenden Faktoren hinaus bedarf es bei der Erforschung der Bedingungen des Lösungsprozesses, wie MISLEVY (2007; 2016) aus Sicht des von ihm entwickelten soziokognitiven Ansatzes zu Recht hervorhebt, der Betrachtung noch weiterer Merkmale, wie insbesondere der Vertrautheit mit dem Gegenstandsfeld des Tests, den ein Proband bearbeitet sowie des situationalen Kontexts, in dem die Testbearbeitung erfolgt.

4.2 Itemschwierigkeit, Motivation und Testkonzeption

Ein weiterer Ursachenkomplex, der bei der Analyse der schwierigkeitsbestimmenden Merkmale der Berücksichtigung bedarf, ist in den motivationalen und volitionalen dispositiven Merkmalen zu sehen, die das Lösungsverhalten der Probanden moderieren können (ALEXANDER, KULIKOWICH & SCHULZE, 1994). Insbesondere unter low-stakes Testbedingungen mit ihrem eher geringen Erfolgsrisiko gewinnen diese Merkmale an Einfluss und erschweren den Rückschluss von Test(kenn)werten (z. B. Aufgabenschwierigkeit, Lösungsleistung) auf die zugrunde liegenden Lösungsprozesse. Häufig werden sie als Testteilnahme- und Testbearbeitungsmotivation mittels distalen, auf Selbsteinschätzungen basierenden Verfahren erhoben (LIU, BRIDGEMAN & ADLER,

2012) und weisen insbesondere im Schulbereich einen positiven Zusammenhang zu den Testwerten auf (u. a. in den PISA-Studien; JUDE & KLIEME, 2010). Ähnliche Befunde stellen sich auch im Hochschulbereich ein (COLE & OSTERLIND, 2008; LIU et al., 2012).

Einen anderen Indikator für die Testmotivation liefert der „response time effort" (RTE) (s. WISE & DEMARS, 2005). Studien haben gezeigt, dass RTE signifikant mit der Selbsteinschätzung der Studierenden korreliert (LIU et al, 2012, S. 353). Ebenfalls zu den indirekten, jedoch dem Motivationskonstrukt etwas näheren Verfahren rechnet die Analyse von Itempositionseffekten (WEIRICH, HECHT & BÖHME, 2014), die über komplexe Testheftdesigns realisiert werden. Die Annahme, dass bei gegebenen kognitiven Merkmalen mit fortschreitender Testbearbeitung die Motivation mit jeder weiteren Aufgabe sinkt und daher die Items am Anfang eines Testheftes vergleichsweise häufiger korrekt beantwortet werden, lässt sich mittels Kontrolle der Lösungshäufigkeit in Abhängigkeit von ihrer Platzierung prüfen. Der Grund für solche Differenzen kann auch in Ermüdungserscheinungen liegen, die ihrerseits auf die Motivation wirken, jedoch differentialdiagnostisch von anderen Motivationsquellen zu unterscheiden wären.[33] Die Umsetzung dieses Messkonzepts erfordert jedoch große Stichproben und elaborierte Analyseverfahren. In zukünftigen Studien sollten – gerade unter low-stakes Testbedingungen – solche Itemrotationen mittels mehrerer Testhefte vorgenommen werden, um für diese Merkmale des Lösungsverhaltens kontrollieren zu können.

4.3 Einige Folgerungen

Insgesamt bleibt festzuhalten, dass die Gestaltung bzw. Rekonstruktion schwierigkeitserzeugender Merkmale von Testaufgaben *ausschließlich entlang inhaltsunabhängiger kognitiver Kriterien* nicht hinreicht. Das dürfte insbesondere für Messinstrumente gelten, mit denen das Verständnis eher eng umschriebener domänenspezifischer Konzepte und Zusammenhänge erfasst werden soll. Hier gilt es, jene Aufgabenmerkmale mit heranzuziehen, in denen die Kompliziertheit und die Komplexität der thematisierten Inhalte zum Ausdruck kommen. So werden etwa in den Wirtschaftswissenschaften Fragestellungen, in denen ein auftretender ökonomischer Effekt multikausal erklärt werden soll (z. B. Stagflation) höhere Schwierigkeitsgrade aufweisen als jene, in denen lediglich eine oder wenige Ursachen unter modellhaften Konstanzbedingungen („ceteris-paribus"-Klausel) erfragt werden, obwohl sie mittels identischer sprachlicher Strukturen präsentiert werden. Analoges dürfte, wie man sich leicht verdeutlichen kann, bspw. im Bereich der Fehlersuche bei mechanisch und/oder elektronisch arbeitenden Geräten gelten. Zumindest alle wissenschaftlich bearbeiteten Domänen lassen sich unter solchen *inhaltlichen* Perspektiven nach schwierigkeitsbestimmenden Kriterien beschreiben, die in der *grammatischen Struktur* der Aufgabenstellungen ebenso wenig sichtbar werden (müssen) wie in der Angabe einer *kognitiven Verarbeitungsstufe* (z. B. sensu BLOOM).

33 Dazu verspricht die prozessbezogene Erfassung von Bio-Daten die zuverlässigsten Informationen (hier z. B. EMG).

Den oben angewandten Kriterien wäre demnach eine fachimmanente Komplexitätsanalyse hinzuzufügen, die es erlaubt, zwischen mehr oder weniger voraussetzungsvollen Konzepten und ihren mehr oder weniger definiten Bezügen zu differenzieren (z. B. Nachfrageänderung als Differenz zwischen zwei Zeitpunkten und ihre Wirkung auf die Angebotsmenge vs. Nachfrageänderung als Funktion von Preiselastizität in ihrer Wirkung auf die Angebotsmenge). Allenfalls im oben skizzierten Schwellenkonzept nach DAVIES und MANGAN (2007) könnte eine erste Annäherung an inhaltsorientierte Modellierungsweisen gesehen werden, indem es den (niedrigeren oder höheren) inhaltserschließenden Status des einzelnen ökonomischen Konzepts ermittelt.[34]

Nach allem dürfte deutlich geworden sein, dass eine genaue, *theoretisch umfassende* Abschätzung der Schwierigkeit von Testitems die Berücksichtigung von vier schwierigkeitsmoderierenden Quellen erforderlich macht: sowohl die *konstruktrelevanten kognitiven*, nämlich denkpsychologisch-formale und inhaltlich-konzeptuelle Merkmale der Aufgaben als auch die *konstruktirrelevanten situationalen*, nämlich motivationale und bearbeitungsstrategische Merkmale der Personen. Zumindest unter einem mikroanalytischen Aspekt bedürfen sie der Berücksichtigung, wenn belastbare individualdiagnostische Befunde gewonnen werden sollen. Das bedeutet andererseits freilich nicht, dass Testresultate des hier besprochenen Typs einem pauschalen Geltungsvorbehalt unterlägen. Vielmehr erweisen sie sich insbesondere dort als theoretisch und praktisch tragfähig und nützlich, wo es um die Ermittlung von Gruppenmerkmalen und Verteilungskennwerten geht (differenziert etwa nach Geschlecht, Studienphase, Migrationshintergrund usw.), wie sie bspw. zur Beschreibung von Stand und Entwicklung gemessener Kompetenzen in Kohorten herangezogen oder zur Fundierung von curricularen und (hochschul-)didaktischen Maßnahmen ausgewertet werden (vgl. z. B. HAPP 2017; SCHMIDT 2017).

Literatur

ABELE, S. (2016). Umgang mit Komplexität: Eine bedeutsame psychische Voraussetzung des domänespezifischen Problemlösens? *Zeitschrift für Berufs- und Wirtschaftspädagogik, 112*(1), 37–59.

ABELE, S., BEHRENDT, S., WEBER, W. & NICKOLAUS, R. (2016). Berufsfachliche Kompetenzen von Kfz-Mechatronikern – Messverfahren, Kompetenzdimensionen und erzielte Leistungen (KOKO Kfz). In K. BECK, M. LANDENBERGER & F. OSER (Hrsg.), *Technologiebasierte Kompetenzmessung in der beruflichen Bildung. Ergebnisse aus der BMBF-Förderinitiative ASCOT* (S. 171–203). Bielefeld: Bertelsmann.

34 Eine weitere Differenzierung, nämlich nach „domain knowledge" und „topic knowledge" (ALEXANDER et al., 1994), könnte einen zusätzlichen Beitrag zur testtheoretischen Bearbeitung des Problems domäneimmanenter Schwierigkeitsabstufungen leisten. Danach werden neben eher grundlegenden Konzepten systematisch auch Konzepte in die Aufgabenstellungen übernommen, die ein spezialisierteres und tieferes Verständnis erfordern und so eine Varianz in den Lösungsleistungen erzeugen, deren Analyse den Gehalt der Testwertinterpretation substanziell zu erhöhen vermag.

ALEXANDER, P. A., KULIKOWICH J. M. & SCHULZE S. K. (1994). The influence of topic knowledge, domain knowledge, and interest on the comprehension of scientific exposition. *Learning and Individual Differences, 6*, 379–397.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION (AERA), AMERICAN PSYCHOLOGICAL ASSOCIATION (APA), & NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION (NCME) (2014). *Standards for Educational and Psychological Testing*. Washington. DC: American Educational Research Association.

ANDERSON, L. W. & KRATHWOHL, D. R. (Hrsg.). (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives.* Boston: Allyn & Bacon (Pearson Education Group).

ARFFMAN, I. (2013). Problems and Issues in Translating International Educational Achievement Test. *Educational Measurement: Issues and Practice, 32*(2), 2–14. DOI: 10.1111/emip.12007

BECK, K. (1991). Economic Literacy in German Speaking Countries and the United States. First Steps to a Comparative Study. *Economia, 1*, 17–23.

BECK, K. & KRUMM, V. (1992). *Economic Literacy in the United States, Germany, and Austria: Results of cross national studies.* Paper presented at the Annual Meeting of the Joint Council on Economic Education and the National Association on Economic Education JCEE/NAEE, Los Angeles, USA. Nov. 09, 1990. *ERIC*. Microfiche No. ED 340 629, 1–65.

BECK, K., KRUMM, V. & DUBS, R. (1998). *Wirtschaftskundlicher Bildungs-Test* (WBT). Göttingen: Hogrefe.

BECK, K., LANDENBERGER, M. & OSER, F. (Hrsg.) (2016). *Technologiebasierte Kompetenzmessung in der beruflichen Bildung*. Bielefeld: Bertelsmann.

BLOOM, B. S. (1956). (Hg.). *Taxonomy of Educational Objectives: The Classification of Educational Goals*. New York: McKay.

BRAY, M., ADAMSON, B., & MASON, M. (2007). *Comparative education research – approaches and methods*. Hong Kong, China: Springer.

BRÜCKNER, S. (2017). *Prozessbezogene Validierung anhand von mentalen Operationen bei der Bearbeitung wirtschaftswissenschaftlicher Testaufgaben.* Landau: Verlag Empirische Pädagogik (im Druck).

BRÜCKNER, S., FÖRSTER, M., ZLATKIN-TROITSCHANSKAIA, O. & WALSTAD, W. B. (2015a). Effects of prior economic education, native language, and gender on economic knowledge of first-year students in higher education. A comparative study between Germany and the USA. *Studies in Higher Education, 40*(3), 437–453.

BRÜCKNER, S. FÖRSTER, M. ZLATKIN-TROITSCHANSKAIA, O., HAPP, R., WALSTAD, W. B., YAMAOKA, M. & ASANO, T. (2015b). Gender Effects in Assessment of Economic Knowledge and Understanding: Differences Among Undergraduate Business and Economics Students in Germany, Japan, and the United States. *Peabody Journal of Education, 90*(4), 503–518.

BRÜCKNER, S. & PELLEGRINO, J. W. (2016). Integrating the Analysis of Mental Operations into Multilevel Models to Validate an Assessment of Higher Education Students' Competency in Business and Economics. *Journal of Educational Measurement, 53(3),* 293–312.

BRÜCKNER, S. & PELLEGRINO, J. W. (2017). Contributions of Response Processes Analysis to the Validation of an Assessment of Higher Education Students' Competency in Business and Economics (Chap. 3). In B. Zumbo & A. Hubley (Eds.), *Understanding and Investigating Response Processes in Validation Research* (pp. 31–52). New York: Springer.

BUCKLES, S. & SIEGFRIED, J. (2006). Using multiple-choice questions to evaluate in-depth learning of economics. *Journal of Economic Education, 37*(1), 48–57.

COLE, J. S., & OSTERLIND, S. J. (2008). Investigating differences between low- and high-stakes test performance on a general education exam. *Journal of General Education, 57*, 119–130.

COUNCIL FOR ECONOMIC EDUCATION (CEE) (2010). *Voluntary national content standards in economics*. New York: Council for Economic Education.

DAMMAN, E., BEHRENDT, S., ŞTEFÁNICĂ, F. & NICKOLAUS, R. (2016). Erreichte Kompetenzniveaus in der ingenieurwissenschaftlichen Grundbildung – Analyse am Beispiel der Technischen Mechanik. *Zeitschrift für Erziehungswissenschaft, 19*(2), 351–374.

DAVIS, B. G. (2001). *Tools for teaching.* San Francisco, CA: Jossey-Bass.

DAVIES, P. & MANGAN, J. (2007). Threshold concepts and the integration of understanding in economics, *Studies in Higher Education, 32*(6), 711–726.

DEUTSCHE GESELLSCHAFT FÜR ÖKONOMISCHE BILDUNG (DEGÖB) (2004). Kompetenzen der ökonomischen Bildung für allgemeinbildende Schulen und Bildungsstandards für den mittleren Schulabschluss. Abruf am 22.07.2016 unter http://degoeb.de/uploads/degoeb/04_DEGOEB_Sekundarstufe-I.pdf

DRAXLER, D. (2005). Aufgabendesign und basismodellorientierter Physikunterricht (Dissertation). Unv. Duisburg-Essen: Universität Duisburg-Essen. Abruf am 22.07.2016 unter http://duepublico.uni-duisburg-essen.de/servlets/DocumentServlet?id=14098

ERICSSON, K. A. (Ed.). (1996). *The road to excellence: The acquisition of expert performance in the arts and sciences, sports, and games.* Mahwah, N. J.: Erlbaum.

FASSOTT, G., & EGGERT, A. (2005). Zur Verwendung formativer und reflektiver Indikatoren in Strukturgleichungsmodellen: Bestandsaufnahme und Anwendungsempfehlungen. In F. BLIEMEL, A. EGGERT, G. FASSOTT, & J. HENSELER (Hrsg.), *Handbuch PLS-Pfadmodellierung. Methode, Anwendung, Praxisbeispiele* (S. 31–47). Stuttgart: Schäffer-Poeschel.

FINCH, H. (2005). The MIMIC Model as a Method for Detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT Likelihood Ratio. *Applied Psychological Measurement, 29*, 278–295.

FÖRSTER, M., ZLATKIN-TROITSCHANSKAIA, O. & HAPP, R. (2015). *Adapting and Validating the Test of Economic Literacy to Assess the Prior Economic Knowledge of First-Year Students in Business and Economic Studies in Germany* (Discussion Paper; Annual Meeting of the American Economic Association). Boston: AEA.

FÖRSTER, M., BRÜCKNER, S. & ZLATKIN-TROITSCHANSKAIA, O. (2015a). Assessing the Financial Knowledge of University Students in Germany. Empirical Research in Vocational Education and Training, *7*(6), 1–20.

FÖRSTER, M., ZLATKIN-TROITSCHANSKAIA, O., BRÜCKNER, S., HAPP, R., HAMBLETON, R. K., WALSTAD, W. B., ASANO, T. & YAMAOKA, M. (2015b). Validating Test Score Interpretations by Cross-National Comparison: Comparing the Results of Students From Japan and Germany on an American Test of Economic Knowledge in Higher Education. *Zeitschrift für Psychologie, 223*(1), 14–23.

FREY, A. & HARTIG, J. (2013). Wann sollten computerbasierte Verfahren zur Messung von Kompetenzen anstelle von Papier- und Bleistiftbasierten Verfahren eingesetzt werden? *Zeitschrift für Erziehungswissenschaft, 16*(Sonderheft 1), 53–57.

GIGERENZER, G. (2008). *Bauchentscheidungen: die Intelligenz des Unbewussten und die Macht der Intuition.* München: Goldmann.

HAMBLETON, R. K. (2001). The Next Generation of the ITC Test Translation and Adaption Guidelines. *European Journal of Psychological Assessment, 17*, 164–172.

HAPP, R. (2017). *Die Entwicklung des volkswirtschaftlichen Grundlagenwissens im Studienverlauf – Effekte von Eingangsvoraussetzungen auf den Wissenserwerb.* Landau: Verlag Empirische Pädagogik (im Druck).

HARKNESS, J. A. (2008). Comparative survey research: Goals and challenges. In E. D. DE LEEUW, J. J. HOX, & D. A. DILLMAN (Hrsg.), *International handbook of survey methodology* (pp. 56–77). New York: L. Erlbaum Associates.

HARTEIS, C. & BILLETT, S. (2013). Intuitive expertise: Theories and empirical evidence. *Educational Research Review, 9*, 145–157. doi:10.1016/j.edurev.2013.02.001

HARTIG, J. (2007). Skalierung und Definition von Kompetenzniveaus. In E. KLIEME & B. BECK (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie* (S. 83–99). Weinheim: Beltz.

HARTIG, J., FREY, A., NOLD, G. & KLIEME, E. (2012). An Application of Explanatory Item Response Modeling for Mode-Based Proficiency Scaling, *Educational and Psychological Measurement, 72*(4), 665–686.

HARTMANN, G. B. (2015). *Gesamtwirtschaftliche Aspekte – Industrie* (9. Aufl.). Rinteln: Merkur Verlag.

HONTHEIM, T. (2016). *Ökonomische Bildung von Jugendlichen in Rheinland-Pfalz – Bedeutung der curricularen Verankerung volkswirtschaftlicher Inhalte in der Sekundarstufe II.* Lehrstuhl für Wirtschaftspädagogik: Mainz (unveröffentlichte Masterarbeit).

INTERNATIONAL TEST COMMISSION (ITC) (2005). ITC Guidelines for Translating and Adapting Tests. Abruf am 22.07.2016 unter http://www.intestcom.org/files/guideline_test_adaptation.pdf

JUDE, N. & KLIEME, E. (2010). Das Programme for International Student Assessment (PISA). In E. KLIEME, C. ARTELT, J. HARTIG, N. JUDE, O. KÖLLER, M. PRENZEL, W. SCHNEIDER & P. STANAT (Hrsg.), *PISA 2009. Bilanz nach einem Jahrzehnt* (S. 11–22). Münster [u. a.]: Waxmann.

KANE, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement, 50*(1), 1–73.

KELAVA, A., & MOOSBRUGGER, H. (2012). Deskriptivstatistische Evaluation von Items (Itemanalyse) und Testwertverteilungen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion.* (S. 75–102). Berlin: Springer.

KORELYAKOV, YU. A. & LANDA, L. N. (1982). On parametric approaches to the analysis and description of thought processes. *Instructional Science, 11*, 29–50.

KRUGMAN, P., OBSTFELD, M. & MELITZ, M. (2011): *Internationale Wirtschaft – Theorie und Politik der Außenwirtschaft* (9. Aufl.). Hallbergmoos: Pearson Studium.

KUHN, C., ZLATKIN-TROITSCHANSKAIA, O., PANT, H., & HANNOVER, B. (2016). Valide Erfassung der Kompetenzen von Studierenden in der Hochschulbildung. Eine kritische Betrachtung des nationalen Forschungsstandes. *Zeitschrift für Erziehungswissenschaften, 16*(1), 1–24. DOI: 10.1007/s11618-016-0673-7.

KUTSCHA, G. (1975). *Ökonomie an Gymnasien. Ziele, Konflikte, Konstruktionen*. München: Kösel.

LEIGHTON, J. P. (2004). Avoiding Misconception, Misuse, and Missed Opportunities: The Collection of Verbal Reports in Educational Achievement Testing. *Educational Measurement: Issues and Practice, 23*(4), 6–15.

LIU, O. L., BRIDGEMAN, B. & ADLER, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher, 41*, 352–362.

LORZ, O. & SIEBERT, H. (2007). *Einführung in die Volkswirtschaftslehre* (15. Aufl.). Stuttgart: Kohlhammer.

MANKIW, N. G. & TAYLOR, M. P. (2012). *Grundzüge der Volkswirtschaftslehre* (5. Aufl.). Stuttgart: Schäffer-Poeschel.

MESSICK, S. (1989b). Validity. In R. L. Linn (Hrsg.), Educational Measurement (3. Aufl., S. 13–103). New York: Macmillan Publishing.

MEYER, J., & LAND, R. (2006). Threshold concepts: an introduction. In J. MEYER & R. LAND (Eds.), *Overcoming barriers to student understanding: Threshold concepts and troublesome knowledge* (S. 3–18). London, New York: Routledge.

MILLMAN, J., BISHOP, C. H., & EBEL, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement, 25*(3), 707–726.

MINISTERIUM FÜR KULTUS, JUGEND UND SPORT (MKJS) BADEN-WÜRTTEMBERG (Hrsg.) (2016). *Bildungsplan 2016: Allgemeinbildende Schulen Gymnasium (Endfassung) –Wirtschaft.* Stuttgart: MKJS.

MINNAMEIER, G. (2000). *Entwicklung und Lernen – kontinuierlich oder diskontinuierlich?* Münster: Waxmann.

MINNAMEIER, G. (2005). *Wissen und inferentielles Denken. Zur Analyse und Gestaltung von Lehr-Lern-Prozessen.* Frankfurt: P. Lang.

MISLEVY, R. J. & HAERTEL, G. D. (2006). Implications of Evidence-Centered Design for Educational Testing. *Educational Measurement: Issues and Practice, 25*(4), 6–20.

MISLEVY, R. J. (2007). Validity by Design. *Educational Researcher, 36*(8), 463–469.

MISLEVY, R. J. (2016). How Developments in Psychology and Technology Challenge Validity Argumentation. *Journal of Educational Measurement, 53*(3), 265–292.

MÖHLMEIER, H., SKORZENSKI, F., WIERICHS, G. & WURM, G. (2015). *Allgemeine Wirtschaftslehre für den Bankkaufmann/die Bankkauffrau* (11. Aufl.). Troisdorf: Bildungsverlag EINS.

MUTHÉN, B., KAO, C.-F. & BURSTEIN, L. (1991). Instructionally Sensitive Psychometrics: Application of a new IRT-Based Detection Technique to Mathematics Achievement Test Items. *Journal of Educational Measurement, 28*(1), 1–22.

MUTHÉN, L. K. & MUTHÉN, B. O. (1998–2012). *Mplus User's Guide*. (Seventh Edition). Los Angeles, CA: Muthén & Muthén.

NEUWEG, G. H. (2015). *Das Schweigen der Könner*. Münxter: Waxmann.

NICKOLAUS, R. (2016). Barrieren bei der Bewältigung berufsfachlicher Aufgaben. *Zeitschrift für Berufs- und Wirtschaftspädagogik, 112*(2),167–183.

NICKOLAUS, R., GSCHWENDTNER, T. & ABELE, S. (2009). *Die Validität von Simulationsaufgaben am Beispiel der Diagnosekompetenz von Kfz-mechatronikern. Vorstudie zur Validität von Simulationsaufgaben im Rahmen eines VET-LSA*. Abschlussbericht für das Bundesministerium für Bildung und Forschung zum Projekt. Stuttgart: Universität Stuttgart, Institut für Erziehungswissenschaft und Psychologie. Abruf am 29.08.2016 unter https://www.bmbf.de/files/Abschluss-Bericht_Druckfassung.pdf

NICKOLAUS, R., ABELE, S., GSCHWENDTNER, T., NITSCHKE, A. & GREIFF, S. (2012). Fachspezifische Problemlösefähigkeit in gewerblich-technischen Ausbildungsberufen. Modellierung, erreichte Niveaus und relevante Einflussfaktoren. *Zeitschrift für Berufs- und Wirtschaftspädagogik, 108*(2), 243–272.

NICKOLAUS, R., ABELE, S. & SCHMIDT, T. (2014). Die Relevanz expliziten und impliziten Wissens für berufsfachliche Leistungen – Forschungsergebnisse in gewerblich-technischen Domänen und ihre Bedeutung für berufliche Qualifizierungsprozesse. In BMBF (Hrsg.), *Bildungsforschung 2020. Zwischen wissenschaftlicher Exzellenz und gesellschaftlicher Verantwortung. Bildungsforschung* (S. 333–346). Bd. 42. Berlin.

OECD (2013). *Assessment of Higher Education Learning Outcomes. Feasibility Study Re-port. Volume 3 – Further Insights.* Paris: OECD Publishing.

OPP, K. D. (1972). *Methodologie der Sozialwissenschaften.* Reinbek: Rowohlt.

PETSCH, D., NORWIG, K. & NICKOLAUS, R. (2015). Berufsfachliche Kompetenzen in der Grundstufe Bautechnik. Strukturen, erreichte Niveaus und relevante Einflussfaktoren. In A. RAUSCH, J. WARWAS, J. SEIFRIED & E. WUTTKE (Hrsg.), *Konzepte und Ergebnisse ausgewählter Forschungsfelder der beruflichen Bildung*. Festschrift für Detlef Sembill (S. 59–88). Schneider Verlag Hohengehren: Baltmannsweiler.

PINDYCK, R. & RUBINFELD, D. (2009): *Mikroökonomie* (7. Aufl.). Hallbergmoos: Pearson Studium.

PINDYCK, R. & RUBINFELD, D. (2014): *Makroökonomie* (6. Aufl.). Hallbergmoos: Pearson Studium.

PRIM, R. & TILMANN, H. (1997). *Grundlagen einer kritisch-rationalen Sozialwissenschaft.* 7. Aufl. Wiesbaden: Quelle & Meyer.

REISS, K. & VERMEER H. J. (2014). *Towards a General Theory of Translational Action: Skopos Theory Explained.* Manchester: St. Jerome.

SCHMIDT, S. (2017). *Veränderungsmessung des fachlichen Wissens von Studierenden – Eine Längsschnittanalyse des Wissenserwerbs in einem latenten Mehrebenenmodell* (Economics Education and Human Resource Management). Springer Gabler: Wiesbaden (im Druck).

SEEBER, S. (2008). Ansätze zur Modellierung beruflicher Fachkompetenz in kaufmännischen Ausbildungsberufen. *Zeitschrift für Berufs- und Wirtschaftspädagogik, 106*(1), 74–97.

SEIFRIED, J. & ZIEGLER, B. (2009). Domänebezogene Professionalität. In O. ZLATKIN-TROITSCHANSKAIA, K. BECK, D. SEMBILL, R. NICKOLAUS & R. MULDER (Hrsg.), *Lehrprofessionalität. Bedingungen, Genese, Wirkungen und ihre Messung* (S. 83–92). Weinheim: Beltz.

SOLANO-FLORES, G., BACKHOFF, E. & CONTRERAS-NIÑO, L. A. (2009). Theory of test translation error. *International Journal of Testing, 9*, 78–91.

SOPER, J. C. & WALSTAD, W. B. (1987). *Test of Economic Literacy* (2. Ed.). New York: Joint Council on Economic Education.

STEGMÜLLER, W. (1984). *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie, Bd. 2*. Berlin: Springer.

WALSTAD, W. B., REBECK, K. & BUTTERS, R. B. (2013a). *Test of economic literacy: Examiner`s manual* (4. Ed.). New York: Council for Economic Education.

WALSTAD, W. B., REBECK, K., BUTTERS, R. B. (2013b). The Test of Economic Literacy: Development and Results. *Journal of Economic Education, 44*(3), 298–309.

WEIRICH, S., HECHT, M., & BÖHME, K. (2014). Modeling Item Position Effects Using Generalized Linear Mixed Models. *Applied Psychological Measurement, 38*(7), 535–548.

WISE, S. L. & DEMARS, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1–17.

WITT, R. (2009). Pädagogische Professionalität und die Differenzierung der Domänen in der beruflichen Bildung. In O. ZLATKIN-TROITSCHANSKAIA, K. BECK, D. SEMBILL, R. NICKOLAUS & R. MULDER (Hrsg.), *Lehrprofessionalität. Bedingungen, Genese, Wirkungen und ihre Messung* (S. 93–103). Weinheim: Beltz.

YAMAOKA, M., WALSTAD, W. B., WATTS, M. W., ASANO, T. & ABE, S. (2010). *Comparative Studies on Economic Education in Asia-Pacific Region.* Tokyo: Shumpusha Publishing.

ZLATKIN-TROITSCHANSKAIA, O., FÖRSTER, M., BRÜCKNER, S. & HAPP, R. (2014). Insights from a German assessment of business and economics competence. In H. COATES (Ed.), *Higher Education Learning Outcomes Assessment – International Perspectives* (S. 175–197). Frankfurt/Main: Peter Lang.

ZLATKIN-TROITSCHANSKAIA, O., PANT, H. A., KUHN, C., TOEPPER, M., & LAUTENBACH, C. (2016a). Assessment Practices in Higher Education and Results of the German Research Program Modeling and Measuring Competencies in Higher Education (KoKoHs). *Journal Research & Practice in Assessment, 11*, 46–54.

ZLATKIN-TROITSCHANSKAIA, O., PANT, H. A., KUHN, C., TOEPPER, M. & LAUTENBACH, C. (2016b). *Messung akademisch vermittelter Kompetenzen von Studierenden und Hochschulabsolventen. Ein Überblick zum nationalen und internationalen Forschungsstand.* Wiesbaden: Springer.

JUN.-PROF. DR. MANUEL FÖRSTER

Johannes Gutenberg-Universität Mainz, FB03, Wirtschaftspädagogik, Jakob Welder-Weg 9, 55128 Mainz, T 06131 39 23234, E-Mail: manuel.foerster@uni-mainz.de

DR. SEBASTIAN BRÜCKNER

Johannes Gutenberg-Universität Mainz, FB03, Wirtschaftspädagogik, Jakob Welder-Weg 9, 55128 Mainz, T 06131 39 22096, E-Mail: brueckner@uni-mainz.de

DR. ROLAND HAPP

Johannes Gutenberg-Universität Mainz, FB03, Wirtschaftspädagogik, Jakob Welder-Weg 9,

55128 Mainz, T 06131 39 22092, E-Mail: roland.happ@uni-mainz.de
UNIV.-PROFESSOR DR. KLAUS BECK
Johannes Gutenberg-Universität Mainz, FB03, Wirtschaftspädagogik, Jakob Welder-Weg 9,
55128 Mainz, T 06131 39 22027, E-Mail: beck@uni-mainz.de
UNIV.-PROFESSORIN DR. OLGA ZLATKIN-TROITSCHANSKAIA
Johannes Gutenberg-Universität Mainz, FB03, Wirtschaftspädagogik, Jakob Welder-Weg 9,
55128 Mainz, T 06131 39 23020, E-Mail: troitschanskaia@uni-mainz.de