



**University of
Zurich**^{UZH}

**Zurich Open Repository and
Archive**

University of Zurich
Main Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2019

**Wirtschaftsbürgerliche Kompetenz Deutschschweizer Gymnasiastinnen und
Gymnasiasten: Kompetenzmodellierung, Testentwicklung und
evidenzbasierte Validierung**

Ackermann, Nicole

Posted at the Zurich Open Repository and Archive, University of Zurich

ZORA URL: <https://doi.org/10.5167/uzh-175377>

Dissertation

Published Version

Originally published at:

Ackermann, Nicole. Wirtschaftsbürgerliche Kompetenz Deutschschweizer Gymnasiastinnen und Gymnasiasten: Kompetenzmodellierung, Testentwicklung und evidenzbasierte Validierung. 2019, University of Zurich, Faculty of Arts.



**Universität
Zürich^{UZH}**

**Wirtschaftsbürgerliche Kompetenz
Deutschschweizer Gymnasiastinnen und Gymnasiasten:
Kompetenzmodellierung, Testentwicklung und
evidenzbasierte Validierung**

Abhandlung
zur Erlangung der Doktorwürde
der Philosophischen Fakultät
der
Universität Zürich

vorgelegt von
Nicole Ackermann

Angenommen im Frühjahrssemester 2019
auf Antrag der Promotionskommission bestehend aus
Prof. em. Dr. Franz Eberle (hauptverantwortlicher Betreuer)
Prof. Dr. Urs Moser

Zürich, 2019

Danksagung

Diese Dissertationsschrift repräsentiert mein ambitioniertes Dissertationsvorhaben, das ich in den letzten drei Jahren verfolgte. Mit diesem konnte – nein, musste – ich mein Repertoire an sozialwissenschaftlichen Forschungsmethoden erweitern und vertiefen, was für einen interdisziplinären Forschungszugang unabdingbar ist. Es hat mir zudem eine differenzierte Perspektive auf die fachdidaktische Forschung, Lehre und Praxis eröffnet, die meine zukünftige berufliche Tätigkeit leiten wird.

Ein grosser Dank gilt meinem „Doktorvater“ Prof. em. Dr. Franz Eberle, der mir durch die Anstellung an seinem Lehrstuhl den wissenschaftlichen Zugang zur Gymnasial- und Wirtschaftspädagogik ermöglichte, mich zum Doktorat ermunterte und mir mit seinen kritischen Fragen wertvolle Hinweise für die inhaltliche und methodische Ausrichtung des Dissertationsvorhabens gab.

Des Weiteren bedanke ich mich aufrichtig bei allen Personen, die wesentlich zur gelingenden Umsetzung des Dissertationsvorhabens beigetragen haben: die Schulleiter*innen, Lehrer*innen und Schüler*innen der Kantonsschulen im Kanton St.Gallen für ihre freiwillige Teilnahme an den Erhebungen; Dr. Christin Siegfried, Isabel Frese, Béatrice Jung, Roland Schmidt, Franz Brügger, Dr. Daniel Jossen, Urs Perler und Beda Riklin für die fachwissenschaftliche und fachdidaktische Beurteilung des Testinstruments; Nihat Yasartürk, Larissa Zollinger und Patrick Peng für die engagierte Mitarbeit bei der Datenkodierung; Dr. Urs Grob für die forschungsmethodische Beratung; Katharina Szybalski, Aarusza Ramachandran und Sanja Stankovic für die unterstützende Zuarbeit bei der Literaturrecherche; Jsabel Menzi und Martin Ghéczy für das sorgfältige Lektorat.

Schliesslich geht mein herzlicher Dank an alle lieben Menschen in meinem persönlichen Umfeld, die mich während der drei Jahre, und besonders auch in den letzten intensiven Monaten, begleitet und unterstützt haben: meine Eltern Therese und Niklaus für freie Kost und Logis sowie die bedingungslose emotionale Zuwendung; Noah und Nina für die unbeschwerten Stunden zwischen LEGOburg und Naturmuseum; das „Chögli“ für Rat und Tat in allen Lebenslagen; die BerninaWG für die ausgelassenen Aktionen in Haus und Garten; meine „Berggspänli“ für die glückseligen Momente auf hohen Gipfeln und luftigen Graten; die Schwestern der Claustra Son Jon in Müstair und des Nouvo Monastero in Poschiavo für den spirituellen Rahmen in den Sommermonaten; Regula, Michael und Mimi für die unkomplizierte und fröhliche Wohngemeinschaft in Interlaken – und allen anderen, die mir wohlwollend und wertschätzend zur Seite standen.

Nicole Ackermann

Zürich, im Juni 2019

Kurzfassung

Die vorliegende Dissertation hat die «wirtschaftsbürgerliche» Kompetenz (WBK) bzw. deren *sozioökonomische Facette* (WBK-soek) zum Gegenstand. Es werden drei Ziele verfolgt: (A) ein *Rahmenmodell* für die WBK entwerfen, (B) einen *Leistungstest* für die Facette WBK-soek entwickeln und (C) die *evidenzbasierte Validierung* der WBK-soek bei Deutschschweizer Gymnasiast*innen.

WBK bezieht sich auf *ökonomisch geprägte Anforderungssituationen* in verschiedenen Lebensbereichen. Im gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereich ergeben sich komplexe *sozioökonomische Problemsituationen* mit kontroversen Lösungsansätzen aus verschiedenen Politikfeldern (z.B. Altersvorsorge, Energieversorgung). Der Leistungstest WBK-T2 operationalisiert das Konstrukt WBK-soek; er beinhaltet vier sozioökonomische Problemsituationen und insgesamt 32 Items. Die Items variieren systematisch im Kognitionsprozess und im Antwortformat. Der WBK-T2 richtet sich an Schüler*innen am Ende der Sekundarstufe II und wurde bei einer Stichprobe von 375 Gymnasiast*innen in einem Deutschschweizer Kanton eingesetzt. Die *Validierung des Testinhalts* erfolgte durch semi-strukturierte Interviews mit Experten, um die Repräsentativität des Konstrukts sowie die sprachliche und fachliche Adäquanz der Items zu evaluieren. Für die *Validierung der internen Struktur* wurde basierend auf der Item-Response-Theorie ein Partial-Credit-Rasch-Modell skaliert; hernach Test-Dimensionalität, Test-Reliabilität, Item-Homogenität und Item-Qualität evaluiert. Für die *Validierung der Beziehung zu externen Merkmalen* wurden inferenzstatistische Analysen gerechnet und die Testwerte hinsichtlich des curricularen Vorwissens, des biologischen Geschlechts, des soziokulturellen Hintergrunds und der Schulform evaluiert.

Die zusammengetragene Evidenz deutet darauf hin, dass die Testwerte der Gymnasiast*innen reliabel und valide hinsichtlich der sozioökonomischen Facette der WBK interpretiert werden können. Um das WBK-Modell vollständig abzubilden und empirisch zu überprüfen, müssten der persönlich-finanzielle und beruflich-unternehmerische Lebensbereich der WBK ebenfalls operationalisiert werden.

Schlüsselwörter: ökonomische Bildung, ökonomische Kompetenz, wirtschaftsbürgerliche Kompetenz, sozioökonomische Problemsituationen; Kompetenzmodellierung, Testentwicklung, evidenzbasierte Validierung; Item-Response-Theorie (IRT), Partial-Credit-Rasch-Modell.

Abstract

This thesis focuses on economic-civic competence (ECC) and its socioeconomic facet (ECC-soec). The goal is threefold: (A) establish a *framework model* for ECC, (B) develop an *achievement test* for ECC-soec and (C) validate ECC-soec evidence-based regarding Matura school students in German-speaking Switzerland.

ECC refers to *economic situations* in diverse life spheres. In the societal/economic life sphere, complex *socio-economic issues* with ambiguous and controversial solutions stem from various policy fields (e.g., retirement provision, energy supply). The achievement test WBK-T2 operationalizes the construct ECC-soec; it includes four socio-economic issues and 32 items in total. The items systematically vary in cognitive process and answer format. The WBK-T2 addresses students at the end of secondary school and was applied to a sample of 375 Swiss Matura school students in a German-speaking canton. For *content validation*, I conducted semi-structured interviews with experts to evaluate the items regarding construct representativeness as well as linguistic and technical appropriateness. For *internal structure validation*, I estimated a partial credit Rasch model based on item response theory to evaluate test dimensionality, test reliability, item homogeneity and item quality. For *external criteria validation*, I computed inferential statistical analyses to evaluate the relationship between test scores and curricular prior knowledge, sex, socio-cultural characteristics and types of school.

Evidence from all these sources suggests that students' test scores may reliably and validly be interpreted as ECC-soec. However, to fully represent and evaluate the ECC framework model, the personal/financial and the professional/entrepreneurial life spheres should be operationalized as well.

Key words: economic education, economic literacy, economic-civic competence, socio-economic issues; competence modelling, test development, evidence-based validation; item response theory (IRT), partial credit Rasch model.

Zusammenfassung

Die vorliegende Dissertation hat die «wirtschaftsbürgerliche» Kompetenz (WBK) bzw. die sozioökonomische Facette der wirtschaftsbürgerlichen Kompetenz (WBK-soek) zum Gegenstand. Mit dieser Dissertation werden drei Ziele verfolgt: (A) die theoriegeleitete Modellierung der WBK (WBK-Modell), (B) die modellbasierte und *Standards*-orientierte Testentwicklung für die WBK-soek (WBK-Test) und (C) die evidenzbasierte Validierung der WBK-soek bei Deutschschweizer Gymnasiast*innen.

Wirtschaftsbürgerliche Kompetenz bezieht sich auf *ökonomisch geprägte Anforderungssituationen* im persönlich-finanziellen, beruflich-unternehmerischen und gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereich, für deren Bewältigung kognitive Prozesse, nämlich Informationsverarbeitung und Problemlösen, erforderlich sind. Diese Definition der wirtschaftsbürgerlichen Kompetenz orientiert sich am Lebenssituationsprinzip der ökonomischen Bildung und an einem kognitiven Kompetenzkonzept. Das neue *Rahmenmodell der wirtschaftsbürgerlichen Kompetenz* (WBK-Modell) soll für gymnasiale und berufliche Schulen gelten. Es umfasst eine *Strukturebene* mit ökonomisch geprägten Anforderungssituationen in drei Lebensbereichen sowie eine *Prozessebene* mit kognitiven Prozessen bezüglich dieser Anforderungssituationen. Im gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereich – der sozioökonomischen Facette der WBK – ergeben sich komplexe *sozioökonomische Problemsituationen* (z.B. Finanzierung der Altersvorsorge, Sicherstellung der Energieversorgung) mit kontroversen Lösungsansätzen aus verschiedenen Politikfeldern. Die Auseinandersetzung mit sozioökonomischen Problemsituationen ist Bestandteil der *gesellschaftlichen Verantwortung*, auf die Jugendliche in einer modernen Gesellschaft und in einem demokratischen Rechtsstaat wie der Schweiz vorbereitet werden sollen. Dieses Bildungsziel findet sich für alle Schulformen auf der Sekundarstufe II in der Schweiz.

Der *revidierte Test zur wirtschaftsbürgerlichen Kompetenz* (WBK-T2) ist ein psychologischer Leistungstest und operationalisiert das Konstrukt *sozioökonomische Facette der WBK* (WBK-soek). Er beinhaltet vier sozioökonomische Problemsituationen (Altersvorsorge, Energieversorgung, Staatsverschuldung und Managervergütungen) und insgesamt 32 Items. Jede Problemsituation beginnt mit einem Einleitungstext, gefolgt von ca. acht Items. Für die Inhaltsspezifikationen des WBK-T2 wurden diverse reale sozioökonomische Problemsituationen hinsichtlich der Kriterien Repräsentativität (politische Relevanz), Wissenschaftlichkeit, Komplexität/Kontroversität/Multiperspektivität, Bekanntheit und Adaptierbarkeit geprüft und schliesslich vier ausgewählt. Für die Formatspezifikationen wurden in den Einleitungstexten Informationen vorstrukturiert sowie sechs Itemtypen definiert, die im Antwortformat und im Kognitionsprozess variieren.

Die Testwerte sollen kriteriumsorientiert als sozioökonomische Facette der WBK (WBK-soek) interpretiert werden und diagnostischen Zwecken dienen. Der WBK-T2 ist für Schülerinnen und Schüler am Ende der Sekundarstufe II konzipiert. Als Zielgruppe fungieren Schüler*innen an gymnasialen Maturitätsschulen und Berufsmaturitätsschulen im Schwerpunkt „Wirtschaft & Recht“ (WuR) sowie an kaufmännischen Berufsfachschulen; als Kontrastgruppe jene an Maturitätsschulen in anderen Schwerpunktfächern sowie an nicht-kaufmännischen Berufsfachschulen. Für das Konstrukt WBK-soek bzw. für den WBK-T2 werden eine Reihe von Annahmen formuliert, die es im Rahmen des Validierungsprozesses zu überprüfen gilt.

Die Validierung erfolgte entlang drei ausgewählter Evidenzquellen: Testinhalt, interne Struktur und Beziehung zu externen Merkmalen. Für die Validierung des *Testinhalts* des WBK-T2 wurden Leitfaden-Interviews mit Expert*innen durchgeführt, um die ursprünglichen Items des WBK-T1 hinsichtlich der domänenspezifischen Repräsentativität sowie der sprachlichen und fachlichen Adäquanz zu prüfen. Die Expert*innen beurteilten die Items mehrheitlich als domänenspezifisch repräsentativ, fachlich adäquat, sprachlich adäquat. Diese Expertenbeurteilung floss in die Revision des WBK-T1 bzw. in die Entwicklung des WBK-T2 ein, bei der ca. 60 % der Items modifiziert bzw. neu konstruiert wurden.

Der WBK-T2 sowie ein Fragebogen zu Individual- und Kontextmerkmalen wurde bei einer Stichprobe von 375 Gymnasiast*innen (42 % WuR, 58 % Nicht-WuR; 62 % weiblich, 38 % männlich) in einem Deutschschweizer Kanton eingesetzt. Die Kodierung der freien Schüler*innen-Antworten wurde von zwei unabhängigen Personen anhand eines detaillierten Kodierungsmanuals vorgenommen; die Inter-Rater-Reliabilität war substantiell.

Für die Validierung der *internen Struktur* wurde basierend auf der Item-Response-Theorie (IRT) ein Partial-Credit-Rasch-Modell mit der *marginal maximum likelihood*-Methode (MML) geschätzt. Die Prüfung der faktoriellen Struktur des Messmodells spricht für Eindimensionalität des Konstrukts WBK-soek bzw. des WBK-T2. Die Skalierbarkeit der Personen- und Itemparameter ist gegeben: Die Itemparameter streuen breit auf der Logit-Skala und es liegen mehr Items im mittleren Logit-Bereich, was eine zuverlässige Schätzung der Personenparameter induziert. Die Prüfung des Testmodells bestätigt tolerierbare Item-Infits für alle Items, ansteigende Schwellenparameter für fast alle polytomen Items sowie mehrheitlich akzeptables *differential item functioning* (DIF) bezüglich des curricularen Vorwissens (WuR/Nicht-WuR) und des biologischen Geschlechts. Die Reliabilität der Personenparameter und der Personen-Testwerte ist als gut zu beurteilen, die Qualität der Items ist aufgrund klassischer Item-Schwierigkeiten und Item-Trennschärfen bei fast allen Items ebenfalls gut. Es konnte anhand von zwei schwierigkeitsbestimmenden Merkmalen („inhaltliche Komplexität“ und „Informationsvorlage“) fünf Kompetenzniveaus bestimmt und beschrieben werden, die eine kriteriumsorientierte Interpretation der Personenfähigkeiten erlauben.

Für die Validierung der *Beziehung zu externen Merkmalen* wurden Mittelwert- und Korrelationsanalysen gerechnet. Zur *kriterialen Validität* lässt sich folgendes feststellen: Schüler*innen im Schwerpunktfach WuR zeigen eine höhere WBK-soek als jene in einem anderen Schwerpunktfach, Schüler*innen mit Erstsprache Deutsch eine höhere als jene mit Zweitsprache Deutsch, Schüler*innen mit Schweizer Staatsbürgerschaft eine höhere als jene mit anderer Staatsbürgerschaft. Schüler*innen an gymnasialen Maturitätsschulen haben eine höhere WBK-soek als jene an kaufmännischen Berufsfachschulen. Zur *divergenten Validität* kann festgehalten werden: Die WBK-soek korreliert schwach positiv mit den Schulfachnoten WuR und Deutsch sowie schwach positiv mit dem Interesse an sozioökonomischen Problemstellungen.

Die zusammengetragene Evidenz deutet darauf hin, dass die Testwerte der Gymnasiast*innen hinsichtlich des Konstrukts WBK-soek valide interpretiert werden können und dass der WBK-T2 zur reliablen Messung des Konstrukts WBK-soek eingesetzt werden kann. Jedoch wird zur weiteren Optimierung des WBK-T2 eine kleine Revision empfohlen. Dabei könnten, im Sinne einer weiteren Evidenzquelle, die Denkprozesse der Testpersonen beim Lösen der Items analysiert werden.

Unbesehen davon eröffnen sich ausgehend von dieser Dissertation vielseitige Forschungsfelder. Die Kompetenzmessung, d.h. die Erfassung der WBK-soek mit dem WBK-T2 könnte zum einen auf die gesamte Sekundarstufe II, insbesondere Berufsmaturitäts- und Berufsfachschulen, ausgedehnt werden, zum anderen im ganzen deutschsprachigen Raum, insbesondere Deutschland, bei vergleichbaren Stichproben vorgenommen werden. Für die fachdidaktische Unterrichtsforschung böte sich der WBK-T2 als Instrument für das *Pre-/Post-Testing* bei problemorientierten Interventionen an.

Um das WBK-Modell vollständig abzubilden und empirisch zu überprüfen, müsste die persönlich-finanzielle und die beruflich-unternehmerische Facette der WBK – in separaten Leistungstests – ebenfalls operationalisiert werden. Zudem müsste in der Schweizer Fachdidaktik eine theoretisch fundierte und differenzierte Konzeption für die Leitidee des «mündigen Wirtschaftsbürgers» und die Zielsetzung der «Bildung des allgemeinen Wirtschafts- und Gesellschaftsverständnisses» für die Sekundarstufe II entwickelt werden, als Ausgangspunkt für die Entwicklung/Revision von Lehrplänen sowie für die Formulierung kompetenzorientierter Bildungsstandards.

INHALTSÜBERSICHT

DANKSAGUNG.....	III
KURZFASSUNG	V
ABSTRACT	VI
ZUSAMMENFASSUNG	VII
INHALTSÜBERSICHT.....	XI
INHALTSVERZEICHNIS	XIII
ABBILDUNGSVERZEICHNIS	XIX
TABELLENVERZEICHNIS	XXI
ABKÜRZUNGSVERZEICHNIS.....	XXIII
1 Einleitung.....	1
TEIL A: KOMPETENZMODELLIERUNG	9
2 Konzeptionen der ökonomischen Bildung.....	9
3 Modellierung und Messung der ökonomischen Kompetenz	37
4 Rahmenmodell für wirtschaftsbürgerliche Kompetenz	55
TEIL B: TESTENTWICKLUNG.....	85
5 Psychologische Tests und Testtheorie	85
6 Standards für pädagogisches-psychologisches Testen	123
7 Entwicklungsgeschichte des WBK-Tests.....	135
8 Spezifikationen des WBK-T2.....	143
TEIL C: EVIDENZBASIERTE VALIDIERUNG	179
9 Validierung des Testinhalts des WBK-T2	183
10 Durchführung und Auswertung des WBK-T2	189
11 Validierung der Internen Struktur des WBK-T2.....	203
12 Validierung der Beziehung des WBK-T2 zu externen Merkmalen	265
TEIL D: SCHLUSSBETRACHTUNG	283
13 Zentrale Ergebnisse und Erkenntnisse	283
14 Limitationen und Perspektiven	293
LITERATURVERZEICHNIS	299
RECHTSTEXTE.....	317
ANHANG.....	319

INHALTSVERZEICHNIS

DANKSAGUNG.....	III
KURZFASSUNG	V
ABSTRACT	VI
ZUSAMMENFASSUNG	VII
INHALTSÜBERSICHT.....	XI
INHALTSVERZEICHNIS.....	XIII
ABBILDUNGSVERZEICHNIS	XIX
TABELLENVERZEICHNIS	XXI
ABKÜRZUNGSVERZEICHNIS.....	XXIII
1 Einleitung.....	1
1.1 Thematische Relevanz	1
1.2 Hintergrund, Forschungsziel und Leitfragen.....	5
1.3 Gliederung der Arbeit.....	6
TEIL A: KOMPETENZMODELLIERUNG	9
2 Konzeptionen der ökonomischen Bildung.....	9
2.1 Ökonomische Bildung (die deutsche Fachdidaktik)	9
2.1.1 Leitbilder und Bildungsziele der ökonomischen Bildung	9
2.1.2 Curriculare Prinzipien für Lernziele und Lerninhalte der ökonomischen Bildung	11
2.1.3 Bildungsstandards und Kompetenzen für die ökonomischen Bildung.....	13
2.1.3.a Der Vorschlag der DEGÖB (2004, 2006, 2009).....	15
2.1.3.b Der Vorschlag des IÖB (Kaminski et al., 2008, 2017).....	16
2.1.3.c Der Vorschlag von Retzmann et al. (2010) und Seeber et al. (2012).....	19
2.1.3.d Der Vorschlag von Weber (2005, 2013, 2014)	21
2.2 Wirtschaftsbürgerliche Bildung (die Schweizer Fachdidaktik).....	23
2.2.1 Leitbild und Bildungsziel der wirtschaftsbürgerlichen Bildung	23
2.2.2 Gestaltung von Lehrplan und Unterricht.....	25
2.2.3 Bildungsstandards für das Fach „Wirtschaft & Recht“ (Dubs, 2004, 2007)	27
2.3 Ökonomische Bildung im Schweizer Bildungssystem	28
2.3.1 Sekundarstufe II: Berufsbildung und Allgemeinbildung.....	29
2.3.1.a Gymnasiale Bildung.....	29
2.3.1.b Berufliche Bildung	30
2.3.2 Sekundarstufe I: Grundbildung.....	33
2.4 Zwischenfazit	34

3	Modellierung und Messung der ökonomischen Kompetenz	37
3.1	Kompetenzkonzepte	37
3.2	Kompetenzdefinitionen in der Bildungsforschung	42
3.3	Kompetenzmodelle für die pädagogisch-psychologische Diagnostik.....	44
3.4	Konzepte und Instrumente für ökonomische Kompetenz	47
3.4.1	Wirtschaftskundliche Kompetenz und WBT (Beck, 1993)	48
3.4.2	Ökonomische Kompetenzen und OEKOMA-Test (Schumann & Eberle, 2014)	49
3.4.3	Wirtschaftskompetenz im kaufmännischen Bereich (Eberle, 2015)	51
3.5	Zwischenfazit	53
4	Rahmenmodell für wirtschaftsbürgerliche Kompetenz	55
4.1	Bisherige Definition der WBK: eine kritische Reflexion	55
4.1.1	Kompetenzbezeichnung.....	56
4.1.2	Domänenbeschreibung	57
4.2	Modifizierte Definition der WBK und neues WBK-Modell.....	59
4.3	Strukturebene des WBK-Modells: Lebensbereiche.....	62
4.3.1	Die drei Lebensbereiche	62
4.3.2	Der gesamtgesellschaftliche/gesamtwirtschaftliche Lebensbereich	67
4.3.2.a	<i>Politikbegriff und Politikzyklus.....</i>	<i>68</i>
4.3.2.b	<i>Gesellschaftliche Verantwortung und politische Teilhabe</i>	<i>71</i>
4.3.2.c	<i>Politikbereiche und sozioökonomische Problemsituationen: Eine Domänenanalyse für die Schweiz, 1990 bis 2007.....</i>	<i>73</i>
4.4	Prozessebene des WBK-Modells: Kognitionsprozesse	76
4.4.1	Informationsverarbeitung bei sozioökonomische Problemsituationen	76
4.4.2	Problemlösen bei sozioökonomischen Problemsituationen.....	80
4.5	Zwischenfazit	83
	TEIL B: TESTENTWICKLUNG.....	85
5	Psychologische Tests und Testtheorie	85
5.1	Psychologische Tests	85
5.1.1	Definition und Merkmale psychologischer Tests	85
5.1.2	Gütekriterien psychologischer Tests.....	87
5.1.2.a	<i>Objektivität</i>	<i>88</i>
5.1.2.b	<i>Reliabilität (Zuverlässigkeit)</i>	<i>89</i>
5.1.2.c	<i>Validität (Gültigkeit).....</i>	<i>90</i>
5.1.2.d	<i>Nebengütekriterien</i>	<i>93</i>
5.2	Testtheorie: Formale Modelle psychologischer Tests	95
5.2.1	Klassische Testtheorie (KTT)	97
5.2.1.a	<i>Annahmen der KTT.....</i>	<i>97</i>
5.2.1.b	<i>Klassische Item- und Faktorenanalyse.....</i>	<i>97</i>
5.2.2	Probabilistische Testtheorie bzw. Item-Response-Theorie (IRT).....	102

5.2.2.a	<i>Annahmen der IRT</i>	102
5.2.2.b	<i>Klassifizierung von IRT-Modellen</i>	103
5.2.3	Das Rasch-Modell (dichotomes 1PL-Latent-Trait-Modell).....	105
5.2.3.a	<i>Modellgleichung</i>	106
5.2.3.b	<i>Modellannahmen/Modelleigenschaften</i>	108
5.2.3.c	<i>Parameterschätzung</i>	111
5.2.3.d	<i>Modellgeltungstests</i>	115
5.2.3.e	<i>Itemfit-Masse</i>	117
5.2.4	Das Partial-Credit-Rasch-Modell (ordinales 1PL-Latent-Trait-Modell).....	119
6	Standards für pädagogisches-psychologisches Testen	123
6.1	Standards für evidenzbasierte Validierung	123
6.1.1	Validierungsprozess	123
6.1.2	Evidenzquellen für Validität	125
6.2	Standards für Testentwicklung.....	127
6.2.1	Testspezifikationen.....	127
6.2.2	Item-Konstruktion.....	128
6.2.2.a	<i>Itemstamm</i>	129
6.2.2.b	<i>Antwortformate</i>	130
6.2.2.c	<i>Item-Kodierung</i>	132
6.2.3	Testheft	133
6.3	Standards für Testanwendung.....	133
7	Entwicklungsgeschichte des WBK-Tests	135
7.1	CoBALIT und WBK-T1.....	135
7.1.1	Hintergrund des Projekts CoBALIT.....	135
7.1.2	Entwicklung und Adaption des WBK-T1	136
7.1.3	Psychometrische Qualität des WBK-T1.....	137
7.2	WBKgym und WBK-T2.....	139
7.2.1	Hintergrund des Projekts WBKgym.....	139
7.2.2	Revision des WBK-T1 und Entwicklung des WBK-T2.....	139
8	Spezifikationen des WBK-T2	143
8.1	Testzweck.....	143
8.1.1	Eingrenzung des Konstrukts.....	143
8.1.2	Testanwendung.....	144
8.1.3	Testpersonen: Zielgruppe und Zielsubgruppen.....	145
8.1.4	Annahmen zum Konstrukt und zur Interpretation der Testwerte	146
8.1.4.a	<i>Annahmen zum Testinhalt</i>	146
8.1.4.b	<i>Annahmen zur internen Struktur</i>	146
8.1.4.c	<i>Annahmen zur Beziehung mit externen Merkmalen</i>	148
8.2	Inhaltsspezifikationen.....	154

8.2.1	Inhaltsbereiche des WBK-T2: Problemsituationen.....	154
8.2.1.a	Kriterien	154
8.2.1.b	Auswahl.....	162
8.2.2	Inhaltselemente der sozioökonomischen Problemsituationen: Items.....	162
8.3	Formatspezifikationen.....	164
8.3.1	Testaufbau und Testlänge.....	164
8.3.2	Einleitungstexte	165
8.3.3	Itemtypen.....	166
8.4	Kodierungsspezifikationen	174
8.5	Testadministration	175
8.5.1	Testverfahren.....	175
8.5.2	Testdauer	176
8.6	Zwischenfazit	177
TEIL C: EVIDENZBASIERTE VALIDIERUNG		179
9	Validierung des Testinhalts des WBK-T2	183
9.1	Experten-Interviews	183
9.2	Kognitive Interviews (Cognitive Pretesting).....	186
9.3	Zwischenfazit	187
10	Durchführung und Auswertung des WBK-T2	189
10.1	Datenerhebung.....	189
10.1.1	Piloterhebung (Standard-Pretesting).....	189
10.1.2	Haupterhebung/Erhebungsdesign.....	190
10.1.3	Stichprobe.....	191
10.1.3.a	Ziehung der Stichprobe	192
10.1.3.b	Beschreibung der Stichprobe.....	195
10.2	Datenauswertung.....	197
10.2.1	Kodierungsverfahren	197
10.2.2	Datenerfassung und Datenbereinigung.....	199
10.2.2.a	Fehlende Werte	199
10.2.2.b	Neue Variablen.....	200
10.2.3	Analyseverfahren und Statistik-Programme	200
11	Validierung der Internen Struktur des WBK-T2.....	203
11.1	Dimensionalität des Messmodells	204
11.1.1	Modellspezifikationen.....	204
11.1.2	Modellfit-Indizes.....	207
11.2	Schätzung und Skalierung der Parameter.....	212
11.2.1	Person-Item-Map	212
11.2.2	Personenparameter und Personen-Testwerte	214
11.2.3	Item-Schwierigkeitsparameter nach Problemsituation.....	215

11.3	Messgenauigkeit des Tests.....	218
11.3.1	Test Information Curve (TIC) und Standardfehler	218
11.3.2	Reliabilitäts-Koeffizienten	219
11.4	Gültigkeit des Testmodells: Item-Homogenität.....	222
11.4.1	Item-Infits	223
11.4.2	Schwellenparameter.....	226
11.4.3	Differential Item Functioning (DIF)	226
11.4.3.a	DIF bzgl. Schwerpunktfach: Gruppe WuR und Gruppe Nicht-WuR	227
11.4.3.b	DIF bzgl. Geschlecht: Jungen und Mädchen	230
11.5	Verteilung der Testwerte und Qualität der Items.....	235
11.5.1	Testwerte-Verteilung.....	235
11.5.1.a	Prüfung auf Normalverteilung.....	236
11.5.1.b	Testwerte nach Problemsituation	236
11.5.1.c	Kriteriumsorientierte Interpretation der Testwerte	238
11.5.2	Klassische Item-Schwierigkeiten.....	239
11.5.2.a	Item-Schwierigkeit nach Antwortformat	241
11.5.2.b	Item-Schwierigkeit nach Itemtypen.....	241
11.5.3	Klassische Item-Trennschärfen.....	242
11.5.3.a	Item-Trennschärfen (Item-Skala-Korrelationen).....	242
11.5.3.b	Kategorien-Trennschärfen (Kategorie-Skala-Korrelationen)	244
11.6	Graduierung des Tests	245
11.6.1	Qualitative Beurteilung der theoretischen Item-Schwierigkeiten anhand schwierigkeitsbestimmender Merkmale.....	246
11.6.1.a	Schwierigkeitsbestimmende Merkmale.....	246
11.6.1.b	Auswahl von Merkmalen und Ausprägungen	248
11.6.1.c	Itemrating und Inter-Rater-Reliabilität.....	249
11.6.2	Quantitative Bestimmung von Kompetenzschwellen	251
11.6.2.a	Bedeutung der einzelnen Merkmale (Korrelationsanalysen).....	252
11.6.2.b	Bedeutung der Merkmalsausprägungen (Varianzanalysen).....	253
11.6.2.c	Prädiktion der Itemparameter durch schwierigkeitsbestimmende Merkmale (Regressionsanalyse).....	255
11.6.2.d	Bestimmung der Schwellen auf der Logit-Skala.....	258
11.6.3	Qualitative Beschreibung der Kompetenzniveaus.....	259
11.7	Zwischenfazit	262
12	Validierung der Beziehung des WBK-T2 zu externen Merkmalen	265
12.1	Kriteriale Validität: WBK-soek von Zielgruppe und Kontrastgruppe	265
12.1.1	Qualitative Analysen zu den Lerninhalten in den Curricula.....	266
12.1.2	Quantitative Analysen zum curricularen Vorwissen (Fach WuR).....	269
12.2	Kriteriale Validität: WBK-soek von Zielsubgruppen	271
12.2.1	Biologisches Geschlecht.....	271

12.2.2 Soziokultureller Hintergrund (Erstsprache, Staatsbürgerschaft).....	273
12.2.3 Schulform/Bildungsgang.....	275
12.2.4 Politiksystem/Land.....	277
12.2.4.a Gymnasiast*innen in Schweiz und Deutschland.....	277
12.2.4.b Kaufmännische Lernende in Schweiz und Deutschland.....	278
12.3 Divergente Validität: WBK-soek, Leistungs- und Interessensmerkmale.....	280
12.4 Zwischenfazit	281
TEIL D: SCHLUSSBETRACHTUNG	283
13 Zentrale Ergebnisse und Erkenntnisse	283
13.1 Ergebnisse entlang der Leitfragen.....	283
13.1.1 Zur Kompetenzmodellierung: WBK-Modell.....	283
13.1.2 Zur Testentwicklung: WBK-T2	285
13.1.3 Zur evidenzbasierten Validierung.....	287
13.2 Empfehlungen für Testrevision des WBK-T2	290
14 Limitationen und Perspektiven	293
14.1 Limitationen	293
14.2 Perspektiven.....	296
LITERATURVERZEICHNIS	299
RECHTSTEXTE.....	317
ANHANG.....	319
A1 Testentwicklung des WBK-T2	320
A2 Testauswertung des WBK-T2	344
A3 Testadministration von WBKgym.....	358
A4 Fragebogen der Individual- und Kontextmerkmale.....	367

Abbildungsverzeichnis

Abbildung 1-1. Struktur der Dissertationsschrift und Argumentation des Forschungsvorhabens	7
Abbildung 2-1. Kompetenzmodell für ökonomische Bildung nach Seeber et al. (2012).....	20
Abbildung 2-2. Kompetenzmodell für sozioökonomische Bildung nach Weber (2005).....	22
Abbildung 2-3: Wirtschaftliche Bildung mit zweigeteilter Zielsetzung nach Dubs (2011, 2013).....	25
Abbildung 3-1. Kompetenz als Kontinuum	41
Abbildung 3-2. Kompetenzstrukturmodell für Diagnostik.....	46
Abbildung 3-3. Kompetenzniveaumodell für Diagnostik.....	46
Abbildung 3-4. Modell der umfassenden Wirtschaftskompetenz im kaufmännischen Bereich.....	52
Abbildung 4-1. WBK-Modell für ökonomisch geprägte Anforderungssituationen mit Lebensbereichen und Kognitionsprozessen	61
Abbildung 4-2. Eidgenössische Abstimmungsvorlagen nach Politikfeldern 1990-2017	75
Abbildung 5-1. Gegenstand der Testtheorie	96
Abbildung 5-2. Annahmen der KTT.....	98
Abbildung 5-3. Annahmen der IRT.....	103
Abbildung 5-4. Klassifikation der IRT-Modelle.....	104
Abbildung 5-5. Item Characteristic Curves für drei dichotome Items	107
Abbildung 5-6. Category Characteristic Curves für vier Kategorien eines ordinalen Items	120
Abbildung 6-1. Antwortformate für Leistungstest.....	130
Abbildung 8-1. Aufbau des WBK-T2.....	165
Abbildung 8-2. Beispiel für Einleitungstext (Problemsituation AHV)	167
Abbildung 8-3. Beispiel des unveränderten Itemtypus I mit Kodierungsschema (Item AHV_5).....	169
Abbildung 8-4. Beispiel des unveränderten Itemtypus II (Item ENE_8)	169
Abbildung 8-5. Beispiel des modifizierten Itemtypus III mit Kodierungsschema (Item AHV_11 und AHV_11m)	170
Abbildung 8-6. Beispiel des modifizierten Itemtypus IV (Item STA_8 und STA_8m)	172
Abbildung 8-7. Beispiel des neuen Itemtypus V (Item ENE_2 und ENE_2m).....	173
Abbildung 8-8. Beispiel des neuen Itemtypus VI (Item ENE_12n)	173
Abbildung 10-1. Kodierungsverfahren für die Constructed-Response-Items des WBK-T2	198
Abbildung 11-1. Messmodell M1 mit einem latenten Faktor und 32 manifesten Indikatoren	206
Abbildung 11-2. Messmodell M2 mit vier latenten Faktoren gemäss Inhaltsbereich	206
Abbildung 11-3. Messmodell M3 mit zwei latenten Faktoren gemäss Kognitionsprozess	207
Abbildung 11-4. Messmodell M4 mit zwei latenten Faktoren gemäss Antwortformat	207
Abbildung 11-5. Person-Item-Maps für Stichprobe WBKgym und WBK-T2.....	213
Abbildung 11-6. Test Characteristic Curve (TCC) für WBK-T2.....	216
Abbildung 11-7. Itemparameter nach Problemsituation.....	216
Abbildung 11-8. Test Information Curve (TIC) für Stichprobe WBKgym im WBK-T2	220
Abbildung 11-9. Personenparameter und Standardfehler für Stichprobe WBKgym im WBK-T2.....	220
Abbildung 11-10. Item-Infits (wMNSQ) und Signifikanz (t-Statistik)	224
Abbildung 11-11. Item Characteristic Curve (ICC) für drei ausgewählte Items des WBK-T2	225

Abbildung 11-12. Grafischer Modelltest für DIF bezüglich Schwerpunktfach.....	228
Abbildung 11-13. Numerischer Modelltest für DIF auf lokaler Ebene bezüglich Schwerpunktfach.....	230
Abbildung 11-14. Grafischer Modelltest für DIF bezüglich Geschlecht (Mädchen, Jungen)	231
Abbildung 11-15. Numerischer Modelltest für DIF auf lokaler Ebene bezüglich Geschlecht	233
Abbildung 11-16. Häufigkeitsverteilung der Testwerte für Stichprobe WBKgym.....	237
Abbildung 11-17. Klassische Item-Schwierigkeiten (Item-Mittelwerte).....	240
Abbildung 11-18. Klassische Item-Mittelwerte nach Itemtypen.....	242
Abbildung 11-19. Klassische Item-Trennschärfen (Item-Skala-Korrelation).....	243
Abbildung 11-20. Finales Itemrating: Item-Anzahl nach Merkmal und Ausprägung	250
Abbildung 11-21. Deskriptive Analyse der Itemparameter nach Merkmal und Ausprägung.....	254
Abbildung 11-22. Kompetenzniveaus im WBK-T2 (kriteriumsorientierte Interpretation).....	261
Abbildung 12-1. Adjustierte Testwerte nach Problemsituationen und Schwerpunktfach	270
Abbildung 12-2. Boxplots für Testwerte nach Schwerpunktfach und Geschlecht	272
Abbildung 12-3. Adjustierte Testwerte nach Problemsituationen und Schulform.....	276

Tabellenverzeichnis

Tabelle 2-1. Strukturierung der Kompetenzbereiche der ökonomischen Bildung	22
Tabelle 3-1. Testaufbau und Testinhalt des OEKOMA-Tests.....	50
Tabelle 4-1. Strukturebene des WBK-Modells: Inhaltsbereiche und domänenspezifische Anforderungssituationen	65
Tabelle 4-2. Aufgabenfelder von Ökonomie und Politik.....	69
Tabelle 4-3. Phasen des Politikzyklus am Beispiel der sozioökonomischen Problemsituation „Altersvorsorge“	70
Tabelle 4-4. Gesellschaftliche Verantwortung und politische Rechte	72
Tabelle 4-5. Exemplarische Kodierung von Abstimmungsvorlagen zu den Politikfeldern „Energie & Umwelt“ sowie „Soziale Sicherheit & Vorsorge“	75
Tabelle 4-6. Politikfelder und sozioökonomische Problemsituationen.....	76
Tabelle 4-7. Prozessebene des WBK-Modells: Informationsverarbeitung am Beispiel der sozioökonomischen Problemsituation „Altersvorsorge“	79
Tabelle 4-8. Prozessebene des WBK-Modells: Problemlösen für sozioökonomische Problemsituationen ..	82
Tabelle 6-1. Exemplarisches Kodierungsschema für Items in einem Leistungstest.....	132
Tabelle 7-1. Beispiele für Adaption des WBK-T1.....	138
Tabelle 7-2. Testinhalte und Testlänge des WBK-T1 und WBK-T2	141
Tabelle 7-3. Benennung der Problemsituationen im WBK-T1 und WBK-T2	141
Tabelle 8-1. Beurteilung sozioökonomischer Problemsituationen für Testentwicklung und -revision.....	158
Tabelle 8-2. Inhaltselemente der Problemsituationen AHV und ENE im WBK-T2	163
Tabelle 8-3. Konstruktionsregeln für Einleitungstexte.....	166
Tabelle 8-4. Konstruktionsschema für Items (Itemtypen).....	168
Tabelle 8-5. Item-Anzahl im WBK-T2 nach Problemsituation und Itemtypus.....	175
Tabelle 8-6. Item-Anzahl im WBK-T2 nach Antwortformat und Kognitionsprozess.....	175
Tabelle 8-7. Evidenzquellen, Ziele und Vorgehen für die Validierung der WBK-soek.....	180
Tabelle 9-1. Kriterien und Interview-Leitfragen für Validierung der WBK-T2 Items	184
Tabelle 9-2. Ergebnisse der Experten-Interviews zum Inhalt des WBK-T2	185
Tabelle 9-3. Ergebnisse der kognitiven Interviews zum Testheft des WBK-T2	187
Tabelle 10-1. Erhebungsdesign des Projekts WBKgym.....	191
Tabelle 10-2. Stundentafel für Fach WuR im Kanton St.Gallen	194
Tabelle 10-3. Grundgesamtheit und Stichprobe nach Klassen und Schüler*innen.....	194
Tabelle 10-4. Stichprobe nach Sozialisationsmerkmalen.....	196
Tabelle 11-1. Modellfit-Indizes für Modelle M1, M2, M3 und M4	209
Tabelle 11-2. Verteilungs- und Reliabilitäts-Indizes für die Dimensionen der Modelle M3 und M4	210
Tabelle 11-3. Reliabilitäts-Koeffizienten.....	220
Tabelle 11-4. Items und Antwortkategorien mit (geringem) Underfit	224
Tabelle 11-5. Items bzw. Antwortkategorien mit ungeordneten Schwellenparametern.....	226
Tabelle 11-6. Numerischer Modelltest auf globaler Ebene für DIF bezüglich Schwerpunktfach.....	228
Tabelle 11-7. Numerischer Modelltest auf lokaler Ebene für DIF bezüglich Schwerpunktfach	230

Tabelle 11-8. Numerischer Modelltest auf globaler Ebene für DIF bezüglich Geschlecht.....	231
Tabelle 11-9. Numerischer Modelltest auf lokaler Ebene für DIF bezüglich Geschlecht.....	233
Tabelle 11-10. Deskriptive Kennwerte der Testwerte.....	237
Tabelle 11-11. Testwerte und Lösungsquote nach Problemsituation.....	237
Tabelle 11-12. Kompetenzkategorien nach Lösungsquote im WBK-T2 (kriteriumsorientierte Interpretation).....	239
Tabelle 11-13. Items und Antwortkategorien mit zu geringer bzw. zu hoher klassischer Schwierigkeit ..	240
Tabelle 11-14. Items und Antwortkategorien mit geringer klassischer Trennschärfe (Typus I) ..	243
Tabelle 11-15. Items und Antwortkategorien mit geringer klassischer Trennschärfe (Typus V und VI) ..	243
Tabelle 11-16. Ratingschema für Beurteilung der theoretischen Schwierigkeit der Items im WBK-T2	250
Tabelle 11-17. Inter-Rater-Agreement und Inter-Rater-Reliabilität nach schwierigkeitsbestimmenden Merkmalen.....	250
Tabelle 11-18. Korrelation zwischen schwierigkeitsbestimmenden Merkmalen und probabilistischen Item-Schwierigkeiten (Itemparameter).....	253
Tabelle 11-19. Varianzanalyse für Itemparameter nach schwierigkeitsbestimmenden Merkmalen	254
Tabelle 11-20. Regressionsanalyse der Itemparameter auf Ausprägungen der schwierigkeitsbestimmenden Merkmale	257
Tabelle 11-21. Bedeutsame schwierigkeitsbestimmende Merkmale für die Items des WBK-T2	257
Tabelle 12-1. Beurteilung des potenziellen curricularen Vorwissens für sozioökonomische Problemsituationen der Zielgruppe und Kontrastgruppe auf der Sekundarstufe II.....	268
Tabelle 12-2. Mittelwertunterschiede in WBK-soek nach Schwerpunktfach.....	270
Tabelle 12-3. Mittelwertunterschiede in WBK-soek nach Geschlecht	272
Tabelle 12-4. Mittelwertunterschiede in WBK-soek nach Geschlecht innerhalb Schwerpunktfach	272
Tabelle 12-5. Mittelwertunterschiede in WBK-soek nach Erstsprache	274
Tabelle 12-6. Mittelwertunterschiede in WBK-soek nach Staatsbürgerschaft	274
Tabelle 12-7. Mittelwertunterschiede in WBK-soek nach Schulform.....	276
Tabelle 12-8. Mittelwertunterschiede in WBK-soek nach Land für Gymnasiast*innen	279
Tabelle 12-9. Mittelwertunterschiede in WBK-soek nach Land für kaufmännische Lernende	279
Tabelle 12-10. Korrelation zwischen WBK-soek und externen Merkmalen	281

Abkürzungsverzeichnis

1PL-Modell	Einparametrisches logistisches Modell, auch: Rasch-Modell
2PL-Modell	Zweiparametrisches logistisches Modell, auch: Birnbaum-Modell
3PL-Modell	Dreiparametrisches logistisches Modell
ACER	Australian Council for Educational Research (Australien)
AERA	American Educational Research Association
AHV	„Altersvorsorge“; Problemsituation im WBK-T2 und WBK-T1; zudem: Alters- und Hinterlassenenversicherung (Schweiz)
AIC	Akaike information criterion
APA	American Psychological Association
ASCOT	Technology-based Assessment of Skills and Competencies in Vocational Education and Training (dt.: Technologiebasierte Kompetenzmessung in der beruflichen Bildung)
BBG	Bundesgesetz über die Berufsbildung, kurz: Berufsbildungsgesetz (Schweiz)
BFS	Bundesamt für Statistik (Schweiz)
BIC	Bayes Information Criterion
BiVo	Verordnung des SBFI über die berufliche Grundbildung [Beruf] EFZ, kurz: Bildungsverordnung (Schweiz)
BMBF	Bundesministerium für Bildung und Forschung (Deutschland)
B-Profil	Basis-Profil der kaufmännischen Grundbildung (Schweiz)
BWL	Betriebswirtschaftslehre
CAP	Comparative Agendas Project
CC-Funktion	category characteristic function (dts: kategoriencharakteristische Funktion)
CCC	category characteristic curve (dts: kategoriencharakteristische Kurve)
CEE	Concil for Economic Education (USA)
CFA	confirmatory factor analysis (dt.: Konfirmatorische Faktorenanalyse)
CFI	comparative fit index
CML-Funktion	conditional maximum likelihood function (dt. bedingte Maximum-Likelihood-Funktion)
CoBALIT	Competencies in the Field of Business and Administration, Learning, Instruction, and Transition (dt.: Kompetenzen in der kaufmännischen Berufsausbildung)
CR-Test	Critical-Ratio-Test
D-EDK	Deutscheschweizerische Konferenz der kantonalen Erziehungsdirektoren (Schweiz)
DEGÖB	Deutsche Gesellschaft für ökonomische Bildung (Deutschland)
DESI	Deutsch Englisch Schülerleistungen International
DIF	differential item functioning (dt. Differentielle Item-Funktionen)
EAP	expected a posteriori estimates
EAP/PV	expected a posteriori/plausible value
EBA	Eidgenössisches Berufsattest (Schweiz)
EDK	Schweizerische Konferenz der kantonalen Erziehungsdirektoren (Schweiz)
EFA	exploratory factor analysis (dt.: Konfirmatorische Faktorenanalyse)
EFZ	Eidgenössisches Fähigkeitszeugnis (Schweiz)

ENE	„Energieversorgung“; Problemsituation im WBK-T2 und WBK-T1
E-Profil	Erweitertes Profil der kaufmännischen Grundbildung (Schweiz)
ETS	Educational Testing Service (USA)
EUR	„Eurokrise“; Problemsituation im WBK-T1
FRW	Finanz- und Rechnungswesen
gymCH	Gymnasiast*innen in der Schweiz; Zielsubgruppe für WBK-T2
gymDE	Gymnasiast*innen in Deutschland; Zielsubgruppe für WBK-T2
gymMS	Schüler*innen an gymnasialen Maturitätsschulen; Zielsubgruppe für WBK-T2
IC-Funktion	item characteristic function (dt.: itemcharakteristische Funktion)
ICC	item characteristic curve (dt.: itemcharakteristische Kurve)
IRT	Item-Response-Theorie (auch: probabilistische Testtheorie)
ISCO	International Standard Classification of Occupations
ISEI	International Socio-Economic Index
IV	Invalidenversicherung (Schweiz)
J+S	Jugend+Sport; Sportförderprogramm des Bundes (Schweiz)
JML-Funktion	joint maximum likelihood function (dt. gemeinsame Maximum-Likelihood-Funktion), auch: unconditional maximum likelihood function (dt. Unbedingte Maximum-Likelihood-Funktion)
JUG	„Jugendverschuldung“; Problemsituation im WBK-T1
kfmBS	Schüler*innen an kaufmännischen Berufsfachschulen; Zielsubgruppe für WBK-T2
kfmCH	Kaufmännische Lernende in der Schweiz; Zielsubgruppe für WBK-T2
kfmDE	Kaufmännische Lernende in Deutschland; Zielsubgruppe für WBK-T2
KMK	Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, kurz: Kultusministerkonferenz (Deutschland)
K-S-Test	Kolmogoroff-Smirnov-Test
KT	Klassische Testtheorie
LAN	„Agrarhandel“; Problemsituation im WBK-T1
LINCA	Learning and Instruction for Commercial Apprentices (dt.: Lehr-Lernprozesse im kaufmännischen Bereich)
LLTM	Linear-Logistisches-Testmodell
M	Mittelwert einer Variablen bzw. einer Verteilungsfunktion
MAR	Reglement über die Anerkennung von gymnasialen Maturitätsausweisen, kurz: Maturitäts-Anerkennungsreglement (Schweiz)
MAN	„Managervergütungen“; Problemsituation im WBK-T2 und WBK-T1
max	maximaler Wert einer Variablen bzw. Verteilungsfunktion
min	minimaler Wert einer Variablen bzw. Verteilungsfunktion
MLE	maximum likelihood estimate (dt. Maximum-Likelihood-Schätzer)
MRCML	multidimensional random coefficients multinomial logit model
MML-Funktion	marginal maximum likelihood function (dt. marginale Maximum-Likelihood-Funktion)
MNSQ	mean squares

NCME	National Council on Measurement in Education (USA)
NMG	Natur, Mensch, Gesellschaft; Fachbereich des Lehrplan 21 (Schweiz)
OdA	Organisation der Arbeitswelt (Schweiz)
OECD	The Organisation for Economic Co-operation and Development (dt.: Organisation für wirtschaftliche Zusammenarbeit und Entwicklung)
OEKOMA	Ökonomische Kompetenzen von Maturandinnen und Maturanden
PISA	Programme for International Student Assessment
PTT	Probabilistische Testtheorie (auch: Item-Response-Theorie)
PV	plausible values
RECHT/STAAT	Rechts- und Staatskunde
RLP-ABU	Rahmenlehrplan für den allgemeinbildenden Unterricht in der beruflichen Grundbildung (Schweiz)
RLP-BM	Rahmenlehrplan für die Berufsmaturität (Schweiz)
RLP-MS	Rahmenlehrplan für die Maturitätsschulen (Schweiz)
RMSEA	root mean square error of approximation
SBFI	Staatssekretariat für Bildung, Forschung und Innovation (Schweiz)
SD	Standardabweichung zum Mittelwert einer Variablen
SE	Standardfehler einer Verteilung (engl. standard error)
Sek I	Sekundarstufe I
Sek II	Sekundarstufe II
SPEDLOGSWISS	Verband schweizerischer Speditions- und Logistikunternehmen (Schweiz)
SKKAB	Schweizerische Konferenz der kaufmännischen Ausbildungs- und Prüfungsbranchen (Schweiz)
SOEP	Sozioökonomische Problemsituationen
STA	„Staatsverschuldung“; Problemsituation im WBK-T2 und WBK-T1
STEU	„Steuerwettbewerb“; Problemsituation im WBK-T1
SuS	Schülerinnen und Schüler
SWISSMEM	Verband schweizerischer Maschinen-, Elektro- und Metall-Industrie (Schweiz)
SWOT	strengths, weaknesses, opportunities, threats (dt.: Stärken, Schwächen, Chancen, Risiken)
S-W-Test	Shapiro-Wilk-Test
TCC	test characteristic curve (dt.: testcharakteristische Kurve)
TEL	Test of Economic Literacy
TFL	Test of Financial Literacy
TIC	test information curve (dt.: Testinformationskurve)
TIMSS	Third International Mathematics and Science Study
VAR	Varianz zum Mittelwert einer Variablen
VET	Vocational Education and Training
VWL	Volkswirtschaftslehre
WAH	Wirtschaft, Arbeit Haushalt; Kompetenzbereich im Lehrplan 21 (Schweiz)
WBK	Wirtschaftsbürgerliche Kompetenz

WBK-soek	Sozioökonomische Faette der wirtschaftsbürgerlichen Kompetenz
WBK-Q1	Fragebogen zu den Individual- und Kontextmerkmalen der wirtschaftsbürgerlichen Kompetenz, Version 1
WBK-T1	Test zur wirtschaftsbürgerlichen Kompetenz, originale Version 1
WBK-T2	Test zur wirtschaftsbürgerlichen Kompetenz, revidierte Version 2
WBT	Wirtschaftskundlicher Bildungs-Test
WLE	weighted likelihood estimate (dt.: gewichteter Maximum-Likelihood-Schätzer)
wMNSQ	weighted mean squares
W&G	Wirtschaft und Gesellschaft; Lernbereich an kaufmännischen Berufsfachschulen (Schweiz)
WuR	Wirtschaft und Recht; Unterrichtsfach an gymnasialen Maturitätsschulen und Berufsmaturitätsschulen (Schweiz)

1 Einleitung

1.1 Thematische Relevanz

Der Titel dieser Dissertation – «wirtschaftsbürgerliche Kompetenz» – ist etwas holprig, denn er benennt gleich zwei unklare bzw. unscharfe Begriffe, die einer Erklärung und Begründung bedürfen: «wirtschaftsbürgerlich» und «Kompetenz».

Der Begriff «Kompetenz» wird im Alltag vieldeutig verwendet,¹ in der Bildungsforschung ist damit „Fähigkeit“ im weiteren Sinne gemeint (Maag Merki, 2009, S. 492f.). Doch auch in der Bildungsforschung gibt es kein einheitliches und allgemein anerkanntes Kompetenzkonzept. Ausführlich diskutiert wurde in der Forschungsgemeinschaft z.B. das Verhältnis von Kompetenz und Performanz, die Binnenstruktur der Dispositionen (kognitiv, motivational, affektiv), die Bedeutung des Kontextes (allgemein, spezifisch). In dieser Dissertation wird Kompetenz als kontextspezifische kognitive Leistungsdisposition verstanden, die sich funktional auf gleiche oder zumindest ähnliche Anforderungssituationen bezieht (Weinert, 2001a, zit. nach Hartig & Klieme, 2006, S. 128f.). Im vorliegenden Falle der «wirtschaftsbürgerlichen» Kompetenz sind dies *ökonomisch geprägte Anforderungssituationen*.

Der Begriff «wirtschaftsbürgerlich» wurde im Schweizer Bildungskontext von den Wirtschaftspädagogen Rolf Dubs (vgl. Dubs, 1985, 2001, 2011, 2013) und Franz Eberle (vgl. Eberle, 2006b; Eberle, 2015) sowie dem Wirtschaftsethiker Peter Ulrich (vgl. Ulrich, 1993, 2001) geprägt. Mit «wirtschaftsbürgerlicher Bildung» ist die «Bildung des allgemeinen Wirtschaft- und Gesellschaftsverständnisses» (Dubs, 1985, S. 65; 2011, S. 194; 2013, S. 15) gemeint, also eine «wirtschaftliche Allgemeinbildung» (ebd.), die allen Menschen zukommen soll. Davon ist die wirtschaftsberufliche Bildung abzugrenzen, die auf eine berufsspezifische Qualifikation abzielt. Insofern kann «wirtschaftsbürgerliche Bildung» mit dem im deutschen Bildungskontext geläufigen Begriff der ökonomischen Bildung bzw. der ökonomischen Allgemeinbildung gleichgesetzt werden. Für die theoretische Verortung von «wirtschaftsbürgerlichen» Anforderungssituationen wird in dieser Dissertation entsprechend auf Konzeptionen und Standards ökonomischer Bildung zurückgegriffen.

Ökonomische Allgemeinbildung soll junge Menschen zur Bewältigung ökonomisch geprägter Lebenssituationen befähigen (vgl. Albers, 1988, 1995; Kaminski, 1996; Kaminski, Brettschneider, & Schnell, 2017; G. Seeber, Retzmann, Remmele, & Jongebloed, 2012; B. Weber, 2005, 2014). Solche Lebenssituationen stellen sich dem Menschen im persönlichen, beruflichen und gesellschaftlichen

¹ Der Duden nennt u.a. folgende Synonyme: Fähigkeit, Fertigkeit, Qualifikation, Sachverstand, Begabung, Befugnis, Zuständigkeit (Duden, 2019).

Bereich. Die Leitidee der ökonomischen Allgemeinbildung ist «der mündige Wirtschaftsbürger», der diese ökonomisch geprägten Lebenssituationen tüchtig (i.S.v. sachkundig), selbstbestimmt und verantwortungsvoll bewältigt (Albers, 1995, S. 2f.). Diese Leitidee bzw. dieses Bildungsziel der ökonomischen Bildung findet sich auch in vielen neueren fachdidaktischen Konzeptionen der ökonomischen Bildung wieder und ist weitgehend unumstritten.

Hingegen wird die curriculare Ausgestaltung, d.h. die Auswahl und Gewichtung der Lernziele/Lerninhalte für eine ökonomische Allgemeinbildung sowohl in der Bildungsforschung als auch in der Bildungspolitik kontrovers diskutiert (vgl. z.B. DEGÖB, 2009; Dubs, 2004, 2007; Eberle, 2007; G. Seeber et al., 2012; B. Weber, 2005, 2014). Es können drei curriculare Prinzipien für die ökonomische Bildung unterschieden werden (Reetz, 1984, 2003): das Wissenschaftsprinzip (Orientierung an Fachdisziplinen), das Situationsprinzip (Orientierung an Lebenssituationen) und das Persönlichkeitsprinzip (Orientierung an persönlicher Lebensbewältigung und gesellschaftlicher Teilhabe). Die meisten Kompetenzmodelle für *ökonomische Bildung* rekurren auf das Situationsprinzip und verwenden als Referenzsystem entweder den „Gegenstandsbereich Wirtschaft“ (Dubs, 2011, 2013; Kaminski et al., 2017; Kaminski, Eggert, & Burkard, 2008) oder die „ökonomische Perspektive“ (Hedtke, 2019; Retzmann, Seeber, Remmele, & Jongbloed, 2010; G. Seeber et al., 2012; B. Weber, 2005, 2013).

Im Schweizer Bildungssystem ist ökonomische Allgemeinbildung auf der Sekundarstufe II seit vielen Jahren curricular verankert, und zwar an gymnasialen und beruflichen Schulen – obschon bei ersteren politisch nicht ganz unumstritten (vgl. Eberle, 2006b; Eberle & Brüggemann, 2013, S. 115f.). An Gymnasien wird das Fach „Wirtschaft & Recht“ unterrichtet (EDK, 1994); die „Einführung in Wirtschaft & Recht“² muss obligatorisch besucht werden (Art. 9 Abs. 5^{bis} MAR), zudem kann „Wirtschaft & Recht“ als Schwerpunkt oder Ergänzung gewählt werden (Art. 9 Abs. 3 und Abs. 4 MAR). An kaufmännischen Berufsfachschulen besteht ein eigener Lernbereich „Wirtschaft & Gesellschaft“ (SKKAB, 2011a) und an nicht-kaufmännischen Berufsfachschulen wird der Aspekt Wirtschaft im Lernbereich „Allgemeinbildender Unterricht“ behandelt (SBFI, 2006); zudem kann für die Berufsmaturität das Fach „Wirtschaft & Recht“ als Schwerpunkt oder Ergänzung gewählt werden (SBFI, 2012).

Kompetenzmodellierung und -messung ist in der Bildungsforschung in den letzten Jahren *en vogue* geworden, nicht zuletzt auch durch die gross angelegten Schulleistungsstudien wie z.B. PISA (vgl. z.B. OECD, 2013, 2017a). Auch in der deutschsprachigen Berufs- und Wirtschaftspädagogik hat

² Das Fach „Einführung in Wirtschaft & Recht“ wurde mit der Maturitätsreform 1995 als obligatorisch für alle Schüler*innen eingeführt. Es bildete zusammen mit den Fächern „Geschichte“ und „Geografie“ das Maturitätsfach „Geistes- und Sozialwissenschaften“. Bei der Maturitätsreform 2007 jedoch wurden die drei Fächer separiert und dem Fach „Einführung in Wirtschaft & Recht“ den Status eines Maturitätsfachs entzogen (Eberle, 2006b; Eberle & Brüggemann, 2013, S. 115f.).

die Kompetenzforschung einen *Boom* erlebt, woraus umfangreiche Forschungsprogramme und zahlreiche Studien entstanden sind. In Anlehnung an die oben vorgenommene Differenzierung von ökonomischer Allgemeinbildung und Berufsbildung können drei Forschungsströmungen unterschieden werden. Die erste Strömung umfasst Forschung zu ökonomischen Kompetenzen, häufig angelehnt an die Konzepte der *economic literacy* und *financial literacy* aus dem angelsächsischen Raum (vgl. z.B. Aprea, Leumann, & Gerber, 2014; Aprea, Wuttke, Leumann, & Heumann, 2015; K. Beck, 1993; K. Beck & Krumm, 1998; Rudeloff, 2019; Schumann & Eberle, 2014a; Schumann, Kaufmann, Eberle, Jüttler, & Ackermann, 2017). Die zweite Strömung fokussiert auf berufliche und insbesondere kaufmännische Kompetenzen (vgl. z.B. K. Beck, Landenberger, & Oser, 2016; Guggemos & Schönlein, 2015; Holtsch & Eberle, 2018; Lehmann & Seeber, 2007; Michaelis, 2017; Nickolaus, Gschwendtner, & Geißel, 2008; Winther, 2010; Winther & Achtenhagen, 2008, 2009b). Die dritte Strömung schliesslich befasst sich mit pädagogisch-didaktischen, fachdidaktischen und fachwissenschaftlichen Kompetenzen von (kaufmännischen) Berufsschullehrer*innen (vgl. z.B. Findeisen, 2017; Holtsch, 2018; Kuhn, 2014; Kuhn et al., 2014).

Für diese Dissertation ist der Forschungsstand der erstgenannten Strömung massgeblich, also die Modellierung und Messung von allgemeinen *ökonomischen Kompetenzen*. Die meisten Studien in diesem Bereich fokussieren auf *economic literacy*. Sie orientieren sich bei der Kompetenzmodellierung und Testentwicklung an einer Fachlogik, d.h. die Domäne „Ökonomie“ wird über eine fachwissenschaftliche Struktur (VWL, BWL) beschrieben.³ Die verwendeten psychometrischen Leistungstests erfassen entweder nur volkswirtschaftliche Inhalte (z.B. Wirtschaftskundlicher Bildungstest, WBT) (vgl. K. Beck, 1989, 1993; K. Beck & Krumm, 1998) oder volks- und betriebswirtschaftliche Inhalte (z.B. OEKOMA-Test) (vgl. Schumann & Eberle, 2014a; Schumann et al., 2010). Neuere Studien in diesem Bereich adressieren explizit die wirtschaftsbürgerliche Kompetenz (vgl. Eberle, Schumann, Kaufmann, Jüttler, & Ackermann, 2016; Holtsch & Eberle, 2018; Schumann et al., 2017). Dabei ist die Kompetenzmodellierung und Testentwicklung nach einer Themenlogik ausgerichtet, d.h. die Domäne „Ökonomie“ wird über reale wirtschaftliche Problemstellungen beschrieben. Das in diesen Studien vorgeschlagene Modell der umfassenden Wirtschaftskompetenz gilt für den kaufmännischen Bereich und verbindet kaufmännische und wirtschaftsbürgerliche Kompetenz. Für die Erfassung der wirtschaftsbürgerlichen Kompetenz wurde ein neuer Leistungstest entwickelt, der wirtschaftsbürgerliche Kompetenztest (WBK-T1), der mehrheitlich wirtschaftspolitische, aber auch finanzielle Problemsituationen umfasst (Eberle et al., 2016; Schumann et al., 2017). Die Studien im Bereich der *financial literacy* orientieren

³ Ökonomie als Fachwissenschaft gliedert sich in die zwei Fachbereiche Volkswirtschaft (VWL) und Betriebswirtschaft (BWL). Innerhalb der Volkswirtschaft wird üblicherweise in Mikroökonomie, Makroökonomie und politische Ökonomie unterschieden.

sich ebenfalls an einer Themenlogik; sie definieren ihre Domäne über empirisch beobachtbare Lern- und Handlungsfelder des persönlich-finanziellen Lebensbereichs (vgl. Aprea & Wuttke, 2016; Davies, 2015; OECD, 2012, 2017b; Rudeloff, 2019; Schürkmann & Schuhen, 2013).

Aus den obigen Ausführungen lässt sich folgendes Forschungsdesiderat ableiten: Es fehlt an einem Rahmenmodell für «wirtschaftsbürgerliche» Kompetenz, das auf die gesamte Sekundarstufe II, d.h. auf die gymnasiale *und* auf die berufliche Bildung anwendbar ist. Dieses Rahmenmodell sollte theorie- und empiriegeleitet «wirtschaftsbürgerliche», oder eben ökonomisch geprägte Anforderungssituationen, bündeln und diese mit Anforderungsniveaus verknüpfen. Mit dem WBK-T1 besteht zwar ein Testinstrument für die Erfassung der «wirtschaftsbürgerlichen» Kompetenz. Dieser muss jedoch hinsichtlich des neuen Rahmenmodells validiert sowie aufgrund der bisherigen psychometrischen Ergebnisse evaluiert und ggf. revidiert werden. Schliesslich sollte das revidierte Testinstrument bei einer Zielgruppe von gymnasialen und/oder beruflichen Schüler*innen eingesetzt werden, um die Interpretation der Testergebnisse hinsichtlich der «wirtschaftsbürgerlichen» Kompetenz zu validieren. In der Modellierung, Operationalisierung und evidenzbasierten Validierung des Konstrukts «wirtschaftsbürgerliche» Kompetenz besteht also ein Forschungsdesiderat, das mit diesem Dissertationsvorhaben angegangen wird. Sofern diese „Herausforderung“ von der Autorin bewältigt werden kann, stünden schliesslich ein theoretisch fundiertes Rahmenmodell sowie ein optimiertes Testinstrument zur Verfügung, mithilfe dessen a) «wirtschaftsbürgerliche» Kompetenz von Gymnasiast*innen und Berufsschüler*innen grossflächig untersucht und b) die Wirkung traditioneller und komplexer Lehr-Lernarrangements⁴ im Wirtschaftsunterricht auf den «wirtschaftsbürgerlichen» Kompetenzzuwachs hin beleuchtet werden könnte.

⁴ Komplexe Lehr-Lern-Arrangements in den Fächern „Wirtschaft & Recht“ bzw. „Wirtschaft & Gesellschaft“ beziehen sich häufig auf das didaktische Konzept des problemorientierten Lernens (Eberle, 2006; A. Weber, 2005, 2007).

1.2 Hintergrund, Forschungsziel und Leitfragen

Die vorliegende Dissertation entstand im Rahmen des Forschungsprojekts «Wirtschaftsbürgerliche Kompetenz Deutschschweizer Gymnasiast*innen (WBK_{gym})» am Lehrstuhl für Gymnasial- und Wirtschaftspädagogik der Universität Zürich, das von der Autorin initiiert und durchgeführt wurde. Aus dem Forschungsprojekt WBK_{gym} sind zudem zwei Masterarbeiten entstanden (Kavadarli, 2019; Peng, 2018) und noch weitere Publikationen zu erwarten, die z.T. inhaltlich und methodisch an die Dissertation anknüpfen (Ackermann & Kavadarli, 2019, in progress; Ackermann & Peng, 2018, in progress; Ackermann & Siegfried, under review; Siegfried & Ackermann, in progress, under review).

Die Dissertation hat die „wirtschaftsbürgerliche Kompetenz“ (WBK) bzw. die sozioökonomische Facette der wirtschaftsbürgerlichen Kompetenz (WBK-soek) zum Gegenstand. Das Ziel der Dissertation ist dreifach: (A) die theorie- und empiriegeleitete Modellierung der wirtschaftsbürgerlichen Kompetenz (WBK-Modell); (B) die modellbasierte und *Standards*-orientierte Testentwicklung für die sozioökonomische Facette der WBK (WBK-Test); (C) die evidenzbasierte Validierung der Testergebnisse hinsichtlich ihrer Interpretation als sozioökonomische Facette der WBK.

Aus diesem dreifachen Forschungsziel ergeben sich neun Leitfragen, die in der vorliegenden Dissertation bearbeitet werden. Die Leitfragen (1) bis (4) beziehen sich auf die Kompetenzmodellierung.

- (1) *An welche fachdidaktischen Konzeptionen für ökonomische Bildung lassen sich «wirtschaftsbürgerliche» Anforderungssituationen theoretisch anknüpfen?*
- (2) *An welche allgemeinpädagogischen und wirtschaftspädagogischen Kompetenzkonzepte kann das Konstrukt «wirtschaftsbürgerliche» Kompetenz theoretisch angelehnt bzw. von welchen muss es abgegrenzt werden?*
- (3) *Wie lassen sich Inhaltsbereiche und Kognitionsprozesse für «wirtschaftsbürgerliche» Anforderungssituationen zu einem Rahmenmodell wirtschaftsbürgerlicher Kompetenz (WBK-Modell) zusammenführen?*
- (4) *Welche realen sozioökonomischen Problemsituationen ergeben sich im gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereich?*

Die Leitfragen (5) und (6) beziehen sich auf die Testentwicklung.

- (5) *Welche fachdidaktischen und psychometrischen Kriterien sollten bei der Testevaluation und -revision des WBK-T1 berücksichtigt werden, um die Qualität des WBK-T2 zu optimieren?*

- (6) *Welche Spezifikationen müssen bei der Testentwicklung des WBK-T2 vorgenommen werden, um die sozioökonomische Facette der WBK repräsentativ und adäquat abzubilden?*

Die Leitfragen (7) bis (9) beziehen sich auf die evidenzbasierte Validierung.

- (7) *Testinhalt: Inwiefern sind die Inhaltsbereiche und Inhaltselemente des WBK-T2 für die sozioökonomische Facette der WBK repräsentativ und adäquat?*
- (8) *Interne Struktur: Welche psychometrische Qualität lässt sich dem gesamten WBK-T2 (Dimensionalität, Skalierbarkeit, Reliabilität, Graduierbarkeit) und den einzelnen Items (Homogenität, Trennschärfe, Schwierigkeit) nach klassischen und probabilistischen Kriterien attestieren?*
- (9) *Beziehung zu externen Merkmalen: Wie unterscheiden sich verschiedene Personengruppen (Zielgruppe und Kontrastgruppe, Zielsubgruppen) in der sozioökonomischen Facette der WBK? Wie hängen externe Leistungs- und Interessensmerkmale der Testpersonen mit der sozioökonomischen Facette der WBK zusammen?*

1.3 Gliederung der Arbeit

Die Dissertationsschrift gliedert sich entlang ihrer dreifachen Zielsetzungen in drei Hauptteile (vgl. Abbildung 1-1). TEIL A widmet sich der Kompetenzmodellierung. In Kapitel 2 werden diverse fachdidaktische Konzeptionen und bestehende Standards der ökonomischen Bildung nachgezeichnet sowie die Verankerung und Ausgestaltung der ökonomischen Bildung in der Schweiz erläutert. In Kapitel 3 werden die vielfältigen Kompetenzkonzepte und -definitionen diskutiert sowie bestehende Konzepte, Modelle und Instrumente für ökonomische Kompetenz vorgestellt und verglichen. In Kapitel 4 schliesslich wird die Definition der wirtschaftsbürgerlichen Kompetenz präzisiert und darauf basierend ein neues Rahmenmodell für wirtschaftsbürgerliche Kompetenz (WBK-Modell) entworfen; die Inhaltsbereiche und die Kognitionsprozesse für wirtschaftsbürgerliche Anforderungssituationen werden theoretisch und empirisch hergeleitet und detailliert beschrieben.

TEIL B widmet sich der Testentwicklung. In Kapitel 5 werden Merkmale psychologischer Tests und Testtheorie, insbesondere IRT-Modelle erläutert. Kapitel 6 stellt die Standards fürs pädagogische-psychologische Testen vor, insbesondere jene zur Validierung, Testentwicklung und Testdurchführung. Kapitel 7 erzählt die Entwicklungsgeschichte des Tests zur wirtschaftsbürgerlichen Kompetenz (WBK-Test), und zwar von der originalen Version (WBK-T1) bis zur revidierten Version (WBK-T2). In Kapitel 8 werden die Spezifikationen des WBK-T2 bezüglich des Inhalts, des Formats und der Kodierung detailliert beschrieben sowie der Testzweck, insbesondere die Eingrenzung des Konstrukts auf „sozioökonomische Facette der WBK“ (WBK-soek), begründet.

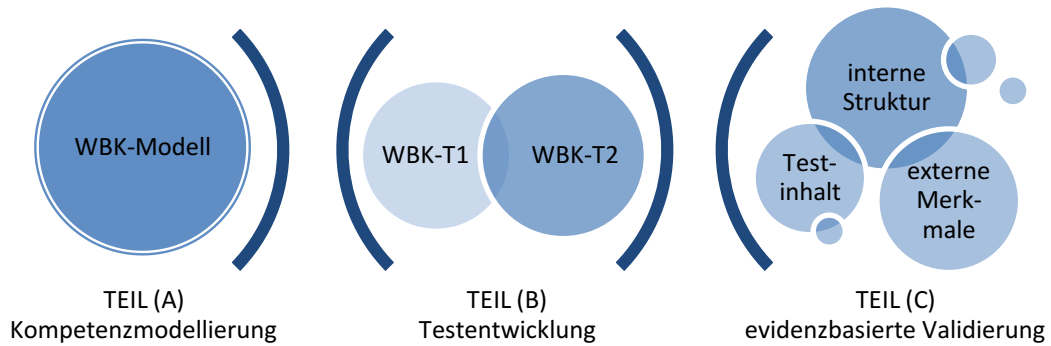


Abbildung 1-1. Struktur der Dissertationsschrift und Argumentation des Forschungsvorhabens

TEIL C widmet sich der evidenzbasierten Validierung. In Kapitel 9 wird der Testinhalt des WBK-T2 mittels Experten-Interviews und kognitiver Interviews validiert. Kapitel 10 beschreibt die Testdurchführung und -auswertung, insbesondere das Erhebungsdesign, die Stichprobe, das Kodierungsverfahren und die Analyseverfahren. In Kapitel 11 werden die umfangreichen Analysen zur Validierung der internen Struktur des WBK-T2 präsentiert und diskutiert. Kapitel 12 schliesslich beschreibt die Validierung der Beziehung des WBK-T2 zu externen Merkmalen, insbesondere zu verschiedenen Personengruppen.

TEIL D ist eine Schlussbetrachtung. In Kapitel 13 werden die zentralen Ergebnisse und Erkenntnisse entlang der Leitfragen zusammengefasst. Kapitel 14 diskutiert Limitationen der vorliegenden Dissertation und eröffnet Perspektiven für weitere Kompetenzforschung und für neue Unterrichtsforschung mit dem WBK-T2.

Der Anhang enthält Dokumente zur Testentwicklung, Testauswertung und Testadministration.

TEIL A: KOMPETENZMODELLIERUNG

2 Konzeptionen der ökonomischen Bildung

In diesem Kapitel wird der Leitfrage (1) nachgegangen (vgl. Kapitel 1.2):

(1) An welche fachdidaktischen Konzeptionen für ökonomische Bildung lassen sich «wirtschaftsbürgerliche» Anforderungssituationen theoretisch anknüpfen?

Fachdidaktische Konzeptionen ökonomischer Bildung sind idealtypische und normative Antworten auf die Frage einer „angemessenen“ ökonomischen Bildung. Sie legen zum einen Leitbilder und Bildungsziele fest und schlagen zum anderen Lernziele/Lerninhalte vor, in jüngster Zeit auch Kompetenzen (*performance standards*), die der Entwicklung von Lehrplänen und der Gestaltung von Lehr-Lern-Arrangements zugrunde gelegt werden. Da die Domäne „Ökonomie“ schwierig zu fassen ist, bestehen denn auch verschiedene und z.T. divergierende fachdidaktische Konzeptionen der ökonomischen Bildung. So stehen mitunter fachdidaktische Konzeptionen, die eher wirtschaftswissenschaftlich ausgerichtet sind, jenen gegenüber, die eher sozialwissenschaftlich ausgerichtet sind.

Dieses Kapitel skizziert als Erstes diverse Konzeptionen der ökonomischen Bildung aus der deutschen Fachdidaktik (Kapitel 2.1) und als Zweites, quasi als Pendant, die Konzeption der wirtschaftsbürgerlichen Bildung aus der Schweizer Fachdidaktik (Kapitel 2.2). Schliesslich wird die rechtliche Verankerung und die curriculare Ausgestaltung der ökonomischen Allgemeinbildung in der Schweiz beschrieben (Kapitel 2.3).

2.1 Ökonomische Bildung (die deutsche Fachdidaktik)

Dieses Kapitel behandelt die Konzeptionen der ökonomischen Bildung aus der deutschen Fachdidaktik. Es erläutert prägende Leitbilder und Bildungsziele (Kapitel 2.1.1), curriculare Prinzipien für die Ermittlung, Auswahl und Begründung von Lernzielen/Lerninhalten (Kapitel 2.1.2) sowie bisher vorgeschlagene Kompetenzmodelle (Kapitel 2.1.3).

2.1.1 Leitbilder und Bildungsziele der ökonomischen Bildung

In der deutschen Fachdidaktik finden sich Konzeptionen der ökonomischen Bildung bzw. der ökonomischen Allgemeinbildung seit den 1970er-Jahren (vgl. z.B. Albers, 1988, 1995; Kaminski, 1996; Kruber, 1995; May, 1978; Reetz, 1984). Diese Konzeptionen betonen allesamt den allgemeinbildenden Charakter der ökonomischen Bildung, unterscheiden sich jedoch in der Differenziertheit ihrer Leitbilder und in der Ausrichtung ihrer Bildungsziele. Nachfolgend werden die Konzeption von Hans-Jürgen Albers und Hans Kaminski skizziert, da sich ihre Leitbilder bzw. Zielsetzung der ökonomischen Bildung auch heute noch in vielen fachdidaktischen Rahmenmodellen

und Lehrplänen für ökonomische Bildung wiederfinden (vgl. G. Seeber et al., 2012; B. Weber, 2005, 2014) und weitgehend unumstritten sind.

Bei Albers (1988, 1995) besteht Bildung und entsprechend auch ökonomische Bildung aus den zwei Elementen „Persönlichkeit“ und „Lebenssituationen“, die er aus dem «Spannungsverhältnis zwischen Mensch und Welt» (Albers, 1995, S. 2) ableitet.⁵ Bildung wird, in Anlehnung an Robinsohn (1967), als «Ausstattung zum Verhalten in der Welt» (Robinsohn, 1967, S. 13, zit. nach Albers, 1995, S. 3) gedacht. Sie soll also zur Entwicklung einer Persönlichkeit beitragen, die fähig ist, mannigfaltige Lebenssituationen zu bewältigen.

«Ökonomische Bildung soll [...] zur Bewältigung der ökonomisch geprägten Lebenssituationen befähigen (Albers, 1995, S. 2).

Die zu bewältigenden Lebenssituationen sind also *ökonomisch geprägt* und stellen sich dem Menschen im privaten, beruflichen und gesellschaftlichen Bereich. Das *Leitbild* der ökonomischen Bildung ist «der mündige Wirtschaftsbürger» (Albers, 1995, S. 3), der diese Lebenssituationen nicht irgendwie, sondern *tüchtig, selbstbestimmt* und *verantwortungsvoll* bewältigt. Die Bewältigung der Lebenssituationen wird also unter dem Anspruch der Mündigkeit⁶ nach drei „Gütekriterien“ beurteilt (Albers, 1995, S. 3f.): *Tüchtigkeit* bezieht sich auf die Fähigkeit, die eigenen wechselnden Lebenssituationen sachgerecht und effizient anzugehen. *Selbstbestimmung* äussert sich in der Fähigkeit, das eigene Leben bewusst zu gestalten und die eigenen Interessen zu vertreten, Abhängigkeiten zu verringern und sich von übermässigen Zwängen zu befreien. *Verantwortlichkeit* ist die Bereitschaft, das eigene Handeln in einem grösseren gesellschaftlichen Zusammenhang zu sehen und die eigene gesellschaftliche Verantwortung anzuerkennen.

Zur Bewältigung von ökonomisch geprägten Lebenssituationen sind nach Albers (1995, S. 4f.) zwei unterschiedliche Fähigkeiten notwendig; er bezeichnet sie als Zielkategorien ökonomischer Bildung: *Urteilsfähigkeit* und *Handlungsfähigkeit*. In privaten und beruflichen Lebenssituationen dominiert die Handlungsfähigkeit, d.h. Sachverhalte analysieren und abwägen, Entscheidungen treffen und konkrete Handlungen vollziehen. Im gesellschaftlichen Lebensbereich dominiert die Urteilsfähigkeit, d.h. Sachverhalte analysieren und beurteilen, Position beziehen, ggf. als Kollektiv entscheiden und handeln.

⁵ In dieser Konzeption der ökonomischen Bildung lässt sich ein neuhumanistischer Bildungsbegriff erkennen, der wesentlich von Wilhelm von Humboldt (1767–1835) geprägt wurde. «Bildung als freie Wechselwirkung von Mensch und Welt, die sich angemessen nur als Wechselwirkung von Selbständigkeit und Empfänglichkeit umschreiben lässt. [...] Bildung vollzieht sich [...] in einer offenen Dialektik von Welterfahrung und Weltentwurf» (Benner & Brüggemann, 2010, S. 194f.).

⁶ Albers (1995, S. 3) verweist auf die Unschärfe des Begriffs «Mündigkeit» als Leitidee der gebildeten *Persönlichkeit*, aber geht nicht weiter darauf ein.

Nach Kaminski (1996; Kaminski et al., 2017) hat Bildung, in Anlehnung Robinsohn (1967) und Klafki (1985, 1998), die grundlegende Aufgabe, den Menschen «zur Bewältigung gegenwärtiger und zukünftiger Lebenssituationen zu befähigen» (Kaminski et al., 2017, S. 33). Er rekurriert ebenfalls auf Mündigkeit als Ziel der Allgemeinbildung und der ökonomischen Allgemeinbildung. Mündigkeit umfasst zwei Dimensionen, die individualistische und die soziale (Kaminski et al., 2017, S. 36f.). Die *individualistische* Dimension bezieht sich im Sinne der Aufklärung auf Selbstbestimmung und Eigenverantwortung. Die *soziale* Dimension bezieht sich auf gesellschaftliche Eingliederung und Teilhabe sowie auf gesellschaftliche Verantwortung.⁷

«[Ökonomische Bildung soll] Kinder und Jugendliche zur Bewältigung ökonomisch geprägter Lebenssituationen [...] befähigen, zur Teilhabe an und Mitgestaltung von Gesellschaft sowie zur Reflexion der ethischen Implikationen wirtschaftlichen Handelns und der wirtschaftsethisch fundierten Lösungsansätze für individuelle und gesellschaftliche Probleme» (Kaminski et al., 2017, S. 37).

Aus dem Bildungs- und Mündigkeitsbegriff leitet Kaminski et al. (2017, S. 37) drei zentrale *Leitbilder* und Situationsfelder der ökonomischen Bildung ab: «der mündige Verbraucher» im Situationsfeld *Konsum*, «der mündige Erwerbstätige» im Situationsfeld *Arbeit und Beruf* sowie «der mündige Wirtschaftsbürger»⁸ im Situationsfeld *Wirtschaftsgesellschaft*. Allerdings sind diese Situationsfelder bzw. Rollen nicht scharf voneinander trennbar, da sich jeder Mensch typischerweise in allen Situationsfeldern bewegt und jede Rolle wahrnimmt.

2.1.2 Curriculare Prinzipien für Lernziele und Lerninhalte der ökonomischen Bildung

Curriculare Prinzipien sind Kriterien für die Ermittlung, Auswahl und Begründung von *relevanten* Lernzielen/Lerninhalten oder Kompetenzbereichen in Lehrplänen (Reetz, 2003, S. 101). In der Fachdidaktik der ökonomischen Bildung werden zur Lösung dieses „Relevanzproblems“ drei curriculare Prinzipien vorgeschlagen (Reetz, 1984, 2003): das Wissenschaftsprinzip, das Situationsprinzip und das Persönlichkeitsprinzip. Alle drei Prinzipien beziehen sich auf den zu bildenden Menschen und sind somit miteinander verbunden bzw. voneinander abhängig (Reetz, 2003, S. 101f.). Es geht also nicht um eine isolierte Verwendung der Prinzipien, sondern um eine ergänzende (Kaminski et al., 2017, S. 90). Zudem muss für die als relevant erachteten Lernziele/Lerninhalte ermittelt und begründet werden, wie *intensiv* sie im Fähigkeitsspektrum ausgeprägt sein sollen, d.h. welche kognitiven Prozesse damit verbunden sind (Reetz, 2003, S. 101).

⁷ Die soziale Dimension der Mündigkeit kann auch als politische Dimension ausgelegt werden. Die politische Dimension bezieht sich auf die politische Stellungnahme und das Urteil hinsichtlich kollektiv verbindlicher Entscheidungen. Diese politische Dimension der Mündigkeit gilt für die (politische) Bildung in einer Demokratie als zentral (Bünger, 2013, S. 18).

⁸ An dieser Stelle sei angemerkt, dass das Leitbild des «mündigen Wirtschaftsbürgers» von Albers (1988, 1995) breiter ist als dasjenige von Kaminski (1996; Kaminski et al., 2017). Bei Albers umfasst es alle drei Lebensbereiche, bei Kaminski bloss den wirtschaftsgesellschaftlichen.

Das *Wissenschaftsprinzip* bestimmt die Inhalte nach vorliegenden wissenschaftlichen und anderen kulturellen Objektivationen (Reetz, 2003, S. 101). Es orientiert sich an der Fachdisziplin bzw. den Fachdisziplinen (z.B. akademischen Lehrbücher, akademische Lehrveranstaltungen). Historisch betrachtet ist diese „gelehrte Bildung“, die sich seit den 1970er Jahren in den Lehrplänen niederschlägt, eine Gegenbewegung zur „volkstümlichen Bildung“ der 1950er/1960er Jahre (Kaminski et al., 2017, S. 85f.). Die Wissenschaftsorientierung in den Lehrplänen wird u.a. mit der Wissenschaftspropädeutik (wissenschaftliches Denken) und mit der zunehmenden Verwissenschaftlichung aller Lebensbereiche begründet, die es notwendig machen, «die Lerninhalte sowohl in ihrer inhaltlichen Struktur wie in ihrer Wahrheitsqualität an den Wissenschaften zu orientieren» (Reetz, 2003, S. 106f.).

In der ökonomischen Bildung sind die Fachdisziplinen die *Wirtschaftswissenschaften* im engeren und die *Sozialwissenschaften* im weiteren Sinne. Die ökonomische Bildung kann entweder darauf ausgelegt sein, die Erkenntnisse der Wirtschaftswissenschaft in reduzierter und vereinfachter Form zu erlernen, eine sog. «Abbild-Didaktik» (Kaminski et al., 2017, S. 84), oder sozioökonomische Problemstellungen mittels wirtschaftswissenschaftlicher Konzepte und Modelle zu analysieren und die ihr Erklärungs- und Gestaltungspotenzial kritisch zu beurteilen (B. Weber, 2013, S. 6). Die deutsche fachdidaktische Forschungsgemeinschaft hat sich darüber in den letzten Jahren gespalten: Das eine Lager proklamiert eine ökonomische Bildung, die auf Wirtschaftswissenschaft als Bezugsdisziplin fokussiert und ihre Zielsetzungen und Inhalte aus der Volkswirtschaft und teilweise aus der Betriebswirtschaft ableitet (vgl. Kaminski et al., 2017; Kaminski et al., 2008; Retzmann et al., 2010; G. Seeber et al., 2012). Das andere Lager plädiert für eine *sozioökonomische* Bildung, die sich an den Sozialwissenschaften als Bezugswissenschaft orientiert und einen pluralistischen und integrativen Ansatz verfolgt (vgl. Engartner & Krisanthan, 2013; Hedtke, 2005, 2006, 2014, 2019; Tafner, 2018; B. Weber, 2005, 2014).

Das *Situationsprinzip* richtet die Inhalte nach den gegenwärtigen und zukünftigen Lebenssituationen, die für die Lernenden bedeutsam sind bzw. bedeutsam werden können (Reetz, 2003, S. 101; vgl. auch Robinsohn, 1967). Es geht also mitunter um das Bestimmen von spezifischen Anforderungen, die für die Bewältigung dieser Situationen notwendig sind. Die Situationen sollen danach ausgewählt, welche Bedeutung ein Gegenstand in den spezifischen Lebenssituationen und in der Wissenschaftsstruktur hat (Kaminski et al., 2017, S. 87f.). Kritisiert wird am Situationsprinzip, dass zwar relevante „Situationen“ heuristisch ermittelt werden können, aber dass damit drei Problemkomplexe ungelöst bleiben (vgl. Tramm & Reetz, 2010): das *Normenproblem*, Situationen sind nicht exogen, sondern werden im Lebensumfeld konstruiert und interpretiert; das *Antizipationsproblem*, zukünftige Situationen sind nicht verlässlich vorhersehbar; das

Theorieproblem, der Zusammenhang zwischen Anforderungssituationen, Kompetenzen und Lerninhalten ist *per se* nicht klar, sondern muss theoretisch begründet werden.

In der ökonomischen Bildung hat der Lebenssituationsansatz eine lange und ausgeprägte Tradition (vgl. z.B. Albers, 1995; Kaminski, 1996; Steinmann, 1997). Die Lebenssituationen sind *ökonomisch geprägte Lebenssituationen*, die in verschiedenen Lebensbereichen auftreten, in denen der Mensch verschiedene Rollen einnimmt, z.B. Verbraucher*in, Erwerbstätige*r und Wirtschaftsbürger*in (vgl. Kaminski et al., 2017, S. 37; Retzmann et al., 2010, S. 14f.; G. Seeber et al., 2012, S. 73). Ökonomische Bildung kann nun darauf ausgerichtet sein, die bestehenden Anforderungssituationen anzunehmen und sich an sie anzupassen oder aber diese Anforderungssituationen zu verstehen und auf der individuellen und politischen Ebene nach Lösungen zu suchen (B. Weber, 2013, S. 6).

Das *Persönlichkeitsprinzip* richtet die Inhalte nach der Persönlichkeitsentwicklung des Lernenden aus (Reetz, 2003, S. 101). Es orientiert sich an einem Bildungsideal, d.h. den normativen Vorstellungen des gebildeten Menschen. Für die ökonomische Bildung ist dies das Leitbild des mündigen Wirtschaftsbürgers bzw. der mündigen Wirtschaftsbürgerin. Die ökonomische Bildung zielt sodann auf seine persönliche Lebensbewältigung und seine gesellschaftliche Teilhabe ab (B. Weber, 2013, S. 6).

2.1.3 Bildungsstandards und Kompetenzen für die ökonomische Bildung

Bildungsstandards und Kompetenzen sind in den deutschsprachigen Ländern seit den 2000er-Jahren Kernelement, aber auch Streitpunkt von Bildungsreformen. Allerdings bestehen in den verschiedenen Ländern unterschiedliche Konzepte von Bildungsstandards (Maag Merki, 2007, S. 17, S. 20). Im Allgemeinen können drei verschiedene Konzepte von Bildungsstandards unterschieden werden (Maag Merki, 2007, S. 21):

- *Performance Standards*: beziehen sich auf die Lernergebnisse bzw. auf die Leistungen der Lernenden (Output-Orientierung), d.h. welche *fachlichen Kompetenzen* bis zu einem bestimmten Zeitpunkt erworben werden; idealerweise werden Kompetenzbereiche (Anforderungsbereiche) und Kompetenzniveaus (Anforderungsniveaus i.S.v. Mindeststandards, Maximalstandards) formuliert. Im internationalen Vergleich werden die meisten Bildungsstandards als *Performance Standards* formuliert.⁹

⁹ In der Schweiz enthält der Lehrplan 21 für die Grundschule sechs Fachbereiche (z.B. Deutsch, Mathematik, Natur/Mensch/Gesellschaft). Jeder Fachbereich ist in Kompetenzbereiche gegliedert, je nach fachdidaktischer Tradition sind dies Fähigkeiten (z.B. für Sprachen: Hören, Lesen, Sprechen, Schreiben; für Mathematik: Zahl und Variable). Zu jedem Kompetenzbereich werden Kompetenzen formuliert (z.B. für Sprachen: „Die Schülerinnen und Schüler können vielfältige Schreibprodukte verfassen und Struktur, Inhalt, Sprache und Form auf das Schreibziel ausrichten“. Für jede Kompetenz wird der erwartete Kompetenzerwerb in mehreren Kompetenzstufen beschrieben, die Kompetenzstufen unterscheiden sich durch Zunahme an Fakten-, Konzept- und Prozesswissen oder durch die Zunahme an Komplexität der Anwendungssituation.

- *Content Standards*: beziehen sich auf Lernziele und Lerninhalte des schulischen Lernens, d.h. welche Inhalte in bestimmten Zeitphasen behandelt werden.
- *Opportunity-to-learn-Standards*: beziehen sich auf die Ausstattung von Schulen, auf die Gestaltung von Unterricht und auf die Qualifikation von Lehrer*innen (Input-Orientierung).

In Deutschland wurde in der „Expertise zur Entwicklung nationaler Bildungsstandards“ von Klieme, Avenarius, et al. (2007) ein Konzept für Bildungsstandards vorgeschlagen, das den *Performance Standards* zuzuordnen ist:

«Bildungsstandards [...] greifen allgemeine Bildungsziele auf. Sie benennen die Kompetenzen, welche die Schule ihren Schülerinnen und Schülern vermitteln muss, damit bestimmte zentrale Bildungsziele erreicht werden. Die Bildungsstandards legen fest, welche Kompetenzen die Kinder oder Jugendlichen bis zu einer bestimmten Jahrgangsstufe erworben haben sollen. Die Kompetenzen werden so konkret beschrieben, dass sie in Aufgabenstellungen umgesetzt und prinzipiell mit Hilfe von Testverfahren erfasst werden können» (Klieme, Avenarius, et al., 2007, S. 19).

Nach diesem Konzept haben Bildungsstandards zwei Funktionen. Erstens bieten sie Schulen einen *Referenzrahmen* mit verbindlichen Bildungszielen (Klieme, Avenarius, et al., 2007, S. 90ff.): Bildungsstandards konkretisieren den Bildungsauftrag der Schule und formulieren verbindliche Anforderungen an das Lehren und Lernen in der Schule. Sie haben insofern Konsequenzen für die Entwicklung von Bildungsplänen. Zweitens sind standard-bezogene Testverfahren ein Instrument für die *Qualitätssicherung im Bildungssystem* (Klieme, Avenarius, et al., 2007, S. 99ff.), und zwar auf der Ebene des Systems (sog. Bildungsmonitoring) und auf der Ebene der einzelnen Schule (sog. Schulevaluation) sowie ein Instrument für die *Individualdiagnostik* von schulischen Leistungen. Basierend auf den Bildungsstandards werden Lernergebnisse erfasst und bewertet sowie hinsichtlich der anzustrebenden Kompetenzen evaluiert.

Gute Bildungsstandards lassen sich entlang der folgenden Merkmale beschreiben (Klieme, Avenarius, et al., 2007, S. 24ff.):

- *Fachlichkeit*: beziehen sich auf einen *bestimmten Lernbereich* bzw. auf ein bestimmtes Unterrichtsfach (auch: Domäne);
- *Fokussierung*: konzentrieren sich auf den *Kernbereich des Lernbereichs*, decken nicht die gesamte Breite und Verzweigung des Lernbereichs ab;
- *Kumulativität*: beziehen sich auf Kompetenzen, die bis zu einem bestimmten Zeitpunkt in der Lernbiografie aufgebaut worden sind, zielen also auf systematisch vernetztes Lernen;

- Verbindlichkeit für alle: drücken *Mindestanforderungen* aus, die von allen Lernenden schulformübergreifend erwartet werden;
- Differenzierung: unterscheiden zwischen Kompetenzniveaus, die über und unter den Mindestanforderungen liegen, machen Lernentwicklungen nachvollziehbar;
- Verständlichkeit: sind klar und knapp formuliert;
- Realisierbarkeit: mit realistischem Aufwand zu erreichen, obschon eine Herausforderung für Lehrende und Lernende.

Die Debatte um Bildungsstandards und ihre Kompetenzen (vgl. Klieme, Avenarius, et al., 2007) und der sog. „PISA-Schock“ stellte die deutschen Fachdidaktiken vor die Aufgabe, ihre Domäne kompetenztheoretisch zu fundieren und ihre Bildungsziele zu standardisieren. In der Folge erarbeiteten verschiedene fachdidaktische Forschungsgruppen „Kompetenzmodelle“ für die ökonomische Bildung. Diese Kompetenzmodelle für ökonomische Bildung unterscheiden sich zum einen in der Beschreibung und Eingrenzung der Domäne und zum anderen in der Beschreibung und Verknüpfung der Kompetenzbereiche.

2.1.3.a Der Vorschlag der DEGÖB (2004, 2006, 2009)

Die „Deutsche Gesellschaft für ökonomische Bildung“ (DEGÖB) entwickelte, ausgehend von mehreren Expertenrunden und einer Fachtagung, Kompetenzen der ökonomischen Bildung und Standards. Die Standards der DEGÖB waren die ersten für die ökonomische Bildung in Deutschland und stellten einen Meilenstein in der deutschen Wirtschaftsdidaktik dar. Sie beziehen sich auf den Grundschulabschluss (DEGÖB, 2006), den mittleren Schulabschluss (DEGÖB, 2004) und den höheren Schulabschluss (gymnasiale Oberstufe) (DEGÖB, 2009). Die DEGÖB definiert ökonomische Bildung wie folgt:

«Ökonomische Bildung ist also das individuelle Vermögen sich in ökonomisch geprägten Lebenssituationen und Entwicklungen einer immer schneller sich verändernden Wirtschaftswelt zu orientieren, zu urteilen, zu entscheiden, zu handeln und mitzugestalten. Ökonomische Bildung soll Menschen zu einem mündigen Urteil, zur Selbstbestimmung und zur verantwortlichen Mitgestaltung befähigen» (DEGÖB, 2004, S. 4).

Die DEGÖB konkretisierte den Kern der ökonomischen Bildung durch fünf Fähigkeiten, sog. Kompetenzbereiche, die für alle Lebenssituationen bzw. Rollen gleichermaßen gelten (DEGÖB, 2009, S. 4ff.; übersichtlich bei G. Seeber et al., 2012, S. 29ff.): (A) Entscheidungen ökonomisch begründen; (B) Handlungssituationen ökonomisch analysieren; (C) ökonomische Systemzusammenhänge erklären; (D) Rahmenbedingungen des Wirtschaftens verstehen und mitgestalten; (E) Konflikte

perspektivisch und ethisch beurteilen. Für jeden Kompetenzbereich werden 5 bis 7 Standards formuliert, für den höheren Schulabschluss bis zu deren 13.

Die *formelle Kritik* am Vorschlag der DEGÖB bezieht sich darauf, dass nur drei Merkmale guter Bildungsstandards (vgl. Klieme, Avenarius, et al., 2007, S. 24ff.) umgesetzt wurden (G. Seeber et al., 2012, S. 36): Fachlichkeit, Fokussierung und Kumulativität. Die *inhaltliche Kritik* am Vorschlag der DEGÖB bezieht sich auf die Beschreibung der Kompetenzbereiche und der Standards (G. Seeber et al., 2012, S. 37ff.): unklare Verknüpfung von Leistungsstandards (*performance standards*) und Inhaltsstandards (*content standards*), unvollständige und unklare Kompetenzbeschreibungen, inkonsistente Zuordnung von Standards zu Kompetenzbereichen, überkomplexe und unerreichbare Standards, ungleiche Gewichtung der Kompetenzbereiche (v.a. beim höheren Schulabschluss).

Das grösste Defizit des Vorschlags der DEGÖB wird darin gesehen, dass kein theoretisch fundiertes und empirisch bewährtes Kompetenzmodell vorgelegt wurde. Diese Kritik ist allerdings zur teilweise gerechtfertigt, da dem Vorschlag der DEGÖB zumindest ein implizites Kompetenzmodell zugrunde liegt, das von Retzmann (2005) hermeneutisch nachgezeichnet wurde. Es beinhaltet *agentives Handeln* i.S.v. ökonomischem Handeln (Anbieten, Nachfragen, Produzieren, Konsumieren, Investieren, Sparen, Konkurrieren, Kooperieren etc.) und *reflexives Handeln* i.S.v. ökonomischem Denken (Analysieren, Entscheiden, Urteilen, Begründen). Es beinhaltet verschiedene Perspektiven (Akteur*in auf Märkten, Adressat*in von anderen Akteuren, Beobachter*in, Bürger*in).

2.1.3.b Der Vorschlag des IÖB (Kaminski et al., 2008, 2017)

Die Forschungsgruppe um Kaminski et al. (2008, 2017) hat ein allgemeinbildendes Verständnis von ökonomischer Bildung, das sich auf alle Jahrgangsstufen und auf verschiedene Lebensbereiche bezieht:

«Ökonomische Bildung wird verstanden als die Gesamtheit aller pädagogischen Bemühungen in allgemeinbildenden Schulen, Kinder und Jugendliche mit solchen Kenntnissen, Fähigkeiten, Fertigkeiten, Verhaltensbereitschaften und Einstellungen auszustatten, die sie befähigen, sich mit den ökonomischen Bedingungen ihrer Existenz und deren sozialen, politischen, rechtlichen, technischen, ökologischen und ethischen Dimensionen auf privater, betrieblicher, volkswirtschaftlicher und weltwirtschaftlicher Ebene auseinanderzusetzen» (Kaminski, 1996, S. 18f.).

Die Forschungsgruppe entwickelte ein Kompetenzmodell, das sich an den Merkmalen guter Bildungsstandards orientiert (vgl. Klieme, Avenarius, et al., 2007, S. 24ff.) und diese für die ökonomische Bildung interpretiert (Kaminski et al., 2008, S. 19ff.). Die *Domäne* der ökonomischen Bildung wird über den «Gegenstandsbereich Wirtschaft» beschrieben, d.h. die *Wirtschaftsordnung*, den Wirtschaftskreislauf (private Haushalte, Unternehmen, Staat, Ausland/internationale Wirtschaftsbeziehungen) sowie Arbeit und Beruf. Begründet wird die Wahl dieses Referenzsystems

damit, dass bei einer reinen Orientierung an Lebenssituationen ökonomische Bildung zur rezeptartigen «Lebenshilfe» (Kaminski, 1996, S. 38) oder zur oberflächlichen «Aktualitätenschau» (Dubs, 2011, S. 200) verkommen könne.

«Als wesentliches Element des Referenzsystems wird die jeweils geltende Wirtschaftsordnung angesehen, die den ökonomischen Rahmen für das Handeln der Individuen bildet und sich gleichzeitig als permanente ordnungspolitische Gestaltungsaufgabe darstellt. [...] Die Erschließung von Lebenssituationen kann allerdings nur gelingen, wenn die jeweilige Erkenntnisperspektive einer Domäne offengelegt und damit typische Denk- und Analyseschemata zur Verfügung gestellt werden» (Kaminski et al., 2017, S. 38).

Das Kompetenzmodell beinhaltet drei Kompetenzbereiche, die von vier fachdidaktischen Ordnungsversuchen umklammert werden. Die *Ordnungsversuche* sind Denkansätze, die jeweils alternativ oder kumulativ zur Analyse wirtschaftlicher Phänomene herangezogen werden können (Kaminski et al., 2017, S. 129ff.; Kaminski et al., 2008, S. 13ff.):

- (1) Denken in Kategorien eines ökonomischen Verhaltensmodells (*rational choice*): Entwicklung eines Verhaltensmodells für Individuen; Nutzen/Kosten-Überlegungen unter Beachtung des institutionellen Rahmens (vgl. Kruber, 2005, S. 86ff.).
- (2) Denken in Kreislaufzusammenhängen: Auseinandersetzung mit zentralen Akteuren sowie deren Beziehungen zueinander im Wirtschaftsprozess; Strukturen und Zusammenhänge analysieren und beurteilen (vgl. Kruber, 2005, S. 90ff.).
- (3) Denken in Ordnungszusammenhängen: Erfassen und Ordnen des Institutionen- und Regelsystems einer Wirtschaftsordnung; Wirtschaftspolitik als ordnende Gestaltungsaufgabe verstehen (vgl. Kruber, 2005, S. 93ff.).
- (4) Denken in Kategorien: Identifizieren von Kategorien, die allen wirtschaftlichen Handlungen immanent sind; fachwissenschaftliche Kategorien (Stoffkategorien)¹⁰ (vgl. Dauenhauer, 2001; Kruber, 2000; May & Albers, 2008); ökonomische und politische Themenfelder (vgl. Kruber, 2005, S. 97ff.).

Die *Kompetenzbereiche* werden als Fachwissen, Erkenntnisgewinnung und Bewertung bezeichnet (Kaminski et al., 2017, S. 137ff.; Kaminski et al., 2008, S. 26ff., S. 42ff.), die jedoch domänenunspezifisch beschrieben sind. Zu jedem Kompetenzbereich werden sodann sechs „generelle“ Kompetenzen für die Jahrgangsstufen 4, 6, 8, 10 und 12 formuliert.

¹⁰ *Fachwissenschaftlichen Kategorien* sind (May, 2011, S. 5ff.): Menschliches Handeln ist bedürfnisgetrieben; die Knappheit der Güter zwingt den Menschen zu wirtschaftlichem Handeln; wirtschaftliches Handeln ist konfliktgeprägt; wirtschaftliches Handeln ist entscheidungsbestimmt; wirtschaftliches Handeln ist risikobehaftet; wirtschaftliches Handeln ist nutzen- bzw. gewinnorientiert; wirtschaftliches Handeln impliziert Arbeitsteilung; wirtschaftliches Handeln schafft Interdependenz; wirtschaftliches Handeln bedarf der Koordination; wirtschaftliches Handeln führt zu Ungleichheit; Ungleichheit induziert Leistungsstreben; Fortschritt und Wohlstand; Wohlstand fundiert Freiheit und Macht; jeder ist sein eigener Unternehmer; wirtschaftliches Handeln/Geschehen vollzieht sich in Kreislaufprozessen.

Davon abzugrenzen sind *fachwissenschaftliche Bildungsinhalte* (May, 2011, S. 5ff.; May & Albers, 2008): Bedürfnisse, Güter, Produktionsfaktoren, ökonomisches Prinzip, Markt, Konsumökonomie (Verbraucherrecht, Verbraucherpolitik), Arbeitsökonomie (Arbeit und Qualifikation, Arbeitsrecht, Sozialpartnerschaft), Gesellschaftsökonomie (soziale Sicherung, private Altersvorsorge, soziale Marktwirtschaft, wirtschaftliche Integration).

- (A) Fachwissen (Sach- und Analysekompetenz): Die Schüler*innen verfügen über strukturiertes ökonomisches Wissen, das ihnen das Wiedererkennen von wirtschaftlichen Strukturelementen, Invarianzen und Prozessregeln in der Fülle der ökonomischen Phänomene ermöglicht.
- (B) Erkenntnisgewinnung (Methodenkompetenz): Die Schüler*innen analysieren wirtschaftliche Phänomene mithilfe fachspezifischer Methoden und wenden fachspezifische Arbeitstechniken an.
- (C) Bewertung (Urteilskompetenz): Die Schüler*innen bewerten ökonomische Handlungen und Sachverhalte und reflektieren Wege des Erkennens und Urteilens.

2.1.3.c Der Vorschlag von Retzmann et al. (2010) und Seeber et al. (2012)

Die Forschungsgruppe rund um Retzmann et al. (2010) und G. Seeber et al. (2012) versteht ökonomische Bildung ebenfalls als allgemeinbildend, aber begründet sie funktional:

«Der ökonomisch gebildete Mensch kann [...] seine Interessen in der heutigen Wirtschaft und Gesellschaft mündig vertreten, sachkundig urteilen und verantwortlich handeln. Ökonomische Bildung ist deshalb ein unverzichtbarer Bestandteil der Allgemeinbildung» (Retzmann et al., 2010, S. 12).

«Ökonomische Bildung befähigt den Menschen zur individuellen Lebensführung, zur gesellschaftlichen Teilhabe sowie zur politischen Mitwirkung; sie trägt damit zur Integration der Heranwachsenden in die Gesellschaft bei. Von einer soliden ökonomischen Grundbildung profitieren gerade auch bildungsferne Schichten, da sie berufliche Chancen eröffnet, finanzielle Risiken verdeutlicht und somit zu einer in wirtschaftlicher Hinsicht gelingenden Lebensgestaltung beiträgt» (Retzmann et al., 2010, S. 11).

Die Forschungsgruppe entwickelte, ausgehend von ihrer Kritik am Vorschlag der DEGÖB, ein Kompetenzmodell für ökonomische Bildung, das den Merkmalen guter Bildungsstandards (vgl. Klieme, Avenarius, et al., 2007, S. 24ff.) gerecht wird. Die *Domäne* der ökonomischen Bildung wird über die «ökonomische Perspektive» beschrieben und über die Rollenkonzepte konkretisiert.

«Die Rollenkonzepte erheben nicht den Anspruch auf Vollständigkeit, sondern auf Maßgeblichkeit. Sie strukturieren die ökonomisch geprägte Lebenswelt und nicht die Kompetenzbereiche ökonomischer Bildung, denn es gibt nicht die Kompetenzen eines Verbrauchers, Arbeitnehmers, Unternehmers oder Wirtschaftsbürgers. Kompetenzen sind an ihren Träger gebunden, nicht an die spezifische Situation, in der sie gebraucht werden» (Retzmann et al., 2010, S. 15).

Kompetenzbereiche	Teilkompetenzen
A Entscheidung und Rationalität	A 1 Situationen analysieren A 2 Handlungsalternativen bewerten A 3 Handlungsmöglichkeiten gestalten
B Beziehung und Interaktion	B 1 Interessenkonstellationen analysieren B 2 Kooperationen analysieren, bewerten und gestalten B 3 Beziehungsgefüge analysieren
C Ordnung und System	C 1 Märkte analysieren C 2 Wirtschaftssysteme und Ordnungen analysieren C 3 Politik ökonomisch beurteilen und gestalten

Abbildung 2-1. Kompetenzmodell für ökonomische Bildung nach Seeber et al. (2012)

Quelle: G. Seeber et al. (2012, S. 92) und Retzmann et al. (2010, S. 19).

Aus der Verbindung der Perspektiven und Rollenkonzepte werden drei *Kompetenzbereiche* identifiziert (G. Seeber et al., 2012S. 76, S. 86ff.) (vgl. Abbildung 2-1). Dabei wird betont, dass die Kompetenzbereiche «eine sachliche Klassifizierung» sind und keine Entwicklungsstufen darstellen (G. Seeber et al., 2012, S. 91). Die für den Teilkompetenzen verwendeten Verben «beziehen sich auf ein beschreibbares Spektrum von Fähigkeiten; sie verweisen nicht auf taxonomische Stufen» (ebd., S. 91).

- (A) Entscheidung und Rationalität (des Einzelnen): individuelle Perspektive; ökonomisch begründete Entscheidungen treffen zwischen bestehenden Alternativen, persönliche Wünsche und Werte berücksichtigen (Präferenzen), Vor- und Nachteile von Handlungsalternativen abwägen (Kosten/Nutzen).
- (B) Beziehung und Interaktion (mit anderen): soziale Perspektive; beim wirtschaftlichen Handeln die Wünsche und Werte von anderen berücksichtigen; die Konsequenzen des eigenen Handelns für andere berücksichtigen.
- (C) Ordnung und System (des Ganzen): systemische Perspektive; das Wirtschaftssystem und die politische Ordnung als gestaltbar verstehen.

Innerhalb der Teilkompetenzen werden *Kompetenzanforderungen* für den Abschluss des Primarbereichs, der Hauptschule und des Abiturs definiert (Retzmann et al., 2010, S. 19ff.; G. Seeber et al., 2012S. 91ff.). Für jeden Abschluss werden *exemplarische Aufgaben* mitgeliefert, um das Anforderungsniveau der Kompetenzen zu veranschaulichen (Retzmann et al., 2010, S. 38ff.). Die Aufgaben adressieren alle Rollen, allerdings erfolgt die Auswahl der Lebenssituationen gemäss den Kompetenzen nach dem Prinzip der Horizonterweiterung: Beim Abschluss des Primarbereichs überwiegen Konsumsituationen, beim Abschluss der Sekundarstufe I (Hauptschule) kommen Situationen der Geldanlage, Kreditaufnahme, Berufsausbildung und Unternehmensgründung hinzu, beim Abschluss der Sekundarstufe II (gymnasiale Oberstufe) die Einübung von Methoden der Erkenntnisgewinnung und Hypothesenprüfung.

2.1.3.d Der Vorschlag von Weber (2005, 2013, 2014)

Die Forschungsgruppen aus dem sozialwissenschaftlichen Bereich, so z.B. B. Weber (2005, 2014), Hedtke (2006, 2015, 2019) und Tafner (2016, 2018) kritisieren an den obigen Vorschlägen ihre Fokussierung auf die *Ökonomik* als Bezugswissenschaft für Denk- und Analyseschemata bei gleichzeitiger Perspektive auf die *reale Ökonomie*, d.h. ökonomisch geprägte Lebenssituationen. Bei den obigen Vorschlägen zeigt sich nämlich «eine strukturelle Verengung [...] der Kompetenzbezeichnung. So sollte dazu befähigt werden, Entscheidungen ökonomisch zu begründen statt ökonomisch (geprägte) Entscheidungen begründet zu treffen, Handlungssituationen sollten ökonomisch analysiert werden statt ökonomisch (geprägte) Handlungssituationen zu analysieren» (B. Weber, 2013, S. 9).

Diese Forschenden (statt vieler vgl. A. Fischer & Zurstrassen, 2014) plädieren deshalb für einen sozialwissenschaftlichen bzw. sozioökonomischen Ansatz der ökonomischen Bildung. Angesichts der engen Verflechtung von Wirtschaft, Politik und Gesellschaft in den verschiedenen Lebensbereichen sei es fragwürdig, inwiefern eine Orientierung an der Wirtschaftswissenschaft für die ökonomische Bildung hinreichend sei.

«Der sozialwissenschaftliche Ansatz hält die Frage nach den Bezugswissenschaften ökonomischer Bildung bewusst offen und beantwortet sie pragmatisch. Er behandelt die Wirtschaftswissenschaften nicht als a priori bevorzugte Bezugsdisziplinen. Vielmehr bevorzugt er jeweils die Disziplin, die für die konkreten Ziele und angestrebten Kompetenzen ökonomischer Bildung die am besten geeigneten Wissensbestände anbieten kann» (Hedtke, 2006, S. 95f.)

Kutscha (2014, S. 74) schlägt für die sozioökonomischer Bildung drei *Kompetenzbereiche* vor:

- Wahrnehmungs- und Orientierungsfähigkeit: wirtschaftliche Phänomene im Kontext ihrer sozialen und politischen Handlungsbedingungen differenziert wahrnehmen und verstehen;
- Problemlösungs- und Entscheidungsfähigkeit: vernünftig begründungsfähige Entscheidungen in Bezug auf (ausgewählte und für die Lernenden bedeutsame) ökonomische Handlungspraxen entwickeln;
- Urteils- und Kritikfähigkeit: gesellschaftliche und individuelle Folgen ökonomischer Entscheidungen und wirtschaftspolitischer Maßnahmen kritisch abwägen und selbständig beurteilen.

Eine vergleichende Analyse von B. Weber (2013) zu den obigen Vorschlägen (DEGÖB, 2004, 2006, 2009; Kaminski et al., 2017; Kaminski et al., 2008; Retzmann et al., 2010; G. Seeber et al., 2012) zeigt, dass die Strukturierung der Kompetenzbereiche eine ähnliche Dreiteilung hat (vgl. Tabelle 2-1). Es gibt jeweils drei Perspektiven – Individuum, Beziehung und Ordnungen/Institution –, in jeder Perspektive sind drei Handlungen charakteristisch – Entscheidung, Interaktion und Koordination.

Tabelle 2-1. Strukturierung der Kompetenzbereiche der ökonomischen Bildung

Perspektive	Handlung	Analysegegenstand
Individuum	Entscheidung	- Situation und Entscheidung des Einzelnen, Rationalität; - individuelle Entscheidungsebene zur Bewältigung von wechselnden ökonomisch geprägten Lebenssituationen
Beziehungen (Gruppe, Gemeinschaft)	Interaktion	- Beziehung und Interaktion mit Anderen, Interessen, Ethik; - ethischen Bewertungsebene für wechselnde Rollen als Verbraucher, Erwerbstätiger, Wirtschaftsbürger
Ordnungen und Institutionen	Koordination	- Ordnung und System des Ganzen, Politik; Wirtschafts- und Gesellschaftsordnung, Wirtschaftspolitik; - politische Gestaltungsebene zur gesellschaftlichen Teilhabe

Quelle: Eigene Darstellung i.A.a. B. Weber (2013, S. 9f.).

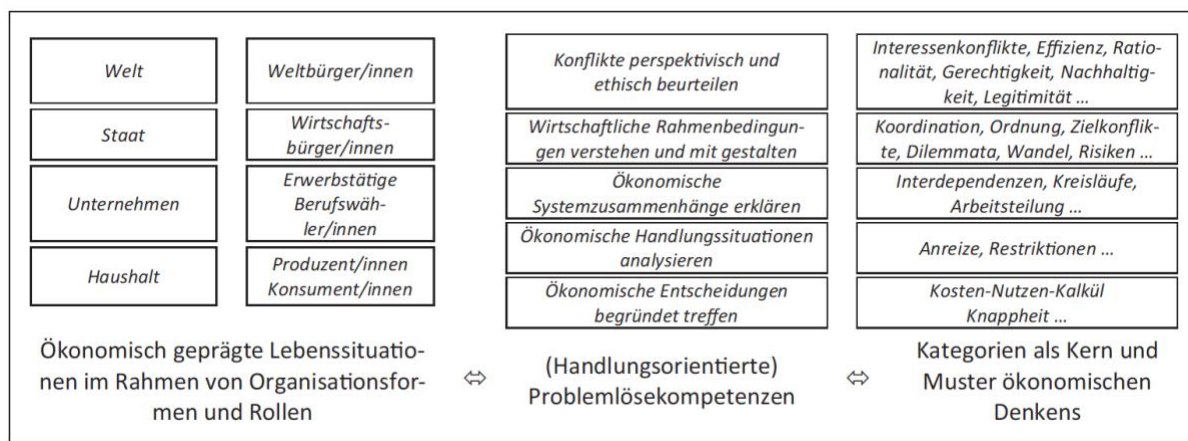


Abbildung 2-2. Kompetenzmodell für sozioökonomische Bildung nach Weber (2005)

Quelle: B. Weber (2013, S. 9).

B. Weber (2005, 2013, 2014) entwickelte ein Kompetenzmodell für sozioökonomische Bildung, das sie im Spannungsfeld zwischen Subjekt, Lebenswelt und Wissenschaft aufspannt und mit dem sie die obigen Vorschläge integrativ verbindet (vgl. Abbildung 2-2). Ausgangspunkt sind ökonomisch geprägte Lebenssituationen, die sich im Rahmen von Organisationsformen (vgl. Kaminski et al., 2017; Kaminski et al., 2008: Wirtschaftsordnung, Wirtschaftskreislauf) und Rollen (vgl. Retzmann et al., 2010; G. Seeber et al., 2012) zeigen. Wissenschaftliche Kategorien geben nicht die Lerninhalte vor, sondern fungieren als *Instrumente* (Denk- und Analyseschemata), um Wissen anzueignen und anzuwenden (B. Weber, 2013, S. 8). Solche *Kategorien ökonomischen Denkens* hat Kruber (1995, S. 98ff; 2005, S. 85ff.) umfassend herausgearbeitet; er nennt z.B. Knappheit, Rationalität (Kosten-Nutzen-Überlegungen), Arbeitsteilung, Marktkoordination, Wirtschaftskreislauf, Interdependenz/Zielkonflikte, Wandel/Instabilität, Staatseingriffe, Interessenkonflikte, Wertbezug, Wirtschaftsordnung. Als wichtigste *Beurteilungskriterien* der Domäne werden *Effizienz* (Zweck-Mittel-Optimierung), *Rationalität* (Kosten-Nutzen-Abwägungen) und *Gerechtigkeit* angesehen.

Die fünf *Kompetenzbereiche* werden als *handlungsorientierte Problemlösekompetenzen* beschrieben. Sie werden durch ihren Bezug zur gesellschaftlichen Mikro- und Makroebene konkretisiert und sind deshalb nicht mit kognitiven Anforderungsniveaus zu verwechseln (B. Weber, 2013, S. 9).

- (1) ökonomische Entscheidungen begründet treffen (Mikroebene: individuelle Entscheidungen);
- (2) wirtschaftliche Handlungssituationen analysieren (Mikroebene: individuelle Entscheidungen);
- (3) ökonomische Systemzusammenhänge erklären (Makroebene: politische Ordnungsgestaltung);
- (4) ökonomische Rahmenbedingungen verstehen und mitgestalten (Makroebene: politische Ordnungsgestaltung);
- (5) Konflikte perspektivisch und ethisch beurteilen (Mikroebene: individuelle Entscheidungen; Makroebene: politische Ordnungsgestaltung).

Auch für dieses Kompetenzmodell gibt es bis anhin keine empirische Erprobung. Jedoch entwickelte Hedtke (2019) kürzlich einen Vorschlag für ein sozioökonomisches Curriculum.

2.2 Wirtschaftsbürgerliche Bildung (die Schweizer Fachdidaktik)

Diese Kapitel behandelt die Konzeption der wirtschaftsbürgerlichen Bildung aus der Schweizer Fachdidaktik. Zuerst wird das Leitbild und das Bildungsziel erläutert (Kapitel 2.2.1), danach die vorgeschlagene Gestaltung von Lehrplan und Unterricht (Kapitel 2.2.2) und schliesslich ein Vorschlag für Bildungsstandards im Fach „Wirtschaft & Recht“ (Kapitel 2.2.3).

2.2.1 Leitbild und Bildungsziel der wirtschaftsbürgerlichen Bildung

In der Schweizer Fachdidaktik finden sich Konzeptionen der ökonomischen Bildung bzw. der ökonomischen Allgemeinbildung seit den 1980er-Jahren und wurden massgeblich vom Wirtschaftspädagogen Rolf Dubs und vom Wirtschaftsethiker Peter Ulrich geprägt.

Dubs (1985, 2001, 2011, 2013) unterscheidet ökonomische Bildung hinsichtlich ihrer normativen Zielsetzung und curricularen Ausgestaltung in wirtschaftsbürgerliche und wirtschaftsberufliche Bildung (vgl. Abbildung 2-3). *Wirtschaftsberufliche* Bildung zielt auf die Entwicklung von berufs- und betriebsspezifischen Fähigkeiten und Fertigkeiten ab und soll auf eine bestimmte berufliche Tätigkeit vorbereiten (Dubs, 1985, S. 65; 2011, S. 194; 2013, S. 15).¹¹ *Wirtschaftsbürgerliche* Bildung ist die Bildung des allgemeinen Wirtschafts- und

¹¹ Im dualen Berufsbildungssystem der Schweiz sind Lernziele/Lerninhalte der wirtschaftsberuflichen Bildung in den schulischen und betrieblichen Lehrplänen verankert.

Gesellschaftsverständnisses. Sie soll dazu befähigen, wirtschaftliche und gesellschaftliche Zusammenhänge zu verstehen und die darin bestehenden/entstehenden Problemstellungen zu beurteilen (Dubs, 1985, S. 65; 2011, S. 194; 2013, S. 15).¹² Entsprechend ist wirtschaftsbürgerliche Bildung als wirtschaftliche Allgemeinbildung zu verstehen und sollte allen Menschen zukommen, nicht nur einer vom Bildungszugang her privilegierten Schicht. Die Zielsetzung dieser Bildung wird wie folgt formuliert:

«Sie [die Bildung des allgemeinen Wirtschafts- und Gesellschaftsverständnisses] muss der jungen Generation dasjenige volkswirtschaftliche und betriebswirtschaftliche Wissen vermitteln, das sie befähigt, gesamtwirtschaftliche und unternehmerische Zusammenhänge zu verstehen, aktuelle Problem und Streitfragen in ihrer gesellschaftlichen, politischen und wirtschaftlichen Interdependenz zu erkennen und zu definieren, darin liegende Zielkonflikte zu erfassen und sie einer reflektierten Lösung zuzuführen» (Dubs, 2011, S. 196).

Nach Ulrich (1993, 2001) soll wirtschaftsbürgerliche Bildung die Menschen dazu befähigen, wirtschaftliche Vorgänge dahingehend zu beurteilen, «ob und wie weit sie [...] in "rein" ökonomischer Perspektive (d.h. jener der marktwirtschaftlichen System- oder Sachlogik) rational [...] [und] in einem umfassenden Sinn (d.h. aus der Perspektive der humanen Lebenswelt) vernünftig sind» (Ulrich, 2001, S. 2). Das Leitbild der wirtschaftsbürgerlichen Bildung sind mündige Wirtschaftsbürger*innen, welche die Funktionslogik des Marktes verstehen und diese «im Kontext *persönlicher Sinnorientierungen* und *gesellschaftlicher Legitimitätsansprüche* kritisch [...] reflektieren» (Ulrich, 2001, S. 3, Hervorhebungen im Original).

Ulrich entwirft eine wirtschaftsbürgerliche Bildung, welche die vorherrschende *Systemökonomie*, d.h. Ökonomie als autonomes marktwirtschaftliches Wirtschaftssystem, um eine *Sozialökonomie*, d.h. Ökonomie als gesellschaftliches Wirtschaftsleben, ergänzt (Ulrich, 2001, S. 5f.). Wirtschaftsbürgerliche Bildung soll eine Perspektive der «Wirtschaft in der Gesellschaft» einnehmen, bei der die Wirtschaft der Gesellschaft dient, nicht umgekehrt. Ulrich kritisiert insbesondere aus einer wirtschaftsethischen Perspektive, dass die neoklassisch geprägte *mainstream economics* sich von Politik und Ethik losgelöst und sich «in einem sozialen und ethischen Vakuum» (Ulrich, 2001, S. 1) bewegt. Diese *mainstream economics* versteht sich als reine „Systemökonomie“, welche die normative Logik des „freien“ Marktes und das „rationale“ Verhalten Menschen konsequent entfalten lässt, aber sie nicht kritisch reflektiert, geschweige denn zähmt (Ulrich, 2001, S. 2).

¹² Im Schweizer Bildungssystem ist eine so definierte wirtschaftsbürgerliche Bildung curricularer Bestandteil an allgemeinbildenden Schulen (gymnasiale Maturitätsschulen, Fachmaturitätsschulen) und an beruflichen Schulen (Berufsfachschulen, Berufsmaturitätsschulen) (vgl. Kapitel 2.3.1).

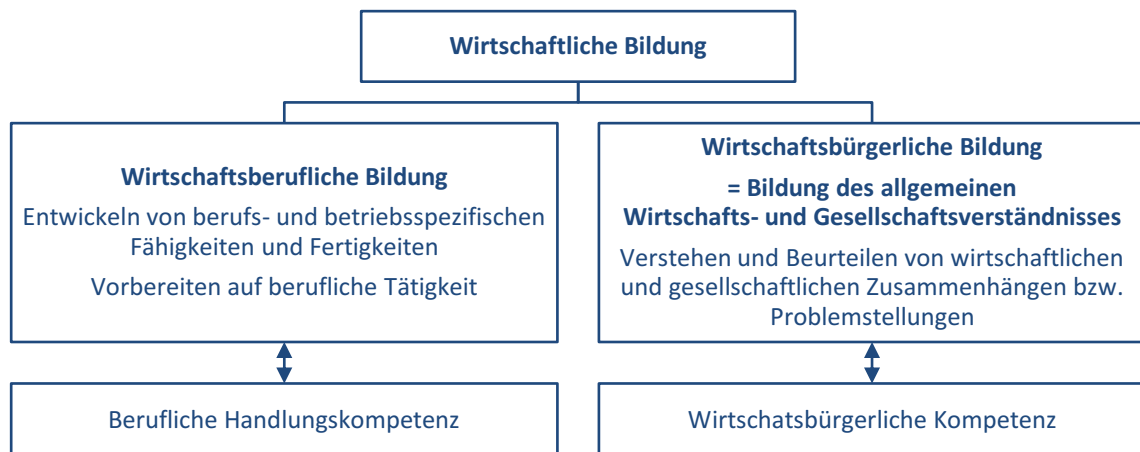


Abbildung 2-3: Wirtschaftliche Bildung mit zweigeteilter Zielsetzung nach Dubs (2011, 2013)

Quelle: Eigene Darstellung, i.A.a. Dubs (1985, S. 65; 2011, S. 194; 2013, S. 15).

2.2.2 Gestaltung von Lehrplan und Unterricht

Dubs beschreibt die „Domäne“ der wirtschaftsbürgerlichen Bildung über die Zeit hinweg inkonsistent, d.h. er hat seine Beschreibung im Laufe der Jahre inhaltlich „verbreitert“. In seinen früheren Arbeiten (vgl. Dubs, 1985, 1998) gehört zur wirtschaftsbürgerlichen Bildung das Verstehen des «gesamtwirtschaftliche[n] und gesamtgesellschaftliche[n]» (Dubs, 1985, S. 65) Geschehens, das in seinen Konsequenzen alle Menschen in verschiedenen Lebensbereichen betrifft. Diese Bildung soll die Menschen «in volkswirtschaftlichen Fragestellungen so entscheidungsfähig zu machen, dass sie im freien Urteil zu einer eigenen Meinung gelangen» (Dubs, 1985, S. 65). Entsprechend verwendet er den Begriff «Wirtschaftsbürgerkunde» bis Ende der 1990er-Jahre synonym für Volkswirtschaftslehre (vgl. Dubs, 1998; aber auch Dubs, 2004).¹³ In seinen späteren Arbeiten (vgl. Dubs, 2001, 2011, 2013) gehört zur wirtschaftsbürgerlichen Bildung das gesamtheitliche Verstehen von «betriebswirtschaftliche[n], volkswirtschaftliche[n], und gesellschaftliche[n] Zusammenhänge[n] und Probleme[n]» (Dubs, 2001, S. 3) sowie das ganzheitliche Lösen dieser Probleme. Dubs plädiert nun dafür, dass wirtschaftsbürgerliche Bildung volkswirtschaftliche und betriebswirtschaftliche Lerninhalte umfassen muss, aber dass ersteren ein grösseres Gewicht beizumessen ist (Dubs, 2011, S. 197f.). Menschen sind, so seine Begründung, in ihrem Alltag unweigerlich mit *unternehmensbezogenen* Fragen konfrontiert, sei es als Angestellte oder als Selbständigerwerbende und brauchen für deren Beurteilung betriebswirtschaftliche Kenntnisse.

¹³ Dubs' Standard-Lehrbuch zur Volkswirtschaftslehre an allgemeinbildenden Schulen auf der Sekundarstufe II heisst denn auch «Volkswirtschaftslehre: Wirtschaftsbürgerkunde für Mittelschulen und zum Selbststudium», erstmals erschienen 1967, letztmals aufgelegt 1998 (vgl. Dubs, 1998). In seinem Vorschlag für ein Kompetenzmodell von 2004 verwendet Dubs die Bezeichnungen Volkswirtschaftslehre und Wirtschaftsbürgerkunde synonym (Dubs, 2004).

Die Vermutung liegt nahe, dass die von Dubs vorgenommene inhaltliche „Verbreiterung“ der Konzeption der wirtschaftsbürgerlichen Bildung bildungspolitisch motiviert war, nämlich um ein obligatorisches Fach „Wirtschaft & Recht“ an den Schweizer Gymnasien zu legitimieren.¹⁴ Dubs (2001, 2011) legt in seinen späteren Aufsätzen nämlich «[die] Rechtfertigung einer Bildung des allgemeinen Wirtschafts- und Gesellschaftsverständnisses und ihre normative Grundlegung» (Dubs, 2001, S. 3ff.; 2011, S. 194ff.) dar. Diese ist m.E. mehr ideologisch gefärbt als theoretisch fundiert, weil sie, gekleidet in eine polemische Sprache, ein „Weltuntergangs-Szenario“ ohne ökonomische Bildung beschreibt.¹⁵ Eberle (2006b, S. 18ff.) hingegen argumentiert „bildungstheoretisch“ und m.E. überzeugend, indem er die obersten Bildungsziele der gymnasialen Maturitätsschulen anführt, nämlich die «allgemeine Studierfähigkeit» und «anspruchsvolle Aufgaben in der Gesellschaft» (Art. 5 Abs. 1 MAR). Zum einen habe das Fach „Wirtschaft und Recht“ eine wissenschaftspropädeutische Funktion: Die Forschungstraditionen der Wirtschaft- und Rechtswissenschaften heben sich deutlich von den natur- oder geisteswissenschaftlichen Traditionen ab, so dass ein Transfer zwischen den Disziplinen nur bedingt bis gar nicht möglich ist. Zum anderen leiste das Fach „Wirtschaft & Recht“ einen Beitrag zur vertieften Gesellschaftsreife: Gesellschaftliche und wirtschaftliche Problemstellungen sind bedeutsame und «anspruchsvolle Aufgaben in der Gesellschaft», zu deren Lösung ökonomische und juristische Sachkenntnisse und Problemlösemethoden erforderlich sind.

Die Konzeption der wirtschaftsbürgerlichen Bildung, insbesondere auch die Beschreibung ihres Inhaltsbereichs beeinflusst die Entwicklung von Lehrplänen und die Gestaltung des Unterrichts (Dubs, 1985, S. 21). Dubs plädiert für einen *systematischen, disziplinenorientierten* Aufbau des Lehrplans und für eine *problemorientierte*¹⁶ Umsetzung des Lehrplans im Unterricht (Dubs, 2011, S. 197ff.; 2013, S. 19ff.). Er nennt dies «Lehrplan der Inselbildung»: Entlang der wissenschaftlichen Systematik sollen mithilfe aktueller Problemstellungen *Wissensstrukturen* aufgebaut und erweitert werden. Im fortgeschrittenen Unterricht sollen die Problemstellungen insbesondere auch interdisziplinär angegangen, d.h. aus wirtschaftlicher, politischer und soziologischer Perspektive beleuchtet werden. Es soll also ein strukturiertes und zusammenhängendes Orientierungswissen

¹⁴ In der Schweiz wurde im Zuge der Maturitätsreform 1995 das obligatorische Fach „Einführung in Wirtschaft und Recht“ an den gymnasialen Maturitätsschulen eingeführt (Art. 9 Abs. 5^{bis} Maturitäts-Anerkennungsreglement). Im Lehrplan des Kantons St.Gallen sind die Lerninhalte für das Einführungsfach „Wirtschaft und Recht“ wie folgt gegliedert: Einführung, Betriebswirtschaftslehre, Recht, Volkswirtschaftslehre (St.Gallen, 2006, S. 81ff.). Mit der Maturitätsreform 1995 wurden zudem die bisherigen Maturitätstypen abgeschafft, dafür ein breiter Kanon an Schwerpunktfächern und Ergänzungsfächern eingeführt, darunter auch „Wirtschaft und Recht“ (Art. 9 Abs. 3 und Abs. 4 Maturitäts-Anerkennungsreglement). Das Schwerpunktfach „Wirtschaft und Recht“ entspricht weitgehend dem vormaligen Typus E (Wirtschaft), der mit der Maturitätsreform 1972 eingeführt wurde (Eberle & Brüggemann, 2013, S. 9).

¹⁵ Das Szenario gestaltet sich wie folgt: Das Geschehen in Wirtschaft und Gesellschaft ist abstrakt und komplex → die Menschen verstehen „es“ nicht → sie fühlen sich verunsichert, unverstanden, ohnmächtig und/oder sind gleichgültig → die Polarisierung der Gesellschaft nimmt zu → Machterhaltung und Machtentfaltung greift um sich → die Demokratie ist gefährdet! (frei nach Dubs, 2011, S. 195).

¹⁶ Zur Gestaltung und Wirkung des problem- und handlungsorientierten Wirtschaftsunterrichts gibt es in der Schweiz bereits einige Studien (vgl. Dörig, 2003; Eberle, 2006a; Eberle et al., 2009).

sowie spezifische Problemlösemethoden vermitteln werden (vgl. auch Dubs, 1997; Eberle, 2006b). Ein breites, strukturiertes Orientierungswissen sei Voraussetzung dafür, wirtschaftliche Problemstellungen als solche zu *erkennen*, über Lösungsmöglichkeiten zu *reflektieren* sowie innovative und kreative Problemlösungen zu *entwerfen* (Dubs, 2011, S. 198f.). Da wirtschaftliche Problemstellungen nicht isoliert zu betrachten, sondern stets in einen grösseren Zusammenhang zu stellen seien, eigne sich insbesondere der exemplarische Unterricht (vgl. Wagenschein, 1973) nicht für den Wirtschaftsunterricht (Dubs, 2011, S. 199f.). Dubs (2011, S. 201f.) betont zudem, dass der Unterricht nicht wertneutral erfolgen dürfe. Die Lernenden sollen sich offen und differenziert mit Wertfragen auseinandersetzen, so z.B. durch moralische Dilemmata (vgl. Oser, 1986).

Abschliessend lässt sich festhalten, dass in der Dubs'schen Konzeption der wirtschaftsbürgerlichen Bildung alle drei curriculare Prinzipien zur Geltung kommen (vgl. Kapitel 2.1.2): das Wissenschaftsprinzip durch die wissenschaftliche Systematik des Lehrplans mit Volkswirtschaftslehre und Betriebswirtschaftslehre, das Situationsprinzip durch Anwendung der Wissensstrukturen auf wirtschaftliche und gesellschaftliche Problemstellungen sowie das Persönlichkeitsprinzip durch die Förderung der Urteilsfähigkeit und Entscheidungsfähigkeit.

2.2.3 Bildungsstandards für das Fach „Wirtschaft & Recht“ (Dubs, 2004, 2007)

Für die wirtschaftsbürgerliche Bildung bzw. für das Fach „Wirtschaft & Recht“ an gymnasialen und beruflichen Schulen gibt es bis anhin kein Kompetenzmodell und keine Bildungsstandards (Dubs, 2007, S. 306f.). Dazu müssen vorgängig eine Reihe offener Fragen beantwortet werden. 1) Soll das Kompetenzmodell disziplinenbezogen für jeden Teilbereich (BWL, VWL, Recht) oder disziplinenintegriert für den gesamten Lernbereich „Wirtschaft & Recht“ entwickelt werden? Dubs (2007, S. 306) plädiert für disziplinenbezogene Kompetenzmodelle, denn disziplinenbezogene Kompetenzen seien eine minimale Voraussetzung für einen erfolgreichen integrativen Unterricht. 2) Wie soll das notwendige breite Orientierungswissen für das Fach „Wirtschaft & Recht“ mit Kompetenzen verbunden werden, die dem Kompetenzbegriff gerecht werden? 3) Welches Bildungsziel («Gestaltungsidee») für das Fach „Wirtschaft & Recht“ soll dem Kerncurriculum und dem Kompetenzmodell zugrunde gelegt werden? Dubs (1997, 2004) identifiziert drei mögliche Gestaltungsideen: die wissenschaftlich-exemplarische, die wirtschaftsgeografische und die wirtschaftsbürgerliche. Er plädiert für die wirtschaftsbürgerliche Gestaltungsidee und begründet sie mit den komplexen gesamtwirtschaftlichen und wirtschaftspolitischen Problemstellungen (Dubs, 2004, S. 45).¹⁷ Die dritte Frage blieb von der Wirtschaftsdidaktik in der Schweiz bis heute unbeantwortet (vgl. Dubs, 1997, 2004, 2007).

¹⁷ Zur wirtschaftsbürgerlichen Gestaltungsidee gelangt Dubs durch «eine kurze Analyse der fachdidaktischen Literatur sowie eine Beurteilung des Bildungszieles durch interessierte Kreise» (Dubs, 2004, S. 44).

Für das Fach „Wirtschaft & Recht“ bzw. für den Teilbereich Volkswirtschaftslehre entwickelte Dubs (2004) das erste – und bisher einzige – Kompetenzmodell nach den Merkmalen guter Bildungsstandards (vgl. Klieme, Avenarius, et al., 2007, S. 24ff.). Mit diesem Kompetenzmodell wollte er die Diskussion zu Bildungsstandards in den Wirtschaftsfächern anregen, aber es gibt m.W. noch heute keine Bildungsstandards für die Fächer „Wirtschaft & Recht“ bzw. „Wirtschaft & Gesellschaft“ auf der Sekundarstufe II. Allerdings hat er das Kompetenzmodell nicht in eine Aufgabensammlung umgesetzt und empirisch geprüft.

Als allgemeines Bildungsziel für die Volkswirtschaftslehre identifiziert Dubs (2004, S. 45) die „wirtschaftsbürgerliche Gestaltungsidee“: Junge Menschen sollen befähigt werden, das Geschehen in der Volkswirtschaft zu verstehen und wirtschaftspolitische Probleme und Streitfragen in sachkompetenter Weise eine eigene Meinung bilden. Die Teilkompetenzen¹⁸ des Kompetenzmodells wurden durch eine empirische Medienanalyse wirtschaftlicher und wirtschaftspolitischer Beiträge in Presse, Radio und Fernsehen ermittelt und zunächst nach der kognitiven Taxonomie (vgl. Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956) geordnet (Dubs, 2004, S. 47). Für die Stufung der Teilkompetenzen wurden zwei Aspekte einbezogen (Dubs, 2004, S. 48): zum einen das Anspruchsniveau (Komplexität) einer Aufgaben-/Problemstellung (richtige Lösung einfach bestimmbar, richtige Lösung unter bestimmten Voraussetzungen bestimmbar, keine eindeutige Lösung bestimmbar), zum anderen das verfügbar Vorwissen/Wissen für eine Aufgaben-/Problemstellung (eindeutige Grundbegriffe ausreichend, Vernetzung von eindeutigen Grundbegriffen, Vernetzung von Zusammenhängen).

2.3 Ökonomische Bildung im Schweizer Bildungssystem

Dieses Kapitel beschreibt die rechtliche Verankerung und curriculare Ausgestaltung der ökonomischen Allgemeinbildung im Schweizer Bildungssystem. Zuerst wird auf die Sekundarstufe II eingegangen, insbesondere auf die gymnasiale und die berufliche Bildung (Kapitel 2.3.1), danach auf die Sekundarstufe I (Kapitel 2.3.2). Die Darstellung zur Sekundarstufe II ist für diese Dissertation richtungsweisend. Die Darstellung zur Sekundarstufe I ist ergänzend zu verstehen, da die Sekundarschulen in den meisten Kantonen „Zubringer“ für das Gymnasium sind; so auch im Kanton St.Gallen, in dem die Datenerhebungen des Projekts WBKgym erfolgten.

¹⁸ Als solche Teilkompetenzen werden genannt (Dubs, 2004, S. 47): Probleme verstehen, Streitfragen und Patentlösungen erkennen, zu einer eigenen Meinung gelangen, sich der eigenen normativen Annahmen bewusst sein, den eigenen Standpunkt auf die Konsistenz und soweit als möglich auf die sachliche Richtigkeit hin analysieren, die Folgen der eigenen Meinung auf kurz- und langfristige gesellschaftliche Entwicklungen differenziert abschätzen.

2.3.1 Sekundarstufe II: Berufsbildung und Allgemeinbildung

Auf der Sekundarstufe II ist ökonomische Allgemeinbildung seit vielen Jahren verankert, dies sowohl an den beruflichen Schulen als auch an den allgemeinbildenden Schulen. Die curriculare Ausgestaltung ist jedoch von der Schulform bzw. vom Bildungsgang abhängig.

2.3.1.a Gymnasiale Bildung

Das Schulwesen in der Schweiz ist Sache der Kantone (Art. 62 Bundesverfassung); sie sind auch für die gymnasialen Maturitätsschulen und die Fachmittelschulen auf der Sekundarstufe II zuständig. Je nach Kanton umfasst die gymnasiale Maturitätsschule die 9. bis 12. Klasse oder die 10. bis 13. Klasse. Das „Reglement über die Anerkennung von gymnasialen Maturitätsausweisen“ (Maturitäts-Anerkennungsreglement, MAR) regelt die schweizerische Anerkennung von kantonalen und kantonal anerkannten gymnasialen Maturitätsausweisen (Art. 1 MAR).

Die obersten Ziele der gymnasialen Maturität werden wie folgt umschrieben: Gymnasiale Maturand*innen haben «jene[] persönliche[] Reife, die Voraussetzung für ein Hochschulstudium ist und die sie auf anspruchsvolle Aufgaben in der Gesellschaft vorbereitet» (Art. 5 Abs. 1 MAR). Diese obersten Ziele des Gymnasiums können mit den zwei Formeln «allgemeine Studierfähigkeit» und «vertiefte Gesellschaftsreife» zusammengefasst werden (Eberle & Brüggenschrock, 2013, S. 10; Eberle et al., 2008, S. 27). Diese beiden Ziele stehen sich auf gleicher Hierarchiestufe gegenüber und haben einen grossen gemeinsamen Überschneidungsbereich: Die meisten Lerninhalte des Gymnasiums sind sowohl für den Erwerb der allgemeinen Studierfähigkeit als auch für denjenigen der vertieften Gesellschaftsreife förderlich. Allerdings haben beide Ziele auch voneinander losgelöste Bereiche. So umfasst v.a. die vertiefte Gesellschaftsreife Kompetenzen, die über die fachbezogene und allgemeine Studierfähigkeit hinausgehen (Eberle & Brüggenschrock, 2013, S. 11).

Im „Rahmenlehrplan für die Maturitätsschulen“ (RLP-MS) werden die allgemeinen Bildungsziele der gymnasialen Maturität und der einzelnen Fächer beschrieben und begründet (EDK, 1994). Die Kantone dürfen basierend auf dem RLP-MS kantonale Lehrpläne für ihre Maturitätsschulen erlassen (Art. 8 MAR). Entsprechend variieren die konkreten Lerninhalte und Stundentafeln der Fächer je nach Kanton.¹⁹

Das Fach „Wirtschaft & Recht“ (WuR) muss als obligatorische Einführung²⁰ besucht werden (Art. 9 Abs. 5^{bis} MAR) und kann als Schwerpunkt (Art. 9 Abs. 3 lit. e MAR) oder Ergänzung (Art. 9 Abs.

¹⁹ Im Kanton St.Gallen beispielsweise findet das obligatorische Einführungsfach WuR im 10. und 11. Schuljahr mit insgesamt 4 Jahreswochenstunden (2 Jahreswochenstunden pro Schuljahr) statt, das Schwerpunktfach vom 9. bis 12. Schuljahr mit insgesamt 15 Jahreswochenstunden (3 bzw. 4 Jahreswochenstunden pro Schuljahr) (St.Gallen, 2006, S. 10). Im Kanton Bern hat das Einführungsfach WuR im 10. Schuljahr insgesamt 2 Jahreswochenstunden, das Schwerpunktfach im 10. und 12. Schuljahr insgesamt 14 Jahreswochenstunden (4 bzw. 5 Jahreswochenstunden pro Schuljahr) (Bern, 2005, S. 85).

²⁰ Die Bezeichnung des Faches als „Einführung in Wirtschaft & Recht“ ist einerseits als Abgrenzung zum Schwerpunkt- und Ergänzungsfach zu verstehen, andererseits als Hinweis auf seine Neuartigkeit, denn diese Lerninhalte werden auf den

4 lit. i MAR) gewählt werden. Das Schwerpunktfach WuR umfasst je nach Kanton 12 bis 18 Jahreswochenstunden, das Ergänzungsfach 3 bis 6 Jahreswochenstunden, das Einführungsfach 1 bis 5 Jahreswochenstunden. Das Fach „Wirtschaft & Recht“ ist fachwissenschaftlich strukturiert und umfasst die drei Teilbereiche Volkswirtschaftslehre (VWL), Betriebswirtschaftslehre (BWL) und Rechtslehre (Recht). BWL und VWL haben Knappheit zum Erkenntnisgegenstand, die Rechtslehre die Gerechtigkeit. Diese Teilbereiche sollen integrativ unterrichtet werden (EDK, 1994, S. 76).

Die allgemeinen Bildungsziele des Faches „Wirtschaft und Recht“ werden wie folgt beschrieben:

«Durch den gymnasialen Unterricht in Wirtschaftswissenschaften sollen die Schülerinnen und Schüler erkennen, dass alles Wirtschaften in Knappheitssituationen erfolgt. Die Jugendlichen werden befähigt, wirtschaftliche und rechtliche Zustände und Prozesse in einem Gesellschaftssystem wahrzunehmen und sich der Wertungen bewusst zu werden, die in jeder Gesellschaftsanalyse enthalten sind. Die Schülerinnen und Schüler werden befähigt, Ziele, Strukturen, Prozesse und Interdependenzen in wirtschaftlichen Systemen zu beurteilen, um dadurch die Gestaltungsmöglichkeiten wirtschaftlichen und politischen Handelns zu erkennen. Die Jugendlichen erkennen den Widerspruch zwischen individueller und kollektiver, kurz- und langfristiger Zielsetzung in der Wirtschaft. Sie gewichten sie nach fachspezifischen und ethischen Prinzipien, um so ihrer menschlichen und staatsbürgerlichen Verantwortung im Alltag zu genügen» (EDK, 1994, S. 75).

An den gymnasialen Maturitätsschulen der Schweiz wird ökonomische Allgemeinbildung also vorrangig im Fach „Wirtschaft & Recht“ (WuR) vermittelt. Interdisziplinäre Anknüpfungspunkte ergeben sich mit den Fächern „Geschichte“ und „Geografie“, die beide als Grundlagenfach besucht werden müssen (Art. 9 Abs. 2 lit. h und lit. i MAR) und als Ergänzungsfach gewählt werden können (Art. 9 Abs. 4 lit. e und lit. f MAR). Das Fach Geschichte umfasst u.a. Wirtschafts- und Sozialgeschichte: «ökonomische und soziale Mechanismen und deren Veränderbarkeit» (EDK, 1994, S. 71). Das Fach Geografie behandelt u.a. Wirtschaftsgeografie «Zusammenwirken und gegenseitige Beeinflussung von Mensch und Natur» (EDK, 1994, S. 117).

2.3.1.b Berufliche Bildung

Die Berufsbildung in der Schweiz ist Sache des Bundes (Art. 63 Bundesverfassung); er erlässt Vorschriften über die Berufsbildung und fördert ein breites und durchlässiges Angebot an Berufsbildung. Das Bundesgesetz über die Berufsbildung (Berufsbildungsgesetz, BBG) regelt sämtliche Berufsbereiche ausserhalb der Hochschulen, insbesondere auch die berufliche Grundbildung und die Berufsmaturität (Art. 2 BBG). Für die Berufsbildung sind Bund, Kantone und Organisationen der

unteren Schulstufen nicht unterrichtet (Eberle, 2006b, S. 21). Diese Neuartigkeit hält jedoch mit der Einführung des Lehrplans 21 und mit dem Übertritt der ersten Kohorten ans Gymnasium nicht mehr stand.

Arbeitswelt (Sozialpartner, Berufsverbände etc.) gemeinsam zuständig (Art. 1 Abs. 1 BBG). Die berufliche Grundbildung dient der Vermittlung und dem Erwerb von Qualifikationen (Fähigkeiten, Kenntnisse, Fertigkeiten), die für die Berufstätigkeit erforderlich sind (Art. 15 Abs. 1 BBG). Die berufliche Grundbildung dauert entweder zwei Jahre und führt zum eidgenössischen Berufsattest (EBA) oder drei bis vier Jahre und führt zum eidgenössischen Fähigkeitszeugnis (EFZ) (Art. 17 Abs. 3 BBG). Die Berufsmaturität umfasst eine berufliche Grundbildung und ergänzt sie um eine erweiterte Allgemeinbildung (Art. 2 Berufsmaturitätsverordnung).

Der „Rahmenlehrplan für den allgemeinbildenden Unterricht in der beruflichen Grundbildung“ (RLP-ABU) gilt für alle Berufe, welche die Allgemeinbildung nicht in einer eigenen Bildungsverordnung (BiVo) regeln (SBFI, 2006, S. 4). Dies sind insbesondere gewerbliche, technische, soziale und künstlerische Berufe, nicht aber kaufmännische. Der allgemeinbildende Unterricht (ABU) erfolgt themen- und handlungsorientiert: «Themenorientiert heisst, dass die Inhalte des Unterrichts in Form von Themen organisiert sind und nicht einer disziplinären Fachlogik folgen. Die Themen nehmen Bezug auf die persönliche, berufliche und gesellschaftliche Realität der Lernenden. Handlungsorientiert heisst, dass die Lernenden ihre Kompetenzen durch eigenes Handeln weiterentwickeln» (SBFI, 2006, S. 5). Der ABU gliedert sich in die zwei Lernbereiche „Sprache und Kommunikation“ sowie „Gesellschaft“, deren Bildungsziele gleichermassen bedeutsam sind (SBFI, 2006, S. 5f.).

Der Lernbereich „Gesellschaft“ umfasst acht Aspekte, u.a. Kultur, Ökologie, Politik, Recht, Technologie und Wirtschaft. Für den Aspekt Wirtschaft werden vier *Bildungsziele* genannt (SBFI, 2006, S. 19): (A) Verantwortungsbewusst konsumieren, (B) Im Zusammenspiel der unterschiedlichen Akteure eigene Rollen analysieren, (C) Das eigene Unternehmen als Produzent und Anbieter in der Gesamtwirtschaft verstehen und (D) Wirtschaftliche Prinzipien und Entwicklungen beurteilen. Der Leitgedanke des Aspekts Wirtschaft beschreibt die wirtschaftlichen Kompetenzen, die im Unterricht zu fördern und von den Lernenden zu entwickeln sind:

«Wirtschaftliche Kompetenz wird auf der Analyse von wirtschaftlichen Vorgängen aufgebaut, welche die Lernenden befähigt, in ihrem persönlichen, beruflichen und gesellschaftlichen Umfeld bewusst und adäquat zu handeln. In der Auseinandersetzung mit individuellen, unternehmerischen, nationalen oder internationalen Wirtschaftsfragen, erweitern die Lernenden ihren Sachverstand. Dadurch vermögen sie persönliche Gestaltungsmöglichkeiten, kontroverse gesellschaftliche Fragen sowie globale Herausforderungen auf der Basis von wirtschaftlichen Argumenten einzuordnen und zu beurteilen» (SBFI, 2006, S. 18).

Der „Bildungsplan Kauffrau/Kaufmann EFZ“ regelt den betrieblichen und schulischen Teil der kaufmännischen Berufsbildung (SKKAB, 2011a). Er umschreibt berufliche Handlungskompetenzen zur Bewältigung beruflicher Anforderungen, die aus «einem spezifischen Bündel von Fachkompetenzen,

Methodenkompetenzen, Sozial- und Selbstkompetenzen – jeweils bestimmt durch Anforderungen einer Situation, Aufgabe oder Problemstellung» (SKKAB, 2011a, S. 7) bestehen und sowohl in der betrieblichen als auch in der schulischen Ausbildung gefordert werden. Die beruflichen Fachkompetenzen sind in fünf Lernbereiche gegliedert (SKKAB, 2011a, S. 8ff.): 1) „Branche und Betrieb“, 2) „Standardsprache (regionale Landessprache)“, 3) „Fremdsprachen (zweite Landessprache oder Englisch)“, 4) „Information/Kommunikation/Administration (IKA)“ sowie 5) „Wirtschaft und Gesellschaft (W&G)“.

Der Lernbereich W&G gliedert sich in vier Teilbereiche (SKKAB, 2011a, S. 10; 2011b): „finanzwirtschaftliche Zusammenhänge“, „betriebswirtschaftliche Zusammenhänge“, „Recht und Staat“ sowie „gesamtwirtschaftliche und gesamtgesellschaftliche Zusammenhänge“. Der Lernbereich W&G hat also eine fachwissenschaftliche Struktur – im Gegensatz zum Lernbereich „Branche und Betrieb“, der nach Lernfeldern strukturiert ist (SKKAB, 2011a, S. 8f.). Der Lernbereich W&G hat eine berufsbildende und eine allgemeinbildende Funktion: Die berufsbildende Funktion zeigt sich v.a. in den Teilbereichen Finanzwirtschaft und Betriebswirtschaft (SKKAB, 2011a, S. 7), die allgemeinbildende v.a. in den Teilbereichen Recht/Staat und Gesamtwirtschaft/Gesamtgesellschaft. Die Fachkompetenzen für W&G werden im „Leistungszielkatalog W&G“ konkretisiert: Es werden für jeden Teilbereich Leistungsziele mit Inhaltselementen und Kognitionsstufen operationalisiert sowie für jedes Leistungsziel eine Stundendotation vorgegeben (SKKAB, 2011b). Der Lernbereich W&G umfasst im E-Profil 13 Jahreswochenstunden bzw. total 520 Stunden, im B-Profil 10 Jahreswochenstunden bzw. total 400 Stunden.

Das Bildungsziel für W&G wird wie folgt beschrieben und begründet:

«Um Grundzusammenhänge, Probleme und Herausforderungen in Unternehmungen, in Wirtschaft und Gesellschaft zu verstehen, brauchen Kaufleute ein fundiertes Wissen und grundlegende Einsichten. In den Bereichen finanzwirtschaftliche Zusammenhänge und Buchhaltung, Betriebs- und Rechtskunde sind sie fähig, Probleme zu erkennen und in ihrem Gestaltungsbereich Lösungen vorzuschlagen, umzusetzen oder zu beurteilen. Sie erklären die wesentlichen Zusammenhänge in der Gesamtwirtschaft und sind sich ihrer Verantwortung und Möglichkeiten als Wirtschafts- und Gesellschaftsbürger bewusst» (SKKAB, 2011b, S. 1).

Der „Rahmenlehrplan für die Berufsmaturität“ (RLP-BM) legt die Fächer der Berufsmaturität für den Grundlagenbereich, den Schwerpunktbereich und den Ergänzungsbereich fest (SBFI, 2012, S. 17ff., S. 53ff., S. 110ff.). Im Grundlagenbereich sind dies „Landessprachen“, „Englisch“ und „Mathematik“; im Schwerpunktbereich sind es „Finanz- und Rechnungswesen“, „Gestaltung/Kunst/Kultur“, „Information und Kommunikation“, „Mathematik“, „Naturwissenschaften“, „Sozialwissenschaften“ sowie „Wirtschaft und Recht“; im Ergänzungsbereich „Geschichte und Politik“, „Technik und Umwelt“ sowie „Wirtschaft und Recht“. Der RLP-BM

beschreibt zudem die allgemeinen Bildungsziele für jedes Fach sowie für jedes Fach Lerngebiete und fachliche Kompetenzen. Das Schwerpunktfach und das Ergänzungsfach „Wirtschaft und Recht“ (WuR) sind in drei Lerngebiete gegliedert, nämlich unternehmerische, volkswirtschaftliche und rechtliche Aspekte (SBFI, 2012, S. 99ff., S. 118ff.). Allerdings unterscheiden sich die beiden Fächer im Umfang und in der Differenziertheit der jeweiligen Aspekte. Das Schwerpunktfach WuR umfasst 300 Stunden, das Ergänzungsfach 120. An dieser Stelle sei hervorgehoben, dass es in der Berufsmaturität kein Grundlagenfach „Wirtschaft und Recht“ gibt – im Gegensatz zur gymnasialen Maturität, wo es ein obligatorisches Einführungsfach „Wirtschaft und Recht“ gibt (vgl. Kapitel 2.3.1.a).

Die allgemeinen Bildungsziele des Schwerpunktfaches „Wirtschaft und Recht“ werden wie folgt beschrieben:

«Wirtschaft und Recht im Schwerpunktbereich hilft den Lernenden, sich in der bestehenden, wandelbaren Wirtschafts- und Rechtsordnung als Mitarbeitende in Unternehmungen und Organisationen, als Familienmitglieder, Konsumentinnen/Konsumenten und Staatsbürgerinnen/Staatsbürger zurechtzufinden sowie einen Beitrag zur weiteren nachhaltigen Entwicklung dieser Ordnung zu leisten. Dieses Ziel wird erreicht, wenn die Lernenden zu einer eigenen, sachlich fundierten und normativ begründeten Meinung bei konkreten betriebs- und volkswirtschaftlichen sowie rechtlichen Problemstellungen gelangen (mündige Wirtschafts- und Rechtsbürgerinnen/Rechtsbürger), eigene Lösungsideen einbringen und von Expertinnen/Experten vorgeschlagene Lösungen einschätzen können. Voraussetzung dazu ist, dass die Lernenden wesentliche betriebswirtschaftliche, volkswirtschaftliche und rechtliche Strukturen und Prozesse verstehen sowie normative Grundlagen von Entscheidungen durchschauen. Dazu eignen sie sich ein ökonomisches und rechtliches Grundlagenwissen an, welches sie auf konkrete Fälle anwenden» (SBFI, 2012, S. 99).

2.3.2 Sekundarstufe I: Grundbildung

Das Schulwesen in der Schweiz ist Sache der Kantone (Art. 62 Bundesverfassung); sie organisieren einen obligatorischen Grundschulunterricht, der allen Kindern offensteht. Der Lehrplan 21 ist der erste gemeinsame Lehrplan der deutschsprachigen Kantone für die obligatorische Grundbildung, die sog. Volksschule (D-EDK, 2014).²¹ Er stützt sich auf den Harmonisierungsauftrag in der Bundesverfassung (Art. 62 Abs. 4 Bundesverfassung) und die „Interkantonale Vereinbarung über die Harmonisierung der obligatorischen Schule“ (HarmoS-Konkordat). Das HarmoS-Konkordat regelt u.a. das Schuleintrittsalter, die Dauer und die Ziele der Schulstufen. Der Lehrplan 21 wurde im Jahr 2014 von der Deutschschweizer Erziehungsdirektoren-Konferenz (D-EDK) verabschiedet und wird ab dem Schuljahr 2017/2018 dezentral in den Kantonen eingeführt.

²¹ Das Pendant zum Lehrplan 21 für die französisch- und italienischsprachigen Kantone ist der Plan d'études romand (PER) (CIIP, 2010).

Der Lehrplan 21 definiert elf Schuljahre und unterteilt sie in drei Zyklen (D-EDK, 2014, S. 3): Der 1. Zyklus umfasst zwei Jahre Kindergarten und die ersten zwei Jahre der Primarstufe (bis Ende 2. Klasse), der 2. Zyklus die weiteren vier Jahre Primarstufe (3. bis 6. Klasse), der 3. Zyklus die drei Jahre der Sekundarstufe I (7. bis 9. Klasse). Der Lehrplan 21 gliedert sich inhaltlich in sechs Fachbereiche (D-EDK, 2014, S. 3f.): „Sprachen“ (Deutsch, Französisch, Englisch, Italienisch), „Mathematik“, „Natur, Mensch, Gesellschaft (NMG)“, „Gestalten“ (bildnerisch, textiles und technisches Gestalten), „Musik“ sowie „Bewegung und Sport“. Für jeden Fachbereich werden Kompetenzen beschrieben, welche die Schüler*innen im Laufe der Schuljahre erwerben sollen. Zusätzlich zu den Fachbereichen werden zwei Modulbereiche vorgegeben: „Medien und Informatik“ sowie „Berufliche Orientierung“.

Im Fachbereich „Natur, Mensch, Gesellschaft (NMG)“ bzw. in dessen Subfachbereich „Wirtschaft, Arbeit, Haushalt“ (WAH) finden sich vier Kompetenzbereiche, die ökonomische Wissens- und Handlungsaspekte widerspiegeln (D-EDK, 2014, S. 82ff.): „Produktions- und Arbeitswelten erkunden“ (WAH.1), „Märkte und Handel verstehen“ (WAH.2), „Konsum gestalten“ (WAH.3) sowie „Haushalten und Zusammenleben gestalten“ (WAH.5). Diese Kompetenzen werden im 3. Zyklus, genauer in der 8. und 9. Klasse, erworben und sind mit total 6 Jahreswochenstunden dotiert.

Vor Einführung des Lehrplans 21 enthielten nur wenige kantonale Lehrpläne auf der Sekundarstufe I ökonomische Lerninhalte. Diese ökonomischen Lerninhalte waren in andere Fächer integriert, so z.B. in „Mathematik“ (Geld und Geldeinheiten, Prozentrechnen, Brutto- und Nettopreise, Rabatt, Kapital und Zinsen, Konsumkredit, Leasing und Ratenzahlung, Fremdwährungen und Wechselkurse) und in „Mensch und Umwelt“ (Taschengeld, Lohn, Budget, Konsum) (vgl. Ackermann & Eberle, 2016). Die Schüler*innen wechselten also i.d.R. ohne curriculares Vorwissen von der Sekundarstufe I auf die Sekundarstufe II.

Inwiefern diese neuen Kompetenzbereiche WAH die ökonomische Kompetenz von Schüler*innen auf der Sekundarstufe I fördern und wie sich dieses curriculare Vorwissen auf die ökonomische Kompetenz ebendieser Schüler*innen auf der Sekundarstufe II auswirkt, ist eine derzeit noch ungeklärte, wenn auch nicht minder spannende Forschungsfrage!

2.4 Zwischenfazit

Fachdidaktische Konzeptionen der ökonomischen Bildung legen normative Leitbilder und Bildungsziele fest, die insbesondere bei der Auswahl und Gewichtung der Lerninhalte sowie bei der Formulierung von Lernzielen und Kompetenzbereichen zum Tragen kommen. Ökonomische Bildung bzw. ökonomische Allgemeinbildung zielt darauf ab, junge Menschen zur Bewältigung *ökonomisch geprägter Lebenssituationen* zu befähigen. Das Leitbild der ökonomischen Bildung ist «der mündige Wirtschaftsbürger», der diese Anforderungssituationen tüchtig (sachkundig), selbstbestimmt und verantwortungsbewusst bewältigt. Diese Lebenssituationen stellen sich im *privaten, beruflichen und gesellschaftlichen* Bereich (Albers, 1988, 1995) bzw. in den drei Situationsfeldern *Konsum, Beruf und*

Arbeit sowie *Wirtschaftsgesellschaft* (Kaminski, 1996; Kaminski et al., 2017). Sie korrespondieren mit unterschiedlichen Rollen des Menschen – Verbraucher*in, Erwerbstätige*r und Staatsbürger*in – in seiner Lebenswelt.

Kompetenzmodelle für ökonomische Bildung verwenden entweder das Referenzsystem „Gegenstandsbereich Wirtschaft“ mit Wirtschaftsordnung und Wirtschaftskreislauf (Kaminski et al., 2017; Kaminski et al., 2008) oder das Referenzsystem „ökonomische Perspektive“ mit ökonomisch geprägten Lebenssituationen und wissenschaftlichen Kategorien als Denk- und Analyseschemata (Retzmann et al., 2010; G. Seeber et al., 2012; B. Weber, 2005, 2013). Je nach Referenzsystem werden die Kompetenzbereiche sehr unterschiedlich definiert. Es besteht mitunter in der Forschungsgemeinschaft kein Konsens darüber, welche Kompetenzbereiche für die ökonomische Bildung ausschlagend sind. Synoptisch werden jedoch folgende Fähigkeiten als massgeblich erachtet: Verstehen, Reflektieren, Urteilen, Entscheiden und Handeln (Albers, 1995; Dubs, 2004, 2011, 2013; Kaminski et al., 2017; Kaminski et al., 2008; G. Seeber et al., 2012; B. Weber, 2005, 2014). Zudem sind diese bestehenden Kompetenzmodelle für ökonomische Bildung noch keiner empirischen Überprüfung unterzogen worden und nur für einige wenige sind exemplarische Aufgaben entwickelt worden (vgl. Retzmann et al., 2010, S. 38ff.).

Aus den skizzierten fachdidaktischen Konzeptionen der ökonomischen Bildung und der wirtschaftsbürgerlichen Bildung lässt sich nun ableiten, a) in welcher *Domäne* «wirtschaftsbürgerliche» Anforderungssituationen zu verorten sind, b) welche *Fähigkeiten* sie erfordern und c) welcher *Wissenschaftsbezug* herzustellen ist.

Die *Domäne* der «wirtschaftsbürgerlichen» Anforderungssituationen ist nach der „ökonomischen Perspektive“ der persönliche²², berufliche und gesellschaftliche Lebensbereich (vgl. Albers, 1988; Albers, 1995) bzw. die drei Situationsfelder *Konsum*, *Beruf* und *Arbeit* sowie *Wirtschaftsgesellschaft* (vgl. Kaminski, 1996; Kaminski et al., 2017). Diese Lebensbereiche bzw. Situationsfelder korrespondieren mit den unterschiedlichen Rollen des Menschen – Verbraucher*in, Erwerbstätige*r und Staatsbürger*in – in seiner Lebenswelt. Die erforderlichen *Fähigkeiten* für «wirtschaftsbürgerliche» Anforderungssituationen sind: das *Verstehen* und *Reflektieren* von wirtschaftlichen und gesellschaftlichen Zusammenhängen, das *Beurteilen* von entsprechenden Problemstellungen bzw. Lösungsansätzen sowie das *Lösen* dieser Problemstellungen (vgl. Dubs, 1985, 2001, 2011, 2013). Die bei «wirtschaftsbürgerlichen» Anforderungssituationen anzuwendenden

²² Von der Autorin wird die Bezeichnung „persönlicher“ statt „privater“ Lebensbereich bevorzugt, da er für die Beschreibung der Domäne treffender ist. Im rechtlichen Kontext wird nämlich „privat“ als Abgrenzung zu „staatlich“ verstanden: Die Rechtsgebiete sind in *Privatrecht* und öffentliches Recht gegliedert (Forstmoser & Vogt, 2012, S. 117ff.), juristische Personen haben eine *privatrechtliche* (z.B. Verein, Stiftung, Aktiengesellschaft) oder eine öffentlich-rechtliche Rechtspersönlichkeit (z.B. Kantonbank, Universität, Post, SBB) (Hausheer & Aebi-Müller, 2016, S. 307ff.), die verfassungsmässigen Grundrechte schützen die *Privatsphäre* des Individuums vor Eingriffen des Staates (Tschannen, 2016, S. 102ff.).

wissenschaftlichen Kategorien ergeben sich aus der Wirtschaftswissenschaft (Volkswirtschaftslehre, Betriebswirtschaftslehre) und anderen Sozialwissenschaften (z.B. Rechtswissenschaft, Politikwissenschaft (vgl. Dubs, 2011, 2014; Kaminski et al., 2017; Kaminski et al., 2008; Kruber, 2000, 2005; G. Seeber et al., 2012; A. Weber, 2005; B. Weber, 2014)).

3 Modellierung und Messung der ökonomischen Kompetenz

In diesem Kapitel wird der Leitfrage (2) nachgegangen (vgl. Kapitel 1.2):

- (2) *An welche allgemeinpädagogischen und wirtschaftspädagogischen Kompetenzkonzepte kann das Konstrukt «wirtschaftsbürgerliche» Kompetenz theoretisch angelehnt bzw. von welchen muss es abgegrenzt werden?*

Dieses Kapitel behandelt wissenschaftliche Kompetenzkonzepte (Kapitel 3.1), geläufige und gegenläufige Kompetenzdefinitionen in der Bildungsforschung (Kapitel 3.2) sowie Kompetenzmodelle für die pädagogisch-psychologische Diagnostik (Kapitel 3.3). Des Weiteren werden in diesem Kapitel bestehende Konzepte und Instrumente für die Modellierung und Messung der *ökonomischen Kompetenz* vorgestellt (Kapitel 3.4) und hinsichtlich ihrer Eignung für die wirtschaftsbürgerliche Kompetenz diskutiert.

3.1 Kompetenzkonzepte

Der Begriff Kompetenz ist in den letzten Dekaden *en vogue* geworden, sowohl in der breiten Öffentlichkeit wie auch in vielen Sozialwissenschaften. Die Bedeutung des Begriffs Kompetenz (lat. *competentia*: Zusammentreffen) ist in der Alltagssprache sehr vielfältig (Duden, 2019): von „Zuständigkeit“ im betrieblich-organisatorischen und öffentlich-rechtlichen Sinne bis zu „Fähigkeit“ im pädagogisch-psychologischen Sinne.²³ Auch in der Wissenschaftssprache hat der Begriff Kompetenz verschiedene Bedeutungen und wird nicht einheitlich verwendet (Hartig, 2008, S. 16). Es existiert eine Vielzahl von Konzepten und Definitionen; diese Tatsache bezeichnete Weinert (2001a, S. 45) nicht unkritisch als «konzeptuelle Inflation». Das wissenschaftliche Bestreben nach einer Kompetenzdefinition erklärt Norris (1991) wie folgt:

«Despite their practicality and apparent simplicity the concepts of competence and standards have had a troubled history. As tacit understandings of the words have been overtaken by the need to define precisely and operationalise concepts, the practical has become shrouded in theoretical confusion and the apparently simple has become profoundly complicated» (Norris, 1991, S. 331f.).

Norris (1991, S. 332ff.) identifiziert drei zentrale Forschungstraditionen, in die sich die vielen verschiedenen Kompetenzdefinitionen einordnen lassen: behavioristische Konzeptionen (*behaviourist constructs*), generische Konzeptionen (*generic construct*) und kognitive Konzeptionen (*cognitive constructs*).

²³ Der Duden nennt u.a. folgende Synonyme: Fähigkeit, Fertigkeit, Qualifikation, Sachverstand, Begabung, Befugnis, Zuständigkeit (Duden, 2019).

Behavioristische Konzeptionen der Kompetenz beschreiben das *Verhalten*, auch bezeichnet als Performanz, und die *Situationen*, in denen es stattfindet. Kompetenz ist also verhaltensbezogen und outcome-orientiert. Sie wird aufgefasst als etwas, was eine Person tut oder fähig sein sollte zu tun. Bei den behavioristischen Konzeptionen der Kompetenz geht es weniger um die Verfügbarkeit von Wissen und Können, sondern mehr um die Anwendung des notwendigen, verfügbaren Wissens und Könnens in einer spezifischen Situation für ein zielbezogenes Verhalten.

Behavioristische Konzeptionen sind v.a. in der Arbeits- und Organisationspsychologie verbreitet, um die besten Kandidat*innen für einen Job anhand einer Sammlung von typischen beruflichen Anforderungssituationen zu selektieren (Blömeke, Gustafsson, & Shavelson, 2015, S. 5). Kompetenz wird in Form von *verhaltensbezogenen Zielen* beschrieben, die der Vorführung, Beobachtung und Bewertung zugänglich sind (Norris, 1991, S. 332). Verhaltensbezogene Ziele müssen eine Reihe von Bedingungen erfüllen (Gagné & Briggs, 1974, S. 79): (1) die Handlung der Person und die Situation, in der die Handlung ausgeübt wird, präzise und transparent beschreiben; (2) die Grenzen der Situation für das zu beobachtende Verhalten benennen; (3) die Fähigkeiten (Art, Umfang, Ausprägung), auf die durch das beobachtete Verhalten geschlossen werden kann, angeben.

Generische Konzeptionen der Kompetenz beschreiben «broad clusters of abilities which are conceptually linked» (Elliott, 1989, S. 98, zit. nach Norris, 1991, S. 332), also ein System vielfältiger allgemeiner Fähigkeiten als Anforderungen für eine berufliche Tätigkeit (Norris, 1991, S. 332f.). Die beschriebene Kompetenz orientiert sich an überdurchschnittlichen Personen (*expert performers*), um sie von durchschnittlichen Personen (*average performers*) zu unterscheiden. Die Beschreibung der Kompetenz wird über empirische Untersuchungen in einem dreischrittigen Verfahren gewonnen: (1) erfolgreiche Personen einer beruflichen Tätigkeit identifizieren, (2) ihre Aktivitäten und Erfolgsfaktoren studieren, z.B. mittels *behavioural event interviews* oder *critical incident technique*, (3) die Fähigkeiten und Fertigkeiten identifizieren, die den Unterschied zwischen erfolgreichen und weniger erfolgreichen Personen ausmacht. Mit diesem Vorgehen kann für eine bestimmte berufliche Tätigkeit eine Liste allgemeiner Kompetenzen generiert werden. Nichtsdestotrotz sind solche allgemeinen Kompetenzen für eine Beobachtung und Bewertung schwierig zuträglich und sie sind i.d.R. auch auf andere berufliche Tätigkeiten anwendbar.

Das Konzept der *Schlüsselqualifikationen* (vgl. Mertens, 1974; Zabeck, 1991) aus dem deutschsprachigen Berufsforschung kann zu den generischen Ansätzen gezählt werden. «Schlüsselqualifikationen sind [...] Kenntnisse, Fähigkeiten und Fertigkeiten, welche nicht unmittelbaren und begrenzten Bezug zu bestimmten, disparaten praktischen Tätigkeiten erbringen, sondern vielmehr a) die Eignung für eine grosse Zahl von Positionen und Funktionen als alternative Optionen zum gleichen Zeitpunkt, und b) die Eignung für die Bewältigung einer Sequenz von (meist unvorhersehbaren) Änderungen im Lauf des Lebens» (Mertens, 1974, S. 40). Schlüsselqualifikationen

ermöglichen es Individuen, sich einer veränderten Umwelt anzupassen und sich einen (Arbeits)Platz in der Berufswelt zu sichern (sog. *employability*). Mertens (1974, S. 41ff.) unterscheidet zwischen vier Bildungselementen: Basisqualifikationen (Qualifikationen höherer Ordnung, z.B. logisches, analytisches und kritisches Denken, kooperatives und kreatives Vorgehen), Horizontalqualifikationen (Wesen von Informationen kennen, Informationen gewinnen, verstehen, verarbeiten), Breiterelemente (ubiquitäre Ausbildungselemente, praktische Anforderungen am Arbeitsplatz, z.B. Sprachkenntnisse, EDV-Kenntnisse), Vintagefaktoren (generationsbedingte Differenzen im Bildungsstand).

Das Konzept der *ganzheitlichen Handlungskompetenz*, mit den überlappenden Facetten der Fach-, Sozial- und Personalkompetenz (vgl. Roth, 1971) ist in der Berufsbildung weit verbreitet. Entsprechend findet es sich als *berufliche Handlungskompetenz* in theoretischen Arbeiten (z.B. Reetz, 1990, 1999), Rechtstexten (Art. 21 Abs. 2 lit. a Berufsbildungsgesetz, Art. 12 Abs. 1 lit. b Berufsbildungsverordnung, Art. 49 Abs. 2 Berufsbildungsverordnung), Bildungsplänen und curricularen Kompetenzmodellen (vgl. z.B. SKKAB, 2011a).²⁴ Dieses Konzept hat sich jedoch in der Berufsbildungsforschung für die empirische Kompetenzdiagnostik nicht bewährt, denn Sozial- und Personalkompetenzen lassen sich nur schwer in operationalisierbare Kompetenzmodelle überführen und mittels standardisierter Testverfahren erfassen.

Kognitive Konzeptionen der Kompetenz unterscheiden zwischen Performanz und Kompetenz (Norris, 1991, S. 333f.). Kompetenz ist eine latente Fähigkeit, die sich in einer *idealen* Situation entfaltet, welche sich aber vom manifesten Verhalten in einer *konkreten* Situationen unterscheiden kann (Maag Merki, 2009, S. 493f.). «Competence is [...] emphasizing what people are ideally able to do» und «Performance is [...] emphasizing what people typically do» (Sternberg & Ben-Zeev, 2001, S. 360, S. 373, zit. nach Winther, 2010, S. 20). Kompetenz ist demnach eine *Voraussetzung* für Performanz in spezifischen Situationen (Maag Merki, 2009, S. 493). Diese Unterscheidung geht zurück auf den Linguisten Chomsky (1965), der Kompetenz als allgemeines Sprachvermögen und Performanz als aktuelle Sprachverwendung verstand.

Nach den kognitiven Konzeptionen wird Kompetenz über kognitive Strukturen einer Person repräsentiert, deren Komponenten und Umfang *per se* nicht definiert ist und einer inhaltlichen Verknüpfung mit konkreten Situationen und Anforderungen bedarf. Bei den *Komponenten* von kognitiven Strukturen wird in der psychologischen Literatur z.B. zwischen deklarativen und prozeduralen Repräsentationen (J. R. Anderson, 1983, 1993, 1996) oder zwischen *conceptual*,

²⁴ Im Bildungsplan Kauffrau/Kaufmann EFZ wird berufliche Handlungskompetenz durch einen Kompetenzwürfel mit den drei Elementen Sachkompetenz, Methodenkompetenz sowie Sozial- und Selbstkompetenz dargestellt (SKKAB, 2011a, S. 7ff.). Im Bereich der Sozial- und Selbstkompetenz sollen folgende Fähigkeiten gefördert und erreicht werden: Leistungsbereitschaft, Kommunikationsfähigkeit, Teamfähigkeit, Umgangsformen, Lernfähigkeit, ökologisches Bewusstsein.

procedural und *interpretive competence* (Gelman & Greeno, 1989; Greeno, Riley, & Gelman, 1984)²⁵ unterschieden. Der *Umfang* der kognitiven Strukturen bezieht sich auf fachliche oder auf fachübergreifende Repräsentationen (Maag Merki, 2009, S. 496f.; Weinert, 2001b, S. 28). Fachliche Kompetenz, auch als *domänenspezifische* bezeichnet, bezieht sich funktional auf einen Fachbereich (Fachdisziplin, Schulfach wie z.B. Erstsprache, Fremdsprachen, Mathematik, Naturwissenschaften) oder einen Gegenstandsbereich (spezifische Lebenssituation, spezifische berufliche Tätigkeiten). Fachübergreifende Kompetenz, auch als *domänenverbundene* bezeichnet, ist für verschiedene Fach- und Gegenstandsbereiche funktional (z.B. Problemlösen, Teamfähigkeit, Selbstregulation).

Kognitive Konzepte sind v.a. in der Bildungsforschung verbreitet, um die Bedingungs- und Entwicklungsfaktoren von Kompetenz, z.B. durch Lerngelegenheiten im schulischen und ausserschulischen Kontext, zu untersuchen (Blömeke et al., 2015, S. 5). McClelland (1973) entwickelte Kompetenz als Gegenbegriff zu Intelligenz und argumentierte, dass im schulischen Kontext nicht Intelligenz, sondern Kompetenz getestet werden solle. Mit Kompetenz meinte er das konkrete Handeln in verschiedenen realen Lebenssituationen, «clusters of life outcomes» (McClelland, 1973, S. 9, zit. nach Maag Merki, 2009, S. 493), das anhand von *criterion tasks* beobachtet werden kann. In den gross angelegten Studien (*large scale assessments*) zur Leistungsmessung in Schulen, allen voran PISA (vgl. z.B. OECD, 2013, 2017a), werden die erfassten Schüler*innenleistungen als Ausprägung spezifischer fachlicher Kompetenzen betrachtet (z.B. Lesekompetenz, Mathematikkompetenz, naturwissenschaftliche Kompetenz, Finanzkompetenz). «The concept of competence has become associated with a drive towards more practicality in education and training placing a greater emphasis on the assessment of performance rather than knowledge» (Norris, 1991, S. 331).

²⁵ *Conceptual competence* ist das abstrakte, regelbasierte Wissen in einer Domäne. *Procedural competence* sind Fähigkeiten und Fertigkeiten, um *conceptual competence* in konkreten Situationen anzuwenden. *Interpretive competence* sind die Fähigkeiten und Fertigkeiten, um die Merkmale der konkreten Anforderungssituation zu beurteilen und um geeignete Problemlösestrategien auszuwählen und nutzen.

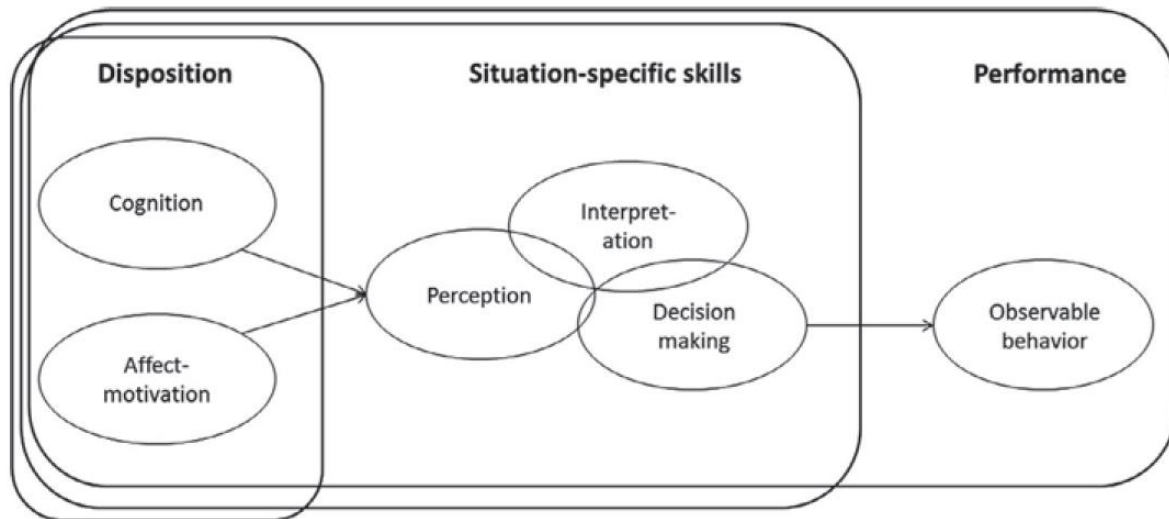


Abbildung 3-1. Kompetenz als Kontinuum

Quelle: Blömeke, Gustafsson, & Shavelson (2015, S. 7).

Ein neuerer Forschungsansatz sieht Kompetenz als horizontales und vertikales *Kontinuum* (vgl. Blömeke et al., 2015). Diese „prozessuale“ Konzeption ist bestrebt, die traditionelle Dichotomie der behavioristischen (holistischen) und kognitiven (analytischen, *latent trait*) Ansätze aufzulösen, indem sie beide verbindet. Kompetenz kann einerseits als horizontales Kontinuum interpretiert werden: Individuelle Dispositionen (kognitive, affektive, motivationale Merkmale) und individuelle Performanz (beobachtbares Verhalten) sind durch *Prozesse* (z.B. Wahrnehmung, Interpretation und Entscheidung) in spezifischen realen Anforderungssituationen strukturell miteinander verbunden (vgl. Abbildung 3-1). «Competence is viewed along a continuum from traits that underlie perception, interpretation, and decision-making skills, which in turn give rise to observed behavior in real-world situations» (Blömeke et al., 2015, S. 3). Andererseits kann Kompetenz als vertikales Kontinuum interpretiert werden, auf dem sich höhere und geringere Kompetenzniveaus abbilden lassen. «[...] competence is also a vertical continuum in terms of performance levels and of developmental stages. [...] competence is a continuous characteristic with higher and lower levels (more or less competent)» (Blömeke et al., 2015, S. 7). Wenn Kompetenz als multidimensionales Konstrukt definiert wird, dann muss theoretisch und empirisch geklärt werden, inwiefern die verschiedenen Dimensionen additiver Natur (Dimensionen können einander kompensieren) oder multiplikativer Natur (Dimensionen können einander nicht kompensieren) sind (Blömeke et al., 2015, S. 7).

Das Kontinuum-Kompetenzkonzept wurde originär für die Hochschulbildung vorgeschlagen, findet aber zunehmend auch in anderen Bildungsbereichen Aufmerksamkeit.

3.2 Kompetenzdefinitionen in der Bildungsforschung

In der Bildungsforschung findet sich für Kompetenz kein gemeinsames Rahmenkonzept und keine einheitliche, allgemein akzeptierte Definition, sondern eine Vielzahl unterschiedlicher Definitionen (Hartig, 2008, S. 16; Klieme, Maag Merki, & Hartig, 2007, S. 5; Maag Merki, 2009, S. 493; Weinert, 2001a, S. 46). Uneinheitliche Definitionen von Kompetenz erschweren die wissenschaftliche Kommunikation und verunmöglichen die Vergleichbarkeit verschiedener empirischer Studien (Hartig, 2008, S. 16). In seiner Expertise trägt Weinert (1999, 2001a) einen Katalog von Kompetenzkonzepten zusammen, nach denen Kompetenz bislang definiert und interpretiert worden war (Weinert, 1999, 2001a, zit. nach Hartig & Klieme, 2006, S. 128f.):

- (1) Kompetenzen als generelle kognitive Leistungsdispositionen, die für die Bewältigung sehr unterschiedlicher Aufgaben benötigt werden;
- (2) Kompetenzen als kontextspezifische kognitive Leistungsdispositionen, die sich funktional auf bestimmte Klassen von Situationen und Anforderungen beziehen, diese spezifischen Leistungsdispositionen lassen sich auch als Kenntnisse, Fertigkeiten oder Routinen charakterisieren;
- (3) Kompetenzen als motivationale Orientierungen, die für die Bewältigung von anspruchsvollen Aufgaben benötigt werden;
- (4) Handlungskompetenz als eine Integration der drei erstgenannten Konzepte, bezogen auf die Anforderungen eines spezifischen Handlungsfeldes wie z.B. eines Berufes;
- (5) Metakompetenzen als das Wissen, die Strategien oder die Motivationen, welche sowohl den Erwerb als auch die Anwendung spezifischer Kompetenzen erleichtern;
- (6) Schlüsselkompetenzen als Kompetenzen im unter (2) genannten funktionalen Sinn, die aber für einen relativ breiten Bereich von Situationen und Anforderungen relevant sind. Hierzu gehören z.B. muttersprachliche oder mathematische Kenntnisse.

Weinert (1999; 2001a, S. 62) selber formuliert in der oben erwähnten Expertise keine Kompetenzdefinition; er tut es aber – fast zeitgleich – in einer anderen Publikation (Weinert, 2001b, S. 27f.):

«[Kompetenz ist] die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten, um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können» (Weinert, 2001b, S. 27f.).

Diese Definition lehnt sich an das unter (4) genannte Konzept der Handlungskompetenz²⁶ an. Sie wird mitunter in einigen Studien zur Modellierung und Messung beruflicher Handlungskompetenzen von Lernenden (vgl. z.B. K. Beck et al., 2016; Holtsch & Eberle, 2018) sowie professioneller Handlungskompetenz von Lehrer*innen (vgl. z.B. Baumert & Kunter, 2006) verwendet. Andere Studien verwenden die unter (2) genannte Definition für die sog. berufsfachliche Kompetenzen von Lernenden (Lehmann & Seeber, 2007; Winther, 2010; Winther & Achtenhagen, 2008).

In den internationalen Schulleistungstudien, z.B. PISA (vgl. z.B. OECD, 2013, 2017a), TIMSS (vgl. Baumert, Bos, & Lehmann, 2000) und DESI (vgl. B. Beck & Klieme, 2007), hat sich hingegen das unter (2) genannte Konzept der kontextspezifischen kognitiven Leistungsdispositionen etabliert. Die von diesem Konzept abgeleiteten Definitionen sind durch zwei wesentliche Restriktionen charakterisiert (Hartig & Klieme, 2006, S. 129; Klieme, Maag Merki, et al., 2007, S. 7). Zum einen ist Kompetenz funktional bestimmt und bereichsspezifisch, d.h. auf einen begrenzten Sektor von Situationen; damit sind allgemeine kognitive Leistungsdispositionen explizit ausgeschlossen, obschon sie ebenfalls zum erfolgreichen Handeln beitragen. Zum anderen ist Kompetenz auf kognitive Dispositionen beschränkt, d.h. motivationale und affektive Dispositionen sind explizit ausgeschlossen.

Die oben aufgelisteten Kompetenzkonzepte haben drei Punkte gemeinsam, die sie vom Konzept der Intelligenz abgrenzen (Hartig & Klieme, 2006, S. 129ff.; Maag Merki, 2009, S. 495f.):

- *Kontextualisierung*: Kompetenz ist eine bereichsspezifische Fähigkeit, um spezifische Situationen und Anforderungen zu bewältigen. Intelligenz ist hingegen eine generalisierbare Fähigkeit, um neue Probleme zu lösen.
- *Lernbarkeit*: Kompetenz ist lernbar und wird durch das Sammeln von Erfahrung mit den spezifischen Situationen und Anforderungen erworben; sie kann grundsätzlich im schulischen Unterricht gefördert werden. Umgekehrt gilt Intelligenz als zeitlich relativ stabil und wird grösstenteils durch biologische Faktoren determiniert, d.h. ist vererbbar.
- *Binnenstruktur* (zur Beschreibung interindividueller Leistungsunterschiede): Beim Kompetenzkonzept ergibt sich die Binnenstruktur aus den spezifischen Situationen (Inhalten) und den damit einhergehenden Anforderungen. Dem gegenüber ergibt sich beim Intelligenzkonzept die Binnenstruktur aus generellen kognitiven und biologischen Prozessen, die für die Intelligenzleistung notwendig sind.

²⁶ Handlungskompetenz bezeichnet alle bei einer Person verfügbaren Handlungskonzepte und Handlungsschemata. Handlungskonzepte werden für spezifische Handlungen entwickelt, Handlungsschemata werden für ähnliche Handlungen generalisiert (Edelmann & Wittmann, 2012, S. 177f.).

Für eine wissenschaftliche Studie, die sich der Kompetenzmodellierung und -messung widmet, ist eine explizite und präzise Definition der interessierenden Kompetenz notwendig. Die Kompetenzdefinition bestimmt bzw. begrenzt die Entwicklung oder Verwendung des Testinstruments, die Methoden der Testauswertung und die Interpretation der Testergebnisse bezüglich Binnenstruktur und Niveaus (Blömeke et al., 2015, S. 3; Hartig, 2008, S. 16). Ausgangspunkt der Kompetenzdefinition ist die Eingrenzung des *Kontexts*, d.h. des Bereichs von Situationen und Anforderungen, in denen sich die Kompetenz zeigen soll (Hartig, 2008, S. 19ff.). Der Kontext wird auch als *Domäne* i.S.v. Fachbereich oder Gegenstandsbereich bezeichnet. Es geht also darum, eine hinreichende und repräsentative Menge *ähnlicher* realer Situationen zu bestimmen, in denen *ähnliche* spezifische Anforderungen bewältigt werden müssen.

«The definition of competence therefore has to start from an analysis of authentic job or societal situations and enumerate the tasks as well as the cognition, conation, affect, and motivation involved»
(Blömeke et al., 2015, S. 5).

Wird der Kontext zu breit oder zu beliebig gewählt, verschwimmt die Grenze zwischen den beiden Konzepten kontextspezifischer und allgemeiner Leistungsdispositionen (siehe oben, (1) und (2)), und die Entwicklung eines Testinstruments zur validen Erfassung der Kompetenz wird nahezu unmöglich (Hartig, 2008, S. 20). Wird hingegen der Kontext zu eng gewählt, wird mit dem Testinstrument isoliertes deklaratives oder prozedurales Wissen abgefragt und als Kompetenz etikettiert (Hartig, 2008, S. 20f.).

3.3 Kompetenzmodelle für die pädagogisch-psychologische Diagnostik

Die Kompetenzdiagnostik erfolgt mit psychologischen Tests, genauer mit standardisierten Leistungstests, wie sie auch in der Psychologie eingesetzt werden (z.B. Intelligenztests) (vgl. Kapitel 5.1). Bei der empirischen Erfassung von Kompetenzen braucht es theoretische Modelle als Grundlage für die Entwicklung von Testinstrumenten (auch: Messinstrumente) und für die Interpretation der Testergebnisse. In der Literatur wird zwischen zwei Arten von Kompetenzmodellen unterschieden (Hartig & Jude, 2007, S. 33ff.; Hartig & Klieme, 2006, S. 132ff.; Klieme, Maag Merki, et al., 2007, S. 11ff.): Kompetenzstrukturmodelle (auch: Komponentenmodelle) und Kompetenzniveaumodelle (auch: Stufenmodelle). Diese Modelle beziehen sich auf unterschiedliche Aspekte des Kompetenzkonstrukts; sie schliessen sich also nicht aus, sondern ergänzen sich idealerweise. Die Wahl des jeweiligen Kompetenzmodells ist abhängig von der Zielsetzung der Kompetenzmodellierung und -messung.

Kompetenzstrukturmodelle beschreiben die inhaltliche Strukturierung (Binnenstrukturierung) der Kompetenz bezüglich Teilkompetenzen, sog. Dimensionen (Hartig & Klieme, 2006, S. 132f.; Klieme, Maag Merki, et al., 2007, S. 12f.) (vgl. Abbildung 3-2). Es wird theoretisch postuliert und

empirisch untersucht, wie die Bewältigung verschiedener Anforderungssituationen (Testitems) miteinander zusammenhängt und inwiefern sich Bündel von Anforderungssituationen (Testkomponenten oder einzelne Tests) bilden lassen.²⁷ Mit Kompetenzstrukturmodellen lassen sich also interindividuelle Unterschiede in der Kompetenz bzw. in den Teilkompetenzen angemessen erfassen und diagnostizieren.

Für die Etablierung eines Kompetenzstrukturmodells ist folgendes Vorgehen angezeigt (Hartig & Jude, 2007, S. 33f.; Hartig & Klieme, 2006, S. 132f.): (1) Basierend auf der klassischen Testtheorie (KTT) eine Faktorenanalyse rechnen (vgl. Kapitel 5.2.1.b) oder basierend auf der Item-Response-Theorie (IRT) ein mehrdimensionales Modell schätzen; (2) die Güte der Messmodelle anhand der χ^2 -Statistik auf Signifikanz prüfen und/oder anhand der Informationskriterien vergleichen; (3) die Korrelationen zwischen den latenten Faktoren schätzen; eine hohe Faktorenkorrelation spricht für eine Dimension, eine geringe für zwei Dimensionen; (4) die Korrelationen zwischen den Indikatoren schätzen; Testitems mit hoher Inter-Item-Korrelation werden derselben Dimension zugeordnet und erfassen dieselbe Teilkompetenz, Testitems mit geringer Inter-Item-Korrelation erfassen entsprechend verschiedene Teilkompetenzen.

Kompetenzniveaumodelle beschreiben die Graduierung der Kompetenz anhand von Abschnitten, sog. Niveaus (Hartig & Klieme, 2006, S. 133ff.; Klieme, Maag Merki, et al., 2007, S. 11f.) (vgl. Abbildung 3-3). Auf der kontinuierlichen Kompetenzskala werden Abschnitte festgelegt und anhand der spezifischen Anforderungen (Testitems) inhaltlich beschrieben.²⁸ Die Kompetenzniveaus geben Auskunft darüber, welche spezifischen Anforderungen eine Person mit hoher Kompetenzausprägung bewältigen kann und welche Anforderungen eine Person mit niedriger Kompetenzausprägung gerade noch oder nicht mehr bewältigen kann. Mit Kompetenzniveaumodellen lassen sich also individuelle Ausprägungen der Kompetenz kriteriumsorientiert interpretieren bzw. diagnostizieren.

²⁷ In PISA 2015 wurden die Lesekompetenz mit drei Teilkompetenzen erfasst: Zugreifen und Abrufen von Informationen abrufen (access and retrieve), Texte integrieren und interpretieren (integrate and interpret), Textinhalte und Textformen reflektieren und evaluieren (reflect and evaluate) (OECD, 2017a, S. 56ff.). In PISA 2015 wurde die Mathematikkompetenz mit vier Teilkompetenzen erfasst: Veränderung und Beziehung (change and relation), Raum und Form (space and shape), Quantität (quantity), Unsicherheit (uncertainty and data) (OECD, 2017a, S. 72ff.). In der DESI-Studie wurde für die Sprachkompetenz im Englischen drei theoretische Kompetenzdimensionen angenommen (Rezeption, Bewusstheit, Produktion) und mit insgesamt sieben Leistungstests erfasst (B. Beck & Klieme, 2007).

²⁸ In PISA 2015 wurde für die Lesekompetenz sieben Niveaus beschrieben (OECD, 2017a, S. 60f.) und für die Mathematikkompetenz sechs Niveaus (OECD, 2017a, S. 78f.).

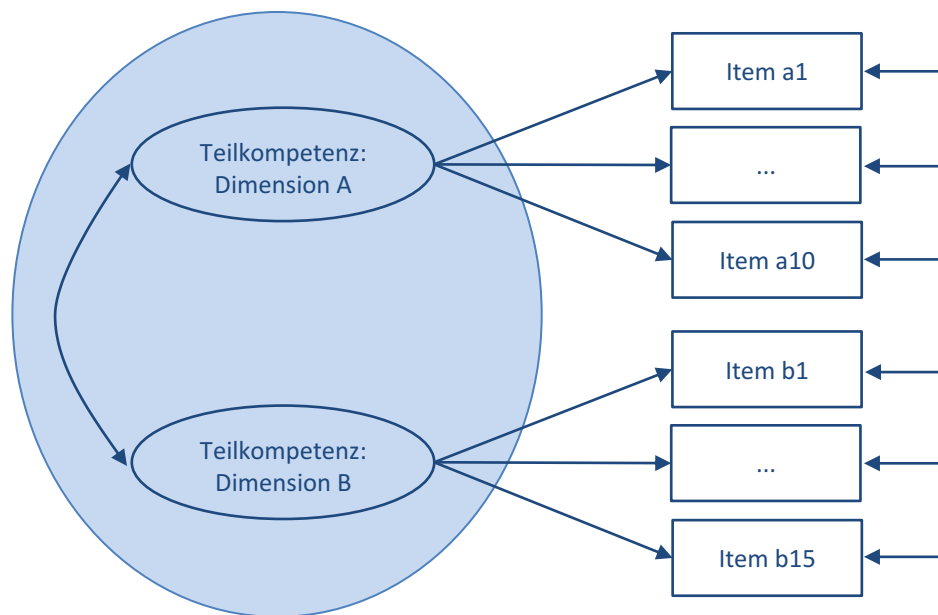


Abbildung 3-2. Kompetenzstrukturmodell für Diagnostik

Quelle: Eigene Darstellung i.A.a. Hartig & Klieme (2006, S. 132).

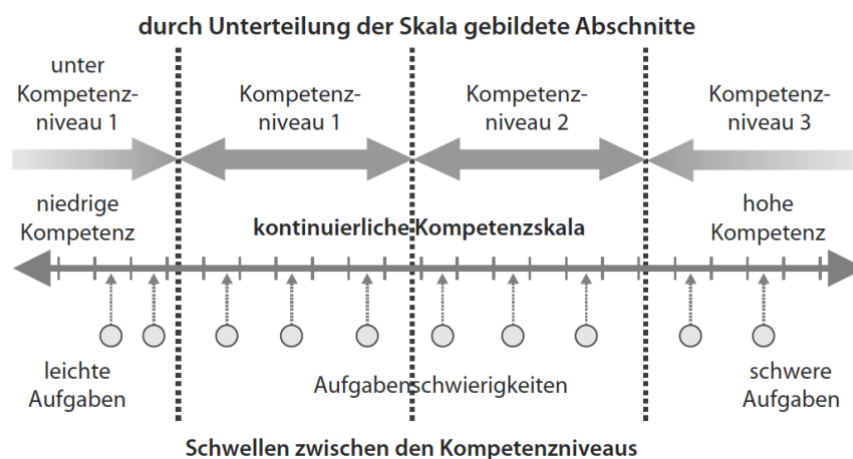


Abbildung 3-3. Kompetenzniveaumodell für Diagnostik

Quelle: Hartig & Klieme (2006, S. 135).

Für die Etablierung eines Kompetenzniveaumodells wird folgendes Vorgehen vorgeschlagen (Hartig & Jude, 2007, S 34ff.; Hartig & Klieme, 2006, S. 133ff.): (1) Basierend auf der Item-Response-Theorie (IRT) mit einem Rasch-Modell die Item-Schwierigkeiten und Personenfähigkeiten schätzen und gemeinsam skalieren (vgl. Kapitel 5.2.3); (2) die Kompetenzskala in Abschnitte unterteilen, entweder willkürlich (z.B. gleich grosse Abstände) oder regelgeleitet (z.B. mittlere Testwerte von Jahrgangsstufen aus der Stichprobe, Regressionskoeffizienten der schwierigkeitsrelevanten Aufgabenmerkmale); (3) die Abschnitte inhaltlich beschreiben anhand der Anforderungen der Testaufgaben.

Ausgehend von einem Kompetenzniveaumodell lässt sich auch ein sog. Kompetenzentwicklungsmodell ableiten, das Anforderungen von Novizen und Experten beschreibt.

3.4 Konzepte und Instrumente für ökonomische Kompetenz

In diesem Kapitel werden Konzepte und Instrumente für *economic literacy* bzw. ökonomische Kompetenz vorgestellt, die für diese Dissertation insofern bedeutsam sind, als dass sie als Abgrenzung oder Anknüpfung fungieren. Es sind dies in chronologischer Reihenfolge: das Konzept der wirtschaftskundlichen Kompetenz und der „Wirtschaftskundliche Bildungs-Test (WBT)“ nach Beck (Kapitel 3.4.1), das Konzept der ökonomischen Kompetenzen und der OEKOMA-Test nach Schumann und Eberle (Kapitel 3.4.2) sowie das Modell der umfassenden Wirtschaftskompetenz im kaufmännischen Bereich nach Eberle (vgl. Kapitel 3.4.3). Im Unterschied zu den oben vorgestellten Kompetenzmodellen der *ökonomischen Bildung*, welche vorderhand die Formulierung und Implementierung schulischer Bildungsstandards bezwecken (vgl. Kapitel 2.1.3), beziehen sich die nachfolgenden Konzepte der *ökonomischen Kompetenz* auf ein allgemeines, z.T. nicht curricular verankertes Bildungsideal²⁹ und bezwecken die Messung der definierten Kompetenz.

Für die empirische Kompetenzmodellierung und -messung in der Berufs- und Wirtschaftspädagogik lassen sich zwei Strömungen unterscheiden. Die eine Forschungsströmung befasst sich mit beruflichen und kaufmännischen Kompetenzen (vgl. z.B. K. Beck et al., 2016; Guggemos & Schönlein, 2015; Holtsch & Eberle, 2018; Lehmann & Seeber, 2007; Nickolaus et al., 2008; Winther, 2010; Winther & Achtenhagen, 2008, 2009b). Berufliche Kompetenzmodelle richten sich an spezifischen beruflichen Anforderungssituationen aus, die in jüngster Zeit vermehrt technologiebasiert mit Simulationstests dargeboten werden (vgl. z.B. K. Beck et al., 2016; Winther & Achtenhagen, 2009a).

Die andere Forschungsströmung befasst sich mit allgemeinen ökonomischen Kompetenzen, also nicht-beruflichen (vgl. z.B. Aprea et al., 2014; K. Beck, 1993; Müller, Fürstenau, & Witt, 2007; Schumann & Eberle, 2014a; Schumann et al., 2017; Schumann, Oepke, & Eberle, 2011; Schürkmann & Schuhen, 2013). Konzepte der *economic literacy* und ökonomischen Kompetenz beziehen sich traditionellerweise auf eine fachwissenschaftliche Systematik.

²⁹ Eine Ausnahme ist der *Test of Economic Literacy* (Soper, 1979; Soper & Walstad, 1987; Walstad & Rebeck, 2001; Walstad, Rebeck, & Butters, 2013b) aus den USA, der sich auf das dortige Framework bzw. die Standards für *Economic Education* (CEE, 2010; Hansen, Bach, Calderwood, & Saunders, 1977; Saunders & Gilliard, 1995) bezieht.

3.4.1 Wirtschaftskundliche Kompetenz und WBT (Beck, 1993)

Die Modellierung und Messung der *economic literacy* hat in den USA eine lange Forschungstradition. Ein erster Meilenstein stellt der Bericht „*Economic education in the schools*“ des *Committee for Economic Development* (CED, 1961) und das „*Framework for Teaching Economics*“ (Hansen et al., 1977; Saunders & Gilliard, 1995) dar. Das *Framework* wurde in den 2000er-Jahren durch die „*Voluntary National Content Standards in Economics*“ ersetzt, das aus 20 *key concepts of economics* besteht, z.B. *productive resources, markets and prices, money and inflation, competition* (CEE, 2010). Ein weiterer Meilenstein war der *Test of Economic Literacy* (TEL) (Soper, 1979; Soper & Walstad, 1987; Walstad & Rebeck, 2001; Walstad et al., 2013b). Die vierte Ausgabe des TEL (TEL4) basiert auf den *Standards* und beinhaltet zu jedem der 20 *key concepts* eins bis drei Items (Walstad et al., 2013b, S. 5ff.). Die *key concepts* werden in zwei Testversionen A und B abgefragt, wobei jede Version 45 Items im gebundenen Antwortformat enthält. Die Items im TEL4 wurden entlang von drei Kognitionsstufen³⁰ konstruiert: *knowledge, comprehension* und *application* (Walstad et al., 2013b, S. 7f., S. 9). Der TEL wurde mittlerweile in zahlreichen US-amerikanischen und internationalen Studien eingesetzt (vgl. z.B. K. Beck & Krumm, 1989; Walstad, 2005), bei denen für einen Grossteil der Schüler*innen ein „ökonomischer Analphabetismus“ identifiziert wurde (S. Seeber, Schumann, & Nickolaus, 2015, S. 171).

Für den deutschsprachigen Raum erfolgte eine Übersetzung und Normierung des TEL als „Wirtschaftskundlicher Bildungs-Test (WBT)“ durch Klaus Beck (K. Beck, 1993; K. Beck & Krumm, 1998). Der WBT war das erste standardisierte Testverfahren zur Erfassung der «wirtschaftskundlichen Grundfähigkeit» (K. Beck, 1993, S. 9). Das dem WBT zugrunde gelegte Konstrukt ist die «wirtschaftskundliche Kompetenz» (ebd., S. 10), die sich als «*ökonomiebezogene Grundfähigkeit* [...] über zwei Funktionen konstituiert, nämlich dem aktualisierbaren *ökonomiebezogenen Wissen* und den darauf operierenden *ökonomiebezogenen Denkleistungen*» (ebd., S. 10, Hervorhebungen im Original).

Der WBT gliedert sich, in Anlehnung an den TEL2 (Soper & Walstad, 1987), in die vier Inhaltsbereiche „Grundlagen“, „Mikroökonomie“, „Makroökonomie“ und „internationale Beziehungen“; insgesamt werden 22 Teilbereiche abgebildet. Die Items differenzieren sich nach fünf Kognitionsstufen der Bloomschen Taxonomie (vgl. Bloom et al., 1956): Wissen, Verstehen, Anwenden, Analyse, Evaluation. Der WBT hat ebenfalls zwei Testversionen mit je 46 Items, wobei 15 Items in beiden Testversionen enthalten sind (sog. Ankeritems) (K. Beck, 1993, S. 16ff.). Der WBT richtet sich hauptsächlich an Schüler*innen auf der Sekundarstufe II (K. Beck, 1993, S. 15).

³⁰ Die Items im TEL1 und der TEL2 hatten sich in Anlehnung an die kognitive Taxonomie von Bloom et al. (1956) auf fünf Kognitionsstufen bezogen. Im TEL3 und TEL4 wurden sie auf drei Kognitionsstufen reduziert (Walstad et al., 2013b, S. 7).

Der WBT wurde Anfang der 1990er-Jahre in einer gross angelegten Studie in fünf deutschen Bundesländern bei ca. 9'000 Schüler*innen aus allgemeinbildenden Schulen (u.a. Gymnasien) und beruflichen Schulen (Branchen Einzelhandel, Industrie, Banken, und Versicherungen) eingesetzt (K. Beck, 1993, S. 33ff.). Die Ergebnisse der Studie zeigten, dass die Schüler*innen im Durchschnitt „nur“ ca. 45 % der Items lösen konnten und dass die Unterschiede zwischen den Schulformen und zwischen den kaufmännischen Ausbildungsberufen erheblich waren (K. Beck, 1993, S. 70).

Kritisiert wird am WBT als Messinstrument für «wirtschaftskundliche Kompetenz», dass er sich zum einen einzig auf Aspekte des Wissens bezieht und insofern an ein *literacy*-Konzept und nicht an ein Kompetenzkonzept anlehnt, und dass er zum anderen ökonomische Bildung auf die Volkswirtschaftslehre verkürzt. So werden im WBT insbesondere gesamtwirtschaftliche Zusammenhänge und wirtschaftspolitische Entscheidungen ausklammert (S. Seeber et al., 2015, S. 171f.).

3.4.2 *Ökonomische Kompetenzen und OEKOMA-Test (Schumann & Eberle, 2014)*

Für den Schweizer Kontext entwickelten Schumann und Eberle (Schumann & Eberle, 2014a; Schumann et al., 2011) ein Konzept für ökonomische Kompetenzen und darauf aufbauend ein Testinstrument zur Erfassung dieser Kompetenzen (OEKOMA-Test). Das zugrunde gelegte Kompetenzkonzept für ökonomische Kompetenzen umfasst in Anlehnung an Weinert (2001b, S. 27f.) fachlich-kognitive und motivational-affektive Komponenten, also «die Gesamtheit des Wissens, der Fertigkeiten und Bereitschaften eines Individuums, wirtschaftliche Problemstellungen erfolgreich und verantwortungsvoll lösen zu können» (Schumann et al., 2011, S. 54). Die *Kerndimension* der ökonomischen Kompetenzen ist «das ökonomische Wissen und Können als Voraussetzung zur Lösung wirtschaftlicher Problemstellungen» (Schumann & Eberle, 2014a, S. 107; Schumann et al., 2011, S. 55) und kann als *economic literacy* im weiteren Sinne verstanden werden. Diese Kerndimension ökonomischer Kompetenzen ist normativ mit einer ökonomischen Grundbildung verknüpft, die auf «eine gelingende individuelle Lebensgestaltung sowie auf soziale und politische Teilhabe des Einzelnen» fokussiert (Schumann & Eberle, 2014a, S. 107).

«Die Referenzfigur ist demnach ein Mensch, welcher als „mündiger Wirtschafts- und Gesellschaftsbürger“ in der Lage ist, wirtschaftsbezogene Problemstellungen zu verstehen, zu analysieren und begründete Schlüsse für (potentielle) Lösungen daraus zu ziehen, also über wirtschaftsbürgerliche Kompetenz verfügt» (Schumann & Eberle, 2014a, S. 107).

Tabelle 3-1. Testaufbau und Testinhalt des OEKOMA-Tests

Inhaltsbereich	Teilbereiche
Volkswirtschaft (VWL)	▪ Mikroökonomie (3 Itemstämme) ▪ Makroökonomie (5 Itemstämme) ▪ Aussenwirtschaft (2 Itemstämme)
Betriebswirtschaft (BWL)	▪ Strategische Unternehmensführung (2 Itemstämme) ▪ Bereiche der Unternehmung (4 Itemstämme) ▪ Corporate Finance (4 Itemstämme)

Quelle: Schumann et al. (2010).

Der OEKOMA-Test erfasst die Kerndimension der ökonomischen Kompetenzen, also ökonomisches Wissen und Können. Er gliedert sich in die zwei Inhaltsbereiche Volkswirtschaft (VWL) und Betriebswirtschaft (BWL) und unterteilt diese wiederum in drei Teilbereiche (Schumann et al., 2010, S. 7f.) (vgl. Tabelle 3-1). Die konkreten Inhalte des OEKOMA-Tests wurden in einem zweischrittigen Verfahren ausgewählt (Schumann & Eberle, 2014a, S. 111f.; Schumann et al., 2010; Schumann et al., 2011, S. 59): (1) Fachanalyse anhand der Unterlagen im Studiengang Wirtschaftswissenschaft an zwei Schweizer Hochschulen; daraus wurde ein Kategoriensystem für eine fachliche Inhaltsstruktur abgeleitet. (2) Medienanalyse von zwei führenden Schweizer Tageszeitungen; die in den Zeitungsartikeln entdeckten ökonomischen Konzepte wurden dem Kategoriensystem zugeordnet und die Inhaltsbereiche und Teilbereiche danach gewichtet. Ausgehend vom Konstrukt der *economic literacy* wurden die Itemstämme aus modifizierten Zeitungsartikeln mit 200 bis 300 Wörtern gebildet und zu jedem Itemstamm vier bis sieben Items formuliert (Schumann & Eberle, 2014a, S. 112). Der OEKOMA-Test umfasst insgesamt 111 Items, wobei ca. 90 % ein Multiple-Choice-Format und die restlichen ein freies Antwortformat haben.

In der OEKOMA-Studie wurden im Jahr 2011 ca. 3'000 Schüler*innen an gymnasialen Maturitätsschulen und an Berufsmaturitätsschulen in der Deutschschweiz untersucht, die in diesem Jahr ihre Abschlüsse erlangten (Schumann & Eberle, 2014a, S. 110f.). Für die Stichprobe wurden vier Straten gezogen: Gymnasialklassen im Schwerpunktfach „Wirtschaft & Recht“, Gymnasialklassen in anderen Schwerpunktfächern, kaufmännische Berufsmaturitätsklassen und nicht-kaufmännische Berufsmaturitätsklassen. Im Ergebnis ergab sich für das Konstrukt „ökonomisches Wissen und Können“ eine dreidimensionale Struktur entlang der Fachbereiche VWL, BWL und FRW (Schumann & Eberle, 2014a, S. 114f.). Zudem zeigten sich erhebliche Unterschiede zwischen den beiden Straten des Gymnasiums (Schwerpunktfach WuR vs. andere) und zwischen den Straten der Berufsmaturität (Profil kaufmännisch vs. andere) (Schumann & Eberle, 2014a, S. 118f.).

An diesem Kompetenzkonzept bzw. an der Kerndimension der ökonomischen Kompetenzen muss m.E. gewürdigt werden, dass es – als damaliges Novum in der Modellierung ökonomischer Kompetenzen – auf «wirtschaftliche Problemstellungen» und deren Lösung fokussiert und solche

realen wirtschaftlichen Problemstellungen als Inhaltsanker für die Testitems verwendet wurden. Kritisiert werden kann hingegen, dass die Inhaltselemente der Items hauptsächlich auf ökonomische Konzepte, i.S.e. *economic literacy* und *managerial literacy*, statt auf ökonomische Probleme ausgerichtet sind, und dass sich die Kognitionsprozesse der Items auf das Reproduzieren ökonomischer Konzepte reduzieren, die grösstenteils im Multiple-Choice-Format abgefragt werden. Der OEKOMA-Test vermag m.E. also nicht komplexe wirtschaftlichen Problemstellungen und deren kontroverse Lösungsansätze zu simulieren, die im Sinne einer «wirtschaftsbürgerlichen Bildung» gefordert werden.

3.4.3 Wirtschaftskompetenz im kaufmännischen Bereich (Eberle, 2015)

Ebenfalls für den Schweizer Kontext, jedoch für die kaufmännische Berufsbildung, entwickelte Eberle (2015) ein Modell der umfassenden *Wirtschaftskompetenz im kaufmännischen Bereich* (vgl. Abbildung 3-4). Das dem Modell explizit zugrunde gelegte Kompetenzkonzept lehnt sich an Weinert (2001b, S. 27f.) an und besteht aus kognitiven Merkmalen (Wissen und Können) sowie psychologischen Merkmalen (Lern- und Leistungsmotivation, Interesse, Einstellungen) (Eberle, 2015, S. 25f., S. 29). Da die kaufmännische Berufsbildung in der Schweiz eine zweifache Zielsetzung – eine berufsbildende und eine allgemeinbildende – hat, umfasst das Modell folgerichtig zwei Domänen, nämlich die kaufmännische und die wirtschaftsbürgerliche (Eberle, 2015, S. 26f., S. 29ff.).

«Der kaufmännischen Domäne ist die Kompetenz zuzuordnen, berufsspezifische kaufmännische Tätigkeiten ausüben bzw. entsprechende Probleme lösen zu können, z.B. die Planung, die Ausführung und das Controlling von professionellem Handeln am Arbeitsplatz. Dazu gehört etwa die korrekte Bearbeitung einer Kundenbestellung» (Eberle, 2015, S. 26).

«Der wirtschaftsbürgerlichen Domäne ist die Kompetenz zuzuordnen, private, wirtschaftspolitisch-volkswirtschaftliche sowie betriebswirtschaftliche Problemstellungen in staatlichen und nichtstaatlichen sozialen Systemen zu verstehen, vorgeschlagene Lösungen zu beurteilen und für einfachere Probleme selbst Lösungen zu entwickeln. [...] Ein aktuelles Beispiel hierfür ist das Verstehen und die Beurteilung des Vorgehens der Schweizer Nationalbank hinsichtlich des Wechselkurses des Schweizer Franken zum Euro» (Eberle, 2015, S. 26f.).

Die beiden Domänen beziehen sich auf zwei unterschiedliche Tätigkeits- und Problemlösebereiche, also auf unterschiedliche Handlungskompetenzen. Eine solche gemeinsame Modellierung der berufsbildenden und der allgemeinbildenden Domäne stellte für den deutschsprachigen Raum ein Novum dar und ist entsprechend komplexer als vergleichbare Kompetenzmodelle der beruflichen Bildung (Eberle, 2015, S. 27f.).

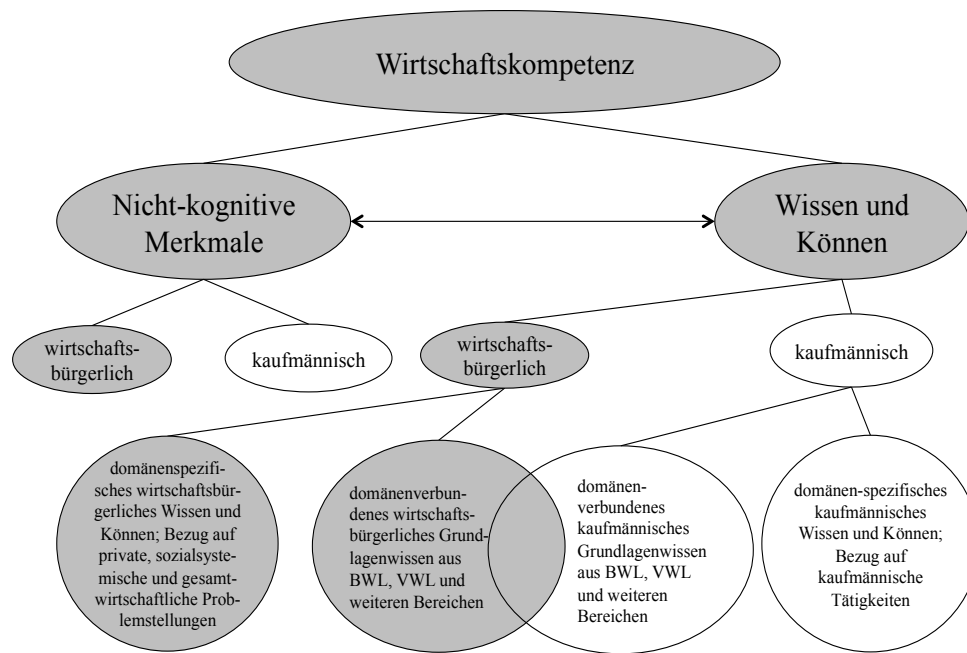


Abbildung 3-4. Modell der umfassenden Wirtschaftskompetenz im kaufmännischen Bereich

Quelle: Eberle (2015, S. 29).

Im Modell der umfassenden Wirtschaftskompetenz wird eine weitere Differenzierung der beiden Domänen vorgenommen, dies in Anlehnung an andere Kompetenzmodelle der beruflichen Bildung (vgl. Winther, 2010): Domänenspezifische Kompetenz bezieht sich auf die Bewältigung spezifischer Anforderungen, domänenverbundene Kompetenz unterstützt die Bewältigung spezifischer Anforderungen in verschiedenen Domänen. Entsprechend überschneiden sich die domänenverbundene kaufmännische und die domänenverbundenen wirtschaftsbürgerliche Kompetenz teilweise (Eberle, 2015, S. 27).

Das Modell der umfassenden Wirtschaftskompetenz im kaufmännischen Bereich bildete die theoretische Grundlage für zwei empirische Forschungsprojekte zur kaufmännischen Berufsbildung in der Schweiz. Im Leading House LINCA wurde u.a. kaufmännische und wirtschaftsbürgerliche Kompetenz von kaufmännischen Lernenden längsschnittlich untersucht (vgl. Holtsch & Eberle, 2018). Für die empirische Erfassung des kaufmännischen Wissens und Könnens wurde ein neues computerbasiertes Testinstrument entwickelt, genannt LINCASIM (Mentele et al., 2014) und in der LINCA-Studie zu drei Messzeitpunkten eingesetzt. Im Schweizer Teilprojekt des Verbundprojekts CoBALIT wurde die wirtschaftsbürgerliche Kompetenz von kaufmännischen Lernenden in zwei ausgewählten Branchen in der Schweiz und in Deutschland querschnittlich untersucht (vgl. Eberle et al., 2016; Schumann et al., 2017). Für die empirische Erfassung des wirtschaftsbürgerlichen Wissens und Könnens wurde ein neues Testinstrument entwickelt, der *wirtschaftsbürgerliche Kompetenztest* (WBK-Test) (vgl. Eberle et al., 2016).

Da sich das vorliegende Modell explizit auf die kaufmännische Berufsbildung bezieht und entsprechend zwei Domänen umfasst und verbindet, eignet es sich nicht als Grundlage für die Modellierung der wirtschaftsbürgerlichen Kompetenz, wie sie in beruflichen *und* gymnasialen Schulen zum Tragen kommt. Zudem ist die Unterteilung der Domänen in domänenspezifisch und domänenverbunden theoretisch unscharf beschrieben und lässt sich wohl auch empirisch nicht trennscharf messen, v.a. wenn das eine das andere voraussetzt.

Nichtdestotrotz bietet die mit diesem Modell mitgelieferte Definition der wirtschaftsbürgerlichen Kompetenz eine gute Ausgangslage für das Anliegen dieser Dissertation, nämlich die generelle Modellierung der wirtschaftsbürgerlichen Kompetenz unabhängig von einer Schulform bzw. einer Ausbildung. Auf die Beschreibung der wirtschaftsbürgerlichen Kompetenz (vgl. Kapitel 4.1) und auf die Operationalisierung der wirtschaftsbürgerlichen Kompetenz mit dem WBK-Test (vgl. Kapitel 7) wird unten ausführlich eingegangen.

3.5 Zwischenfazit

In der Wissenschaft lassen sich ausgehend von den verschiedenen Forschungstraditionen drei Kompetenzkonzepte unterscheiden: behavioristische, generische und kognitive. Kognitive Kompetenzkonzepte sind v.a. in der Bildungsforschung verbreitet, aber auch dort findet sich keine einheitliche und allgemein akzeptierte Kompetenzdefinition. Am häufigsten verwendet wird in empirischen Studien die Definition „kontextspezifische kognitive Leistungsdispositionen“ (Hartig & Klieme, 2006, S. 128; Weinert, 2001a, S. 47). Ebenfalls häufig wird auf die Definition „Handlungskompetenz“, d.h. kognitive, motivationale und soziale Voraussetzungen in einem spezifischen Handlungsfeld (Weinert, 2001b, S. 27f.) rekurriert. Die Kompetenzdiagnostik kann anhand eines Kompetenzstrukturmodells und/oder eines Kompetenzniveaumodells erfolgen, die sich bei der Interpretation der gemessenen Kompetenzen ergänzen können.

Die in der Wirtschaftspädagogik gängigen Konzepte der *economic literacy* bzw. der ökonomischen Kompetenz beziehen sich traditionellerweise auf eine fachwissenschaftliche Systematik, wie z.B. volkswirtschaftliche Themen. Dabei werden komplexe wirtschaftliche Problemstellungen und deren kontroverse Lösungsansätze vernachlässigt, was aber für eine Bildung des allgemeinen Wirtschafts- und Gesellschaftsverständnisses, die sog. «wirtschaftsbürgerliche Bildung», gefordert wird.

Diese Dissertation orientiert sich am Kontinuum-Kompetenzkonzept *sensu* Blömeke et al. (2015) sowie am kontextspezifischen kognitiven Kompetenzkonzept *sensu* Weinert (Weinert, 2001a, S. 46). Von den oben beschriebenen Konzepten für ökonomische Kompetenz eignet sich dasjenige der wirtschaftsbürgerlichen Kompetenz am besten für die in dieser Dissertation angestrebte Kompetenzmodellierung. Es wird eine Arbeitsdefinition für die «wirtschaftsbürgerliche» Kompetenz formuliert und ein Rahmenmodell mit Lebensbereichen und Kognitionsprozessen entworfen (vgl.

Kapitel 4.2). Für die Kompetenzdiagnostik wird schliesslich ein Kompetenzniveaumodell versucht (vgl. Kapitel 11.6).

4 Rahmenmodell für wirtschaftsbürgerliche Kompetenz

In diesem Kapitel wird den Leitfragen (3) und (4) nachgegangen (vgl. Kapitel 1.2):

- (3) *Wie lassen sich Inhaltsbereiche und Kognitionsprozesse für «wirtschaftsbürgerliche» Anforderungssituationen zu einem Rahmenmodell wirtschaftsbürgerlicher Kompetenz (WBK-Modell) zusammenführen?*
- (4) *Welche realen sozioökonomischen Problemsituationen ergeben sich im gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereich?*

Dieses Kapitel liefert zuerst eine kritische Reflexion der bisher verwendeten Definition der WBK (Kapitel 4.1). Ausgehend davon wird der Begriff der WBK präzisiert und ein neues Modell der WBK entworfen, das für die Allgemeinbildung und Berufsbildung anwendbar ist (Kapitel 4.2). Schliesslich werden die Strukturebene (Kapitel 4.3) und die Prozessebene (Kapitel 4.4) des neuen WBK-Modells theoretisch fundiert.

4.1 Bisherige Definition der WBK: eine kritische Reflexion

Dieses Kapitel setzt sich kritisch mit der bisherigen Definition der WBK auseinander. Dabei wird auf die Begriffsbezeichnung (Kapitel 4.1.1) und auf die Domänenbeschreibung (Kapitel 4.1.2) eingegangen. Die bisherige Definition der WBK von Eberle (2015) lautet wie folgt:

«[Wirtschaftsbürgerliche Kompetenz ist] Wissen und Können, das notwendig ist, um private, wirtschaftspolitisch-volkswirtschaftliche sowie betriebswirtschaftliche Problemstellungen in staatlichen und nichtstaatlichen sozialen Systemen zu verstehen, vorgeschlagene Lösungen zu beurteilen und für einfachere Probleme selbst Lösungen zu entwickeln. Es orientiert sich an den in der Realität vorkommenden wirtschaftlichen Problemstellungen» (Eberle, 2015, S. 26).

Diese Definition der wirtschaftsbürgerlichen Kompetenz orientiert sich an einem kontextspezifischen kognitiven Kompetenzkonzept (vgl. Weinert, 2001a, S. 46) (vgl. Kapitel 3.2). Folglich werden in dieser Definition erforderliche kognitive Dispositionen («Wissen und Können»), kontextspezifische Anforderungssituationen («wirtschaftliche Problemstellungen») und Fähigkeitsaspekte bzw. kognitive Prozesse («Probleme verstehen», «Lösungen beurteilen», «Lösungen entwickeln») genannt.

Nichtsdestotrotz birgt diese Definition der WBK einige „Fallstricke“, insbesondere was die Bezeichnung der Kompetenz selbst und die Beschreibung ihrer Domäne angeht. Eine kritische Reflexion dieser bisherigen Definition soll Grundlage bieten für eine präzisere Beschreibung und eindeutigere Operationalisierung des Konstrukts (vgl. Kapitel 8) sowie für eine validere Interpretation

der Testergebnisse hinsichtlich des Konstrukts (vgl. Kapitel 11 und 12). An dieser Stelle sei auch angemerkt, dass die nachfolgende kritische Reflexion die Meinung der Autorin wiedergibt.

4.1.1 Kompetenzbezeichnung

Die Bezeichnung «wirtschaftsbürgerliche» Kompetenz ist m.E. mehrdeutig und daher missverständlich. Mit anderen Worten, sie ist einem intuitiven Verständnis der spezifischen Kompetenz nicht zuträglich. Die Mehrdeutigkeit bezieht sich auf die beiden Wortteile «wirtschafts» und «bürgerlich» sowie auf die Wortkombination «wirtschaftsbürgerlich», die in vielerlei Hinsicht auslegungsbedürftig sind.³¹

Erstens, «wirtschafts» verweist bei einer wörtlichen Auslegung auf *ökonomische* Anforderungssituationen, in denen die Menschen *ökonomisch handeln* sollen, also Fähigkeiten wie z.B. ökonomisch analysieren, ökonomisch beurteilen oder ökonomisch entscheiden gefragt sind. Dies mag eine berechtigte Anforderung an Experten, d.h. Ökonom*innen sein, kann jedoch keine Anforderung an Laien, d.h. an alle Menschen darstellen. Bei der teleologischen Auslegung von «wirtschafts» wäre es denn angebrachter von *ökonomisch geprägten* Anforderungssituationen auszugehen, in denen Fähigkeiten wie z.B. ökonomische Situationen analysieren oder ökonomische Entscheidungen begründet treffen gefragt sind (B. Weber, 2013, S. 9). Diese Auslegung (in Situationen ökonomisch handeln vs. in ökonomischen Situationen handeln) ist jedoch auch der deutschen Fachdidaktik für ökonomische Bildung umstritten (vgl. DEGÖB, 2009; Kaminski et al., 2017; Kaminski et al., 2008; Retzmann et al., 2010; G. Seeber et al., 2012; B. Weber, 2013).

Zweitens, «bürgerlich» bezieht sich bei einer wörtlichen und zeitgenössischen Auslegung auf Staatsbürger*innen, d.h. auf Menschen, die gemäss Verfassung mit politischen Rechten (z.B. Wählen, Abstimmen) ausgestattet sind. Somit sind alle minderjährigen Menschen (d.h. Schüler*innen bis/mit Sekundarstufe II) und alle Menschen ohne Schweizer Staatsbürgerschaft (ca. 20 % der Schweizer Wohnbevölkerung) ausgeschlossen. Ökonomisch geprägte Anforderungssituationen, v.a. diejenigen im persönlich-finanziellen und beruflich-unternehmerischen Bereich, betreffen jedoch *de facto* alle Menschen. Dubs (1985, S. 63ff; 2011, S. 194) fordert denn auch seit langem eine wirtschaftsbürgerliche Bildung für alle Menschen, «damit *jedermann* persönliche wirtschaftliche und wirtschaftspolitische, also berufsunabhängige, Probleme versteht und die vielen allgemeinen wirtschaftlichen Aspekte, welche *jedermann* persönlich betreffen und/oder in der öffentlichen

³¹ Friedrich Carl von Savigny (1779–1861) unterschied in der Jurisprudenz vier Auslegungselemente, sog. *canones* (lat. canon: Regel, Norm, Richtschnur) für Rechtsnormen. Diese Auslegungselemente finden in der heutigen Rechtsprechung im deutschsprachigen Raum noch immer Anwendung, jedoch in etwas abgeänderter Form (Forstmoser & Vogt, 2012, S. 565ff.): grammatikalisches Element (auch: Auslegung nach dem Wortlaut), systematisches Element (auch: Auslegung nach dem Zusammenhang mit anderen Rechtsnormen bzw. nach der inneren Gliederung des Rechtstextes), teleologisches Element (Auslegung nach dem Sinn und Zweck einer Rechtsnorm), realistisches Element (auch: Auslegung nach der Praktikabilität) und historisches Element (Auslegung nach der Entstehungsgeschichte).

Diskussion stehen, mit Sachverstand beurteilen kann» (Dubs, 2011, S. 194, Hervorhebungen der Autorin). Eberle (2015) seinerseits betont, dass «[w]irtschaftsbürgerlich kompetent zu sein, ist nicht nur für Kaufleute bedeutsam, sondern *für alle Menschen*» (Eberle, 2015, S. 26, Hervorhebung der Autorin).

Drittens, die Figur des «Wirtschaftsbürgers» wird in der fachdidaktischen Literatur unterschiedlich verstanden. Bei einigen Forschungsgruppen sind damit Menschen in verschiedenen Rollen gemeint, so z.B. Konsument*innen, Mitarbeiter*innen und Staatsbürger*innen, die sich mit verschiedenen «ökonomisch geprägten Lebenssituationen» bzw. «wirtschaftlichen Problemstellungen» auseinandersetzen (vgl. Albers, 1988, 1995; Dubs, 1985, 2011, 2013; Eberle, 2006b, 2015; Eberle et al., 2016). Hingegen bezeichnen andere Forschungsgruppen mit «Wirtschaftsbürger» nur Menschen in der Rolle als Bürger*innen und in ihrer Beziehung zum Staat, d.h. Menschen, die Steuern/Abgaben bezahlen, Sozialleistungen empfangen, politische Repräsentanten wählen und sich (freiwillig) in der Politik engagieren (vgl. Kamininski, 2017; Kaminski, 1996; Retzmann et al., 2010; G. Seeber et al., 2012; B. Weber, 2013).

4.1.2 Domänenbeschreibung

Die Beschreibung der Domäne der «wirtschaftsbürgerlichen» Kompetenz ist in der bisherigen Definition m.E. zu vage und zu breit gefasst, um daraus ein eindeutig operationalisierbares und valide interpretierbares Konstrukt zu generieren. Zwar werden in der Beschreibung der Inhaltsbereich («wirtschaftliche Problemstellungen») und der Kognitionsbereich («verstehen», «beurteilen», «entwickeln») benannt, doch bleibt v.a. der Inhaltsbereich auslegungsbedürftig.

Erstens ist unklar, was mit «wirtschaftlichen Problemstellungen» gemeint ist. Im engeren Sinne kann es sich hierbei um ökonomische Kategorien i.S.v. Denkmustern handeln (vgl. Dauenhauer, 2001; Kruber, 1995, 2000; May, 1978) (vgl. Kapitel 2.1.3): Knappheit (Ressourcen, Bedürfnisse), Kosten/Nutzen-Überlegungen, Arbeitsteilung, Marktkoordination, Wirtschaftskreislauf, Interdependenzen/Zielkonflikte, Instabilität (z.B. Inflation, Arbeitslosigkeit), soziale Ungleichheiten/ökologische Belastungen, Staatseingriffe, Interessenkonflikte, Wertbezug (Freiheit, Gerechtigkeit, Sicherheit), Wirtschaftsordnung. Im weiteren Sinne können Probleme gemeint sein, die ihren Ursprung im ökonomischen System haben und Konsequenzen in anderen Systemen, insbesondere im politischen System, nach sich ziehen, und *vice versa* (Kruber, 2005, S. 78ff.). Dies ist insbesondere bei «wirtschaftspolitisch-volkswirtschaftlichen Problemstellungen» offensichtlich: Eine wirtschaftspolitische Problemstellung (z.B. Erhöhung der Krankenversicherungsprämien aufgrund steigender Kosten im Gesundheitsbereich) hat sozialpolitische Konsequenzen (z.B. staatliche Prämienverbilligung für Personen in wirtschaftliche bescheidenen Verhältnissen); eine umweltpolitische Problemstellung (z.B. Reduktion der CO₂-Emissionen für den Klimaschutz) hat wirtschaftspolitische Konsequenzen (z.B. verursachergerechte Abgabe auf CO₂-Emissionen).

Zweitens ist unklar, welche «privaten» und «betriebswirtschaftlichen» Problemstellungen (Eberle, 2015, S. 26) gemeint sind. Sofern es sich bei «privat» um persönlich handelt,³² können diese Problemstellungen im finanziellen (z.B. Budget, Schulden, Versicherungen, Vorsorge), im sozialen (z.B. Konflikte am Arbeitsplatz) oder auch im physisch-psychischen (z.B. Erwerbsunfähigkeit, Erwerbslosigkeit) Bereich auftauchen; sie müssen nicht zwingend „wirtschaftlich“ sein. Bei betriebswirtschaftlichen Problemstellungen kommt entweder eine „Aussenbetrachtung“ i.S.v. berufsallgemeines Wissen über Strukturen und Prozesse in Unternehmen in Frage (z.B. die Organisation und Logistik in der eigenen Abteilung und im Unternehmen verstehen) oder aber eine „Innenbetrachtung“ i.S.v. berufsspezifisches Wissen und Können über konkrete betriebliche und unternehmerische Entscheidungen (z.B. Finanzkennzahlen aufbereiten und Investitionsentscheide vorbereiten für die Geschäftsleitung des Unternehmens). Die Domäne der WBK und die sich darin befindlichen Problemstellungen werden bei Eberle (2015; Eberle et al., 2016) nur fragmentarisch und exemplarisch skizziert. Bemerkenswerterweise wird auch nur die Bedeutung der «wirtschaftspolitisch-volkswirtschaftlichen» Problemstellungen angedeutet, jene der «privaten» und «betriebswirtschaftlichen» jedoch ausgeklammert.

«[Bei der wirtschaftsbürgerlichen Kompetenz] geht es im Besonderen auch um die Wahrnehmung der demokratisch-politischen Rechte und Pflichten der Bürgerinnen und Bürger. Ein aktuelles Beispiel hierfür ist das Verstehen und die Beurteilung des Vorgehens der Schweizer Nationalbank hinsichtlich des Wechselkurses des Schweizer Franken zum Euro» (Eberle, 2015, S. 26f.).

«In der stark direkt-demokratisch verfassten Schweiz kommt dieser [umfassenden wirtschaftsbürgerlichen] Kompetenz eine noch größere Bedeutung zu als in ausschließlich repräsentativen (parlamentarischen) Demokratien, da die Bürgerinnen und Bürger hier das Recht haben, in Volksabstimmungen mit ihren Entscheidungen wirtschaftliche und gesellschaftliche Entwicklungen mitzubestimmen sowie eigene Volksinitiativen einzureichen. So konnte beispielsweise im Jahr 2015 unter anderem über Volksinitiativen zur Umwandlung der Mehrwertsteuer in eine Energiesteuer, zur Einführung einer gesamtschweizerischen Erbschaftssteuer und zur reduzierten Besteuerung von Familien abgestimmt werden» (Eberle et al., 2016, S. 94).

Drittens ist unklar, was mit «sozialen Systemen» gemeint ist, in denen die «wirtschaftlichen Problemstellungen» auftreten und inwiefern diese Systeme wirtschaftliche Problemstellungen charakterisieren und differenzieren. Diese Systeme können z.B. als Trägerschaft (öffentlich-rechtlich, privatrechtlich), als Wirtschaftseinheiten (Haushalte, Unternehmen, Staat), als Wirtschaftssektoren/-

³² Die Autorin versteht „privat“ als juristisches Pendant zu „öffentlich“ bzw. „staatlich“.

branchen (Agrarwirtschaft, Handwerk, Industrieproduktion, Dienstleistungen etc.) oder als Lebensbereiche (z.B. Familie/Freunde, Beruf/Arbeit) verstanden werden.

4.2 Modifizierte Definition der WBK und neues WBK-Modell

In diesem Kapitel wird die Definition der WBK präzisiert und ausgehend davon ein neues Rahmenmodell für die WBK entworfen. Die Bezeichnung «wirtschaftsbürgerliche» Kompetenz wird – trotz der oben reklamierten Unschärfe (vgl. Kapitel 4.1.1) – aus forschungsstrategischen Gründen beibehalten. Dies damit die vorliegende Dissertation in der Forschungsgemeinschaft als anknüpfend an frühere theoretische und empirische Arbeiten zur wirtschaftsbürgerlichen Bildung und wirtschaftsbürgerlichen Kompetenz gelesen wird (vgl. z.B. Dubs, 2011, 2013; Eberle, 2015; Eberle et al., 2016; Schumann et al., 2017). Die Beschreibung der Domäne wird konkretisiert, um den oben reklamierten mehrdeutigen Auslegungen für die Reichweite der Domäne entgegenzutreten (vgl. Kapitel 4.1.2). Eine solche Beschreibung soll das theoretische Konstrukt «wirtschaftsbürgerliche» Kompetenz für eine (bessere) empirische Untersuchung zugänglich, d.h. eindeutig operationalisierbar und valide interpretierbar machen (vgl. Kapitel 8, 9, 11 und 12).

Im Sinne einer „Arbeitsdefinition“ wird wirtschaftsbürgerliche Kompetenz (WBK) in dieser Dissertation wie folgt verstanden:

Wirtschaftsbürgerliche Kompetenz bezieht sich auf ökonomisch geprägte Anforderungssituationen im persönlich-finanziellen, im beruflich-unternehmerischen und im gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereich. Die Bewältigung dieser Anforderungssituationen erfolgt durch kognitive Prozesse der Informationsverarbeitung und des Problemlösens unter Einbezug individueller Dispositionen. Sie ist im Ergebnis beobachtbar durch das standardisierte Erfassen von fundiertem Verstehen, systematischem Vergleichen, differenziertem Beurteilen und begründetem Entscheiden.

Zu dieser modifizierten Definition der WBK gibt es folgende Punkte anzumerken:

- Sie orientiert sich an einem kontinuierlichen Kompetenzkonzept (vgl. Blömeke et al., 2015) (vgl. Kapitel 3.1) und an einem kontextspezifischen kognitiven Kompetenzkonzept (vgl. Weinert, 2001a, S. 46) (vgl. Kapitel 3.2). Somit ist sie konzeptionell vergleichbar mit der bisherigen Definition der WBK (vgl. Eberle, 2015, S. 26) (vgl. Kapitel 4.1). Nichtsdestotrotz muss das „kontextspezifische“ und das „kognitive“ dieser modifizierten Definition der WBK detailliert beschrieben sowie theoretisch und empirisch begründet werden. Eine solche Beschreibung und Begründung werden in dieser Dissertation gegeben (vgl. Kapitel 4.3 und 4.4).
- Sie orientiert sich an einer ökonomischen Allgemeinbildung (vgl. Albers, 1988, 1995; Kaminski, 1996; Kaminski et al., 2017; May & Albers, 2008; G. Seeber et al., 2012) bzw. an der Bildung

des allgemeinen Wirtschafts- und Gesellschaftsverständnisses in ihrem weiteren Sinne (vgl. Dubs, 2011, 2013; Eberle, 2015; Ulrich, 1993, 2001) (vgl. Kapitel 2.1.1 und 2.2.1).

- Sie orientiert am Situationsprinzip der ökonomischen Bildung (vgl. Albers, 1995; Kaminski et al., 2017, S. 85ff.; Ochs & Steinmann, 1978; Reetz, 1984, 2003; Steinmann, 1997; B. Weber, 2005, 2013) (vgl. Kapitel 2.1.2) und hat eine themenlogische Binnenstruktur: Ökonomisch geprägte Anforderungssituationen entstehen in verschiedenen Lebensbereichen (vgl. Retzmann et al., 2010; G. Seeber et al., 2012; B. Weber, 2005, 2013) (vgl. Kapitel 2.1.1). Im Gegensatz dazu sind die früheren Konzepte der *economic literacy* bzw. der ökonomischen Kompetenzen (vgl. K. Beck, 1989, 1993) (vgl. Kapitel 3.4.1) an einem Wissenschaftsprinzip der ökonomischen Bildung ausgerichtet und haben eine fachlogische Binnenstruktur: Ökonomische Problemstellungen ergeben sich aus den Fachbereichen Volkswirtschaft bzw. Volkswirtschaft und Betriebswirtschaft.
- Sie bezieht sich auf kognitive Prozesse der Informationsverarbeitung und des Problemlösens (vgl. Bloom et al., 1956; Dörner, 1979; Edelman & Wittmann, 2012; Krathwohl, 2002; Marzano & Kendall, 2007).
- Sie adressiert explizit fachliche und methodische Kompetenz, implizit jedoch auch soziale und personale Kompetenz (vgl. Roth, 1971). Insofern ist sie als ganzheitliche Handlungskompetenz zu verstehen, die sich nicht von Prozessen der Sozialisation, Erziehung und Persönlichkeitsentwicklung loslösen lässt.
- Sie ist für Jugendliche auf der gesamten Sekundarstufe II anwendbar, d.h. für die Allgemeinbildung (z.B. gymnasiale Maturitätsschulen, Fachmaturitätsschulen) und für den allgemeinbildenden Teil der Berufsbildung (z.B. Berufsfachschulen, Berufsmaturitätsschulen). Entsprechend werden als Zielgruppe für den wirtschaftsbürgerlichen Kompetenztest (WBK-T2) Schüler*innen am Ende der Sekundarstufe II festgelegt (vgl. Kapitel 8.1.3). Im Gegensatz dazu ist das Modell der Wirtschaftskompetenz explizit auf die kaufmännische Berufsbildung ausgerichtet und beinhaltet eine wirtschaftsberufliche/kaufmännische und eine wirtschaftsbürgerliche Domäne (vgl. Eberle, 2015; Holtsch & Eberle, 2018) (vgl. Kapitel 3.4.3).

Ausgehend von dieser modifizierten Definition der WBK wird im Rahmen dieser Dissertation ein neues Rahmenmodell für die WBK entworfen. Das neue WBK-Modell umfasst eine Strukturebene mit Anforderungssituationen in den Lebensbereichen (inhaltliche Aspekte der Domäne) und eine Prozessebene mit Kognitionsprozessen bezüglich dieser Anforderungssituationen (kognitive Aspekte der Domäne) (vgl. Abbildung 4-1).

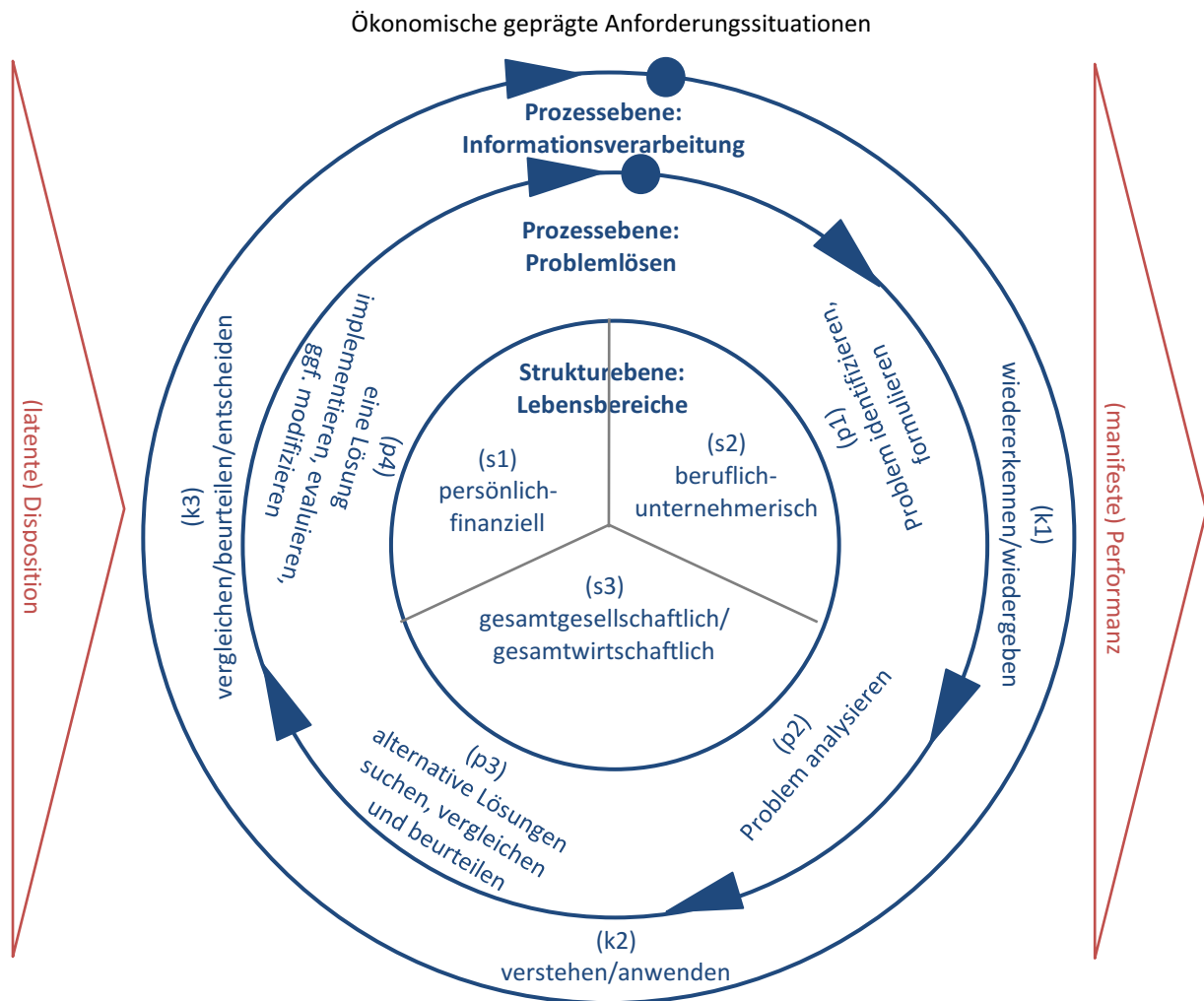


Abbildung 4-1. WBK-Modell für ökonomisch geprägte Anforderungssituationen mit Lebensbereichen und Kognitionsprozessen

Die *Strukturebene* des WBK-Modells referenziert auf das „kontextspezifische“ der Kompetenzdefinition. Sie umfasst *sektoriell* strukturierte Inhaltsbereiche, nämlich die drei Lebensbereiche: (s1) der persönlich-finanzielle Lebensbereich, (s2) der beruflich-unternehmerische Lebensbereich, (s3) der gesamtgesellschaftliche/gesamtwirtschaftliche Lebensbereich (vgl. Kapitel 4.3). In jedem Lebensbereich des WBK-Modells finden sich domänenspezifische, eben ökonomisch geprägte Anforderungssituationen, die es von den Individuen zu bewältigen gilt.

Die *Prozessebene* des WBK-Modells referenziert auf das „kognitive“ der Kompetenzdefinition. Sie umfasst sowohl *hierarchisch* wie auch *zirkulär* strukturierte Kognitionsprozesse bezüglich der Bewältigung der domänenspezifischen Anforderungssituationen: Informationsverarbeitung und Problemlösen (vgl. Kapitel 4.4). Die *Informationsverarbeitung* wird durch drei *hierarchische* Stufen modelliert (vgl. Marzano & Kendall, 2007): (k1) Informationserinnerung (wiedererkennen/wiedergeben), (k2) Informationsverarbeitung (verstehen/anwenden) und (k3) Informationsnutzung (vergleichen/evaluieren/entscheiden). Hierarchisch ist der Prozess der

Informationsverarbeitung, da die Stufen als Anforderungsniveaus gedacht sind; für die Stufe (k3) ist die Stufe (k2) notwendig, für die Stufe (k2) die Stufe (k1).

Das *Problemlösen* wird durch vier *zirkuläre* und *iterative* Phasen modelliert (vgl. Edelman & Wittmann, 2012, S. 190f.; Gollob, Graf-Zumsteg, Bachmann, Gattiker, & Ziegler, 2007; Massing, 1999): (p1) Problem identifizieren und formulieren, (p2) Problemsituation multiperspektiv analysieren; (p3) alternative Lösungsansätze suchen, vergleichen und beurteilen; (p4) eine Lösung implementieren, evaluieren und ggf. modifizieren. Zirkulär und iterativ ist der Problemlöseprozess, da er als heuristische Spirale gedacht ist. Aus der Evaluation in Phase (p4) kann wiederum ein neues Problem entstehen (p1), so dass der Zirkel von vorne beginnt.

Das WBK-Modell verdeutlicht, dass ökonomisch geprägte Anforderungssituationen in den drei Lebensbereichen grundsätzlich gleichartige Kognitionsprozesse erfordern. Jedoch variiert die Extensität und Intensität der Kognitionsprozesse je nach Lebensbereich und Anforderungssituation. So hat bspw. die Phase 4 des Problemlöseprozesses für Anforderungssituationen im persönlich-finanziellen und beruflich-unternehmerischen Lebensbereich durchaus eine grosse Bedeutung, jedoch eine geringere im gesamtwirtschaftlichen/gesamtgesellschaftlichen Lebensbereich – zumindest *für das individuelle Handeln*, nicht aber für das kollektive Handeln.

Nachfolgend wird die im WBK-Modell aufgespannte Strukturebene, d.h. die Lebensbereiche (Kapitel 4.3) und Prozessebene, d.h. die Kognitionsprozesse (Kapitel 4.4) beschrieben sowie theoretisch und empirisch begründet. Dabei wird Hinblick auf die Operationalisierung des Konstrukts sozioökonomische Facette der WBK (vgl. Kapitel 8) auf sozioökonomische Problemsituationen fokussiert.

4.3 Strukturebene des WBK-Modells: Lebensbereiche

In diesem Kapitel werden zunächst die drei Lebensbereiche des WBK-Modells beschrieben und exemplarisch für jede Lebenssituation ökonomisch geprägte Anforderungssituationen benannt (Kapitel 4.3.1). Danach wird der gesamtgesellschaftliche/gesamtwirtschaftliche Lebensbereich an seiner Schnitzzustelle Politik–Ökonomie beleuchtet und konkrete sozioökonomische Problemsituationen mittels einer umfassenden Domänenanalyse identifiziert (Kapitel 4.3.2).

4.3.1 Die drei Lebensbereiche

Ökonomisch geprägte Anforderungssituationen lassen sich über das in der Wirtschaftsdidaktik gängige *Rollenkonzept* zu gleichen oder zumindest zu ähnlichen Anforderungen bündeln. Rollenkonzepte strukturieren nicht die Fachwissenschaft, sondern die konkrete Lebenswelt aus unterschiedlichen Perspektiven. Für das Rollenkonzept ist allerdings nicht die Vollständigkeit der Anforderungssituationen sondern ihre Bedeutsamkeit entscheidend (Retzmann et al., 2010, S. 14f.; G. Seeber et al., 2012, S. 73ff., S. 87).

Ökonomisch geprägte Anforderungssituationen finden sich folglich in drei Lebensbereichen, in denen Menschen in unterschiedlichen Rollen agieren bzw. interagieren (vgl. Eberle, 2015; Retzmann et al., 2010; G. Seeber et al., 2012; B. Weber, 2005, 2013) (vgl. Tabelle 4-1): der persönlich-finanzielle Lebensbereich, der beruflich-unternehmerische Lebensbereich, der gesamtgesellschaftliche/gesamtwirtschaftliche Lebensbereich. Diese Lebensbereiche sind sektoriell zu verstehen, obschon zwischen ihnen Interdependenzen bestehen. Auch tritt ein und derselbe Mensch in verschiedenen Rollen auf bzw. wechselt seine Rolle je nach Situation.

Im *persönlich-finanziellen Lebensbereich* (s1) befinden sich Menschen in der Rolle als Verbraucher*in, genauer als Konsument*in, Sparer*in, Anleger*in, Kreditnehmer*in, Versicherungsnehmer*in. Hier stellen sich ihnen z.B. Anforderungssituationen des Umgangs mit Geld, des nachhaltigen Konsums, des langfristigen Sparens und der privaten Vorsorge (Kaminski et al., 2017, S. 160ff.; Retzmann et al., 2010, S. 15; G. Seeber et al., 2012, S. 87). Zielkonflikte drehen sich z.B. um Gegenwart (Konsum) vs. Zukunft (Sparen/Vorsorge). Persönlich-finanzielle Anforderungssituationen entsprechen m.E. «private[n] Problemstellungen» (Eberle, 2015, S. 26) in der bisherigen Definition der WBK bzw. «persönliche[n] wirtschaftliche[n] Problemen» (Dubs, 2011, S. 194) in der Beschreibung der wirtschaftsbürgerlichen Bildung.

Im *beruflich-unternehmerischen Lebensbereich* (s2) haben Menschen zum einen die Rolle als Arbeitnehmer*in (z.B. als Angestellte*r in einem Produktions-, Handels- oder Dienstleistungsbetrieb). Hier zeigen sich *berufsallgemeine* Anforderungssituationen bei der Berufs- und Studienwahl, bei der Aus- und Weiterbildung, bei der beruflichen Laufbahn („Karriere“) etc. (Kaminski et al., 2017, S. 144ff.; Retzmann et al., 2010, S. 15; G. Seeber et al., 2012, S. 87). Zielkonflikte ergeben sich z.B. bei Karriereplanung vs. Familienplanung, finanzielle Sicherheit vs. kreative Entfaltung. Zum anderen haben Menschen die Rolle als Unternehmer*in (z.B. als Geschäftsinhaber*in bzw. Geschäftsführer*in eines Handwerksbetriebs oder eines Verkaufsladens), wo sich *unternehmensbezogene* Anforderungssituationen bei der administrativen, operativen und strategischen Unternehmensführung stellen.³³ Zielkonflikte entstehen z.B. in den verschiedenen sich überlappenden Umweltsphären des Unternehmens (ökonomisch, sozial, ökologisch, technologisch) sowie zwischen den gegensätzlichen Interessen der Anspruchsgruppen des Unternehmens (Mitarbeitende, Kunden, Lieferanten, Eigenkapitalgeber/Shareholders, Fremdkapitalgeber/Banken, Staat, NGO's). Unternehmensbezogene Anforderungssituationen decken sich m.E. mit «betriebswirtschaftliche[n] Problemstellungen» (Eberle, 2015, S. 26) in der bisherigen Definition der WBK. Hingegen sind berufsallgemeine Anforderungssituationen eher deckungsgleich mit «private[n] Problemstellungen» (Eberle, 2015, S. 26) in der bisherigen Definition der WBK bzw. mit

³³ Eine übersichtliche und verständliche Darstellung von unternehmerischen Strukturen, Prozessen und Zielkonflikten bietet z.B. das St.Galler Management-Modell (Capaul & Steingruber, 2013).

«persönliche[n] wirtschaftliche[n] Problemen» (Dubs, 2011, S. 194) in der Beschreibung der wirtschaftsbürgerlichen Bildung.

An dieser Stelle sei betont, dass es sich hierbei um *berufsallgemeine* bzw. *berufsübergreifende* Situationen als Arbeitnehmer*in oder Unternehmer*in handelt (vgl. Dubs, 1985, S. 63ff.; Dubs, 2011, S. 194). Es sind also explizit *keine berufsspezifischen* Anforderungssituationen gemeint, für die berufliche Handlungskompetenzen erforderlich sind. Berufliche Handlungskompetenzen und/oder berufsfachliche Kompetenzen werden typischerweise in beruflichen Kompetenzmodellen charakterisiert (vgl. z.B. K. Beck et al., 2016; Holtsch & Eberle, 2018; Nickolaus et al., 2008; Winther, 2010). Berufsspezifische Anforderungssituationen im kaufmännischen Bereich sind z.B. Material bewirtschaften, Kunden beraten, Aufträge abwickeln (z.B. Bestellungen, Offerten), finanzielle Informationen verarbeiten (z.B. Buchungen) (SKKAB, 2011a).

Im *gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereich* (s3) schliesslich haben Menschen die Rolle als Staatsbürger*in oder Citoyen-ne, wo sie mit sozioökonomischen Anforderungssituationen konfrontiert sind, d.h. wirtschafts-, sozial- und umweltpolitische Fragestellungen (Kaminski et al., 2017, S. 167ff.; Retzmann et al., 2010, S. 15; G. Seeber et al., 2012, S. 87f.). Konkrete sozioökonomische Anforderungssituationen ergeben sich in verschiedenen Politikfeldern, z.B. Arbeit, Energie & Umwelt, Finanzen & Steuern, Gesundheit, soziale Sicherheit & Vorsorge, Migration & Integration, Verkehr. Aktuelle Beispiele für solche sozioökonomischen Problemsituationen sind z.B. die Krise der internationalen Finanzwirtschaft („Finanzkrise“), die Krise der europäischen Währungspolitik („Euro-Krise“) oder die Krise der europäischen Flüchtlingspolitik („Flüchtlingskrise“). Hier ergeben sich Wertekonflikte, wie z.B. Effizienz, Freiheit, Sicherheit, Gerechtigkeit, Nachhaltigkeit etc. Sozioökonomische Anforderungssituationen korrespondieren m.E. mit «wirtschaftspolitisch-volkswirtschaftliche[n] Problemstellungen» (Eberle, 2015, S. 26) in der bisherigen Definition der WBK bzw. mit «wirtschaftspolitische[n] Problemen» (Dubs, 2011, S. 194) in der Beschreibung der wirtschaftsbürgerlichen Bildung.

Die Rolle als Staatsbürger*in kann mehr oder weniger aktiv ausgestaltet sein, z.B. als gewähltes Mitglied der Exekutive oder Legislative, als Mitglied einer politischen Partei oder einer anderen Interessengruppe – oder schlicht als Abstimmende*r. Weitere Rollen der Menschen im gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereich sind Transferempfänger*in (z.B. Arbeitslosengeld, Sozialhilfe) oder Beitragszahler*in (z.B. Einkommensteuer, AHV/IV-Beiträge) (Retzmann et al., 2010, S. 15; G. Seeber et al., 2012, S. 87f.).

Tabelle 4-1. Strukturebene des WBK-Modells: Inhaltsbereiche und domänenspezifische Anforderungssituationen

Inhaltsbereich	Domänenspezifische Anforderungssituationen: Beschreibung und Beispiele
(s1) persönlich- finanzieller Lebensbereich	▪ Persönlich-finanzielle Anforderungssituationen ▪ Rolle als Verbraucher, d.h. Konsument*in, Sparer*in/Anleger*in, Kreditnehmer*in, Versicherungsnehmer*in ▪ Problemstellungen aus dem Alltag: Geld ausgeben (Konsum), Geld sparen/anlegen, Geld verdienen (Einkommen), Verschuldung vermeiden, private Vorsorge regeln (z.B. Vorsorgekonto, Lebensversicherung, Immobilien), Verträge abschliessen (Arbeit, Miete, Konsumkredit, Versicherung), Steuern deklarieren (Steuererklärung) ▪ Ziele und Zielkonflikte: Gegenwart (Konsum) vs. Zukunft (Sparen/Vorsorge) ▪ Fachwissenschaftlicher Bezug: financial literacy <i>Bsp.: Budget erstellen, Konsumkredit abschliessen, private Vorsorge regeln, Steuererklärung ausfüllen</i>
(s2) beruflich- unternehmerischer Lebensbereich	▪ Berufsallgemeine Anforderungssituationen (nicht: berufsspezifische Anforderungssituationen!) ▪ Rolle als Arbeitnehmer*in, d.h. Auszubildende*r oder Mitarbeiter*in ▪ Problemstellungen zur beruflichen Qualifikation ▪ Ziele und Zielkonflikte: Karriereplanung vs. Familienplanung, finanzielle Sicherheit vs. kreative Entfaltung <i>Bsp.: Berufs-/Studienwahl, Fortbildungen, Umschulung, Anstellung vs. Selbständigkeit</i> ▪ Unternehmensbezogene Anforderungssituationen (vgl. z.B. das St.Galler Management-Modell) ▪ Rolle als Unternehmer*in, d.h. Geschäftsinhaber*in oder Geschäftsführer*in ▪ Problemstellungen der administrativen und operativen Unternehmensführung, Wertschöpfungsprozess: Beschaffung/Einkauf (Make-or-Buy), Produktion (Kosten vs. Qualität, Fokussieren vs. Diversifizieren), Vertrieb/Verkauf (Marketing), Entsorgung; Finanzen (Eigenkapital vs. Fremdkapital), Personal (Gehaltskomponenten, Sozialleistungen, Produktivitäts- und Motivationsanreize, Sozialpartnerschaft) ▪ Problemstellungen der strategischen Unternehmensführung: Gewinn/Rentabilität, Corporate Social Responsibility (CSR), Ökologie/Nachhaltigkeit ▪ Ziele und Zielkonflikte: Umweltsphären (ökonomisch, sozial, ökologisch, technologisch), Anspruchsgruppen (Mitarbeitende, Kunden, Lieferanten, Eigenkapitalgeber/Shareholders, Fremdkapitalgeber/Banken, Staat, NGO's) ▪ Fachwissenschaftlicher Bezug: Betriebswirtschaftslehre, Entrepreneurship, New-Public-Management <i>Bsp.: Make-or-Buy, Kosten vs. Qualität, Fokussieren vs. Diversifizieren</i>
(s3) gesellschaftlicher/ wirtschaftlicher Lebensbereich	▪ Sozioökonomische Anforderungssituationen ▪ Rolle als Staatsbürger*in, d.h. Gesellschafts- und Wirtschaftsbürger*in ▪ Problemstellungen aus verschiedenen Politikfeldern, z.B. Arbeit, Energie & Umwelt, Finanzen & Steuern, Gesundheit, soziale Sicherheit & Vorsorge, Verkehr ▪ Werte und Wertekonflikte: Effizienz, Freiheit, Sicherheit, Gerechtigkeit, Nachhaltigkeit etc. ▪ Fachwissenschaftlicher Bezug: Volkswirtschaftslehre <i>Bsp.: Finanzierung der Altersvorsorge, Reduktion der Staatsverschuldung, Lenkung des Verkehrsaufkommens, Reduktion der Umweltbelastung</i>

Quelle: Eigene Entwicklung, i.A.a. Albers (1995, S. 3), Kaminski et al. (2017, S. 37), Steinmann (1997) und B. Weber (2005, S. 43; 2014, S. 9).

Ökonomisch geprägte Anforderungssituationen in all diesen Lebensbereichen charakterisieren sich durch *Abstraktheit* und *Komplexität* (Eberle, 2006b, S. 18f.; 2015, S. 19). Dass ökonomische Zustände und Entwicklungen abstrakt sind, zeigt sich z.B. angesichts derivater Finanzprodukte (z.B. Hedgefonds), multinationaler Unternehmen, globaler Märkte (z.B. Finanzmarkt), digitaler Kommunikations- und Informationssysteme etc. Dass ökonomische Problemstellungen komplex sind, zeigt sich dadurch, dass Informationen zur Analyse- und Entscheidungsgrundlage z.T. unverständlich, unzuverlässig und/oder einseitig sind, dass es keine „korrekte“ oder „perfekte“ Lösung, sondern mehrerer Lösungsvarianten gibt, und dass für eine *Second-Best-Solution* das Abwägen von Vorteilen und Nachteilen verschiedener Lösungsvarianten erforderlich ist.

Persönlich-finanzielle Anforderungssituationen werden in der Bildungsforschung und Bildungspolitik häufig mit *financial literacy* oder persönlicher Finanzkompetenz verknüpft. *Financial literacy education* steht seit ca. 15 Jahren (vgl. z.B. OECD, 2005a, 2005c), spätestens aber seit der Finanzkrise 2008/2009 auf der bildungspolitischen Agenda. Sie wird hauptsächlich für Schüler*innen auf der Sekundarstufe I von diversen Akteuren gefordert und gefördert (vgl. Ackermann & Eberle, 2016; Ackermann, Ruoss, & Flury, 2018; Pinto, 2013). Mitunter findet sie Eingang in neue Bildungspläne, so z.B. in den Schweizer Lehrplan 21 (D-EDK, 2014) und produziert eine Flut von neuen Lehrmitteln und (digitalen) Lernmaterialien. In der Bildungsforschung wurden für *financial literacy* einige theoretische Rahmenmodelle entwickelt und entsprechende empirischen Untersuchungen durchgeführt (vgl. z.B. Aprea, 2014; Aprea & Wuttke, 2016; Aprea et al., 2015; Brown & Graf, 2013; CEE, 2013; Davies, 2015; Lusardi & Mitchell, 2014; OECD, 2012, 2014a, 2017b; Rudeloff, 2019; Schürkmann, 2017; Schürkmann & Schuhen, 2013). Im „PISA Financial Literacy Framework“ der OECD werden für *financial literacy* vier Inhaltsbereiche spezifiziert und für das „PISA Financial Literacy Framework Assessment“ operationalisiert (OECD, 2017a, S. 146ff.): 1) *money and transactions*, 2) *planning and managing finances*, 3) *risk and reward* und 4) *financial landscape*. In den „National Standards for Financial Literacy“ des *Council for Economic Education* (CEE) werden für *financial literacy* sechs Standards i.S.v. übergeordneten Inhaltsbereichen spezifiziert (CEE, 2013, S. ix, S. 1ff.): 1) *earning income*, 2) *buying goods and services*, 3) *saving*, 4) *using credit*, 5) *financial investing* und 6) *protecting and insuring*. Für die curriculare Einbettung wurden für jeden Standard Kompetenzen formuliert, und zwar für drei verschiedene Altersgruppen (K4, K8, K12). Für die empirische Testung wurde basierend auf den Standards der *Test of Financial Literacy* (TFL) entwickelt (Walstad & Rebeck, 2017). Der TFL wurde mittlerweile für Deutschland adaptiert (Förster, Happ, & Molerov, 2017) und erste Studien damit durchgeführt (Förster, Happ, & Walstad, 2019). In einer neueren Arbeit zur Finanzkompetenz von Lernenden am Ende der Sekundarstufe I wird ein Kompetenzmodell mit fünf Inhaltsbereichen entwickelt und untersucht (Rudeloff, 2019, S. 111ff.): 1) Geld und Zahlungsverkehr, 2) Sparen, 3) Kredite, 4) Versicherungen, 5) Geldpolitik.

Unternehmensbezogene Anforderungssituationen werden in der Bildungsforschung häufig mit *entrepreneurship competence* verknüpft und sind mittlerweile vermehrt Gegenstand empirischer Untersuchungen. Einige Studien zur *entrepreneurship education* befassen sich mit konzeptioneller Grundlegung und empirischer Untersuchung verschiedener unternehmerischer Kompetenzen, z.B. Unternehmensgründung, Krisenbewältigung, Erfolgs- und Versagenserlebnisse (vgl. z.B. Dinh, Haves, & Retzmann, 2018; Morris, 2013; Oser & Volery, 2012; Retzmann & Hausmann, 2012; S. Weber, Oser, Achtenhagen, Fretschner, & Trost, 2014). In anderen Studien werden Lehr-Lern-Konzepte zur Förderung der unternehmerischen Handlungskompetenz entwickelt und erprobt (vgl. z.B. Bijedic, 2013; Riklin, in progress).

Sozioökonomische Anforderungssituationen fanden bislang in der Bildungsforschung wenig Beachtung. Einige wenige Studien untersuchen wirtschaftsbürgerliche Kompetenzen und damit hauptsächlich sozioökonomische Problemsituationen (vgl. z.B. Eberle et al., 2016; Schumann et al., 2017; Siegfried & Ackermann, under review; Sticca, Rohr-Mentele, & Forster-Heinzer, 2018). Jedoch gibt es mehrere Studien zur *economic literacy*, also zum „volkswirtschaftlichen“ Bereich der ökonomischen Bildung und die damit verbundene volkswirtschaftliche Kompetenz (vgl. z.B. K. Beck, 1989, 1993; Brandlmaier et al., 2006; Freundlinger, 1992; Katschnig & Hanisch, 2005; Lehmann & Seeber, 2007; Schumann & Eberle, 2014a). Entsprechend existiert noch keine wissenschaftlich fundierte Analyse und detaillierte Beschreibung des gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereichs und der sich darin befindlichen sozioökonomischen Anforderungssituationen. Eine eher fragmentarische und exemplarische Skizzierung der gesamten Domäne der wirtschaftsbürgerlichen Kompetenz findet sich bei Eberle (2015; Eberle et al., 2016).

4.3.2 Der gesamtgesellschaftliche/gesamtwirtschaftliche Lebensbereich

Anforderungssituationen im gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereich, kurz sozioökonomische Problemsituationen (vgl. Tabelle 4-1) ergeben sich in der Schnittmenge von Politik und Ökonomie.

«Wirtschaft und Politik stellen in der Realität nicht zwei getrennte Lebensbereiche dar, sondern sind aufs Engste miteinander verwoben und interdependent. [...] dass politische Entscheidungen die wirtschaftliche Situation des Individuums beeinflussen [...] bzw. wirtschaftliche Probleme in der Politik thematisiert werden [...]» (Kruber, 2005, S. 78).

Politische Entscheidungen wirken sich auf ökonomische Bereiche (Akteure, Sektoren/Branchen) aus und umgekehrt beeinflussen ökonomische Entwicklungen (z.B. Konjunkturschwankungen) die politischen Gestaltungsmöglichkeiten. Die gesamtwirtschaftliche Situation eines Landes (z.B. Arbeitslosigkeit, Inflation, Staatsverschuldung) beeinflusst wiederum die

Einstellung der Menschen zu politischen Entscheidungsträgern sowie zu politischen Lösungsvorschlägen (z.B. Rentenreform, Atomausstieg, Schuldenbremse) und schliesslich ihr Wahl- und Abstimmungsverhalten (Kruber, 2005, S. 78ff.).

Die Beschreibung des gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereichs und der sich in ihm befindlichen sozioökonomischen Anforderungssituationen wird aufgrund dieser Interdependenzen zwischen Ökonomie und Politik nachfolgend in drei argumentationslogischen Schritten angegangen. Als Erstes wird das Modell des *Politikzyklus* erläutert, in dem sich der politische Problemlöseprozess von komplexen und kontroversen *gesellschaftlichen Problemen* widerspiegelt. Als Zweites wird die Bedeutung der *gesellschaftlichen Verantwortung* und *politischen Teilhabe* des Einzelnen im Schweizer Kontext dargelegt. Als Drittes werden für den Schweizer Kontext Politikfelder und sozioökonomischen Problemsituationen identifiziert.

4.3.2.a Politikbegriff und Politikzyklus

Der Begriff „Politik“ ist schwierig zu definieren und die Domäne „Politik“ schwierig zu fassen, da Politik – ebenso wie Ökonomie (vgl. Kapitel 2.1.3) – ein vielschichtiges, an seinen Randbereichen unscharfes Phänomen ist. Politik erfasst «potentiell alle Ebenen menschlichen Zusammenlebens [...]. Politik ist also ubiquitär» (Weißeno, Detjen, Juchler, Massing, & Richter, 2010, S. 30). Sie ist «die potentielle Disposition über Sachfragen aller Art» (Weißeno et al., 2010, S. 32). Dies zeigt sich anschaulich an den vielen Politikfeldern, auf die Politik anwendbar ist bzw. sein muss, wie z.B. Bildungspolitik, Sozialpolitik, Umweltpolitik, Wirtschaftspolitik, Finanzpolitik, Aussenpolitik.

Politik hat drei Dimensionen (Weißeno et al., 2010, S. 32ff.): *polity* (Ordnungsrahmen Institutionen), *policy* (Ziele, Inhalte, Aufgabenfelder) und *politics* (Prozesse). *Polity* gibt den Ordnungsrahmen für politische Prozesse vor. Der Ordnungsrahmen wird auf nationaler Ebene durch die Rechtsordnung (Verfassung, Gesetze etc.) festgelegt, auf internationaler Ebene durch Staatsverträge. *Policy* ist die inhaltliche Gestaltung gesellschaftlicher Verhältnisse bzw. die Lösung gesellschaftlicher Probleme. Die Gestaltungsaufgabe verdeutlicht sich an den verschiedenen Politikfeldern. *Politics* sind – zumindest in demokratischen Staaten – die politischen Prozesse der Willensbildung und Entscheidungsfindung. Strittige Fragen werden öffentlich ausgetragen (z.B. durch Vernehmlassungsverfahren, parlamentarische Debatten, mediale Diskussionen) und mithilfe von Koalitionen und Kompromissen zwischen den machthebenden Gruppierungen gelöst.

In den *policies* findet sich denn auch der oben erwähnten Überschneidungsbereich zwischen Ökonomie und Politik, wie z.B. Bildungspolitik, Sozialpolitik, Beschäftigungspolitik, Finanzpolitik, Aussenpolitik (Europäische Integration) (Kruber, 2005, S. 78f.) (vgl. Tabelle 4-2). Daneben haben beide Disziplinen ihre spezifischen Bereiche: In der Ökonomie sind dies z.B. Konsum, Arbeit, Kapital (Geld, Kredit), betriebliche Leistungsprozesse und Querprozesse; in der Politik z.B. Demokratie, Grundrechte, Wahlen, Freiheit, Gerechtigkeit, Sicherheit.

Tabelle 4-2. Aufgabenfelder von Ökonomie und Politik

Inhaltsbereich der Politikwissenschaft	Überschneidungsbereich von Ökonomie und Politik	Inhaltsbereich der Wirtschaftswissenschaft
<i>Fachkonzepte Ordnung:</i> Demokratie, Gewaltenteilung, Grundrechte, Rechtsstaat, ...	Bildungspolitik Sozialpolitik Beschäftigungspolitik (Arbeitslosigkeit)	<i>Themen Volkswirtschaft (Denk- und Analyseschemata)</i> Verhaltenstheorie (Rationalität) Institutionentheorie/Ordnungstheorie: Wirtschaftsordnung, Institutionen
<i>Fachkonzepte Entscheidung:</i> Interessengruppen, Parteien, Konflikt, Macht, Regierung, Parlament, Wahlen, ...	Finanzpolitik Aussenpolitik (internationale Beziehungen, europäische Integration) ...	Systemtheorie: Wirtschaftskreislauf, Wirtschaftseinheiten (Haushalte, Unternehmen, Staat, Ausland), Märkte (Gütermarkt Arbeitsmarkt, Kapitalmarkt, Bodenmarkt)
<i>Fachkonzepte Gemeinwohl:</i> Freiheit, Frieden, Gerechtigkeit, Gleichheit, Nachhaltigkeit, Sicherheit, ...		<i>Themen Betriebswirtschaft (Struktur und Prozesse):</i> Betriebliche Leistungsprozesse: Beschaffung/Einkauf, Produktion, Vertrieb/Verkauf (Marketing), Entsorgung Betriebliche Querprozesse: Finanzen, Personal, Organisation

Quelle: Eigene Darstellung i.A.a. Kruber (2005, S. 79, S. 85ff.), Kaminski et al. (2017, S. 92, S. 119) und Weißeno et al. (2010, S. 61ff., S. 108ff., S. 161ff.).

Der Politikzyklus ist ein Modell, welches das politische Geschehen prozesshaft abbildet. Ihm liegt ein empirisch-analytischer Politikbegriff zugrunde, bei dem Politik als System der Willensbildung und Entscheidungsfindung für *komplexe und kontroverse gesellschaftliche Probleme* verstanden wird (Gollob et al., 2007, S. 11; Massing, 1995, S. 61ff.; 1999). Mithilfe des Politikzyklus lassen sich diese politischen Prozesse strukturieren und analysieren. Der Politikzyklus gliedert sich in vier Phasen, die einander gegenseitig bedingen (vgl. Massing, 1995, S. 61ff.; Massing, 1999; ZfD, 2015b): (1) Problemstellung (*policy*, Inhalt), (2) Auseinandersetzung über mögliche Lösungen (*politics*, Prozesse), (3) Entscheidung für eine Lösung (*polity*, Institutionen), (4) Beurteilung der Lösung und Reaktion, die u.U. zu einem neuen Problem führt (*policy*, Inhalt). In den Phasen werden alle drei Politik-Dimensionen sichtbar, aber je nach Phase unterschiedlich stark.

Der Politikzyklus zeigt modellhaft, wie sozioökonomische Problemsituationen im gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereich auf der *politischen Ebene* gelöst werden und inwiefern sie den Einzelnen betreffen. Auf die Bewältigung sozioökonomischer Problemsituationen auf *individueller Ebene* wird weiter unten eingegangen (vgl. Kapitel 4.4.2). Nachfolgend werden die Phasen des Politikzyklus exemplarisch für die sozioökonomische Problemsituation „Altersvorsorge“ illustriert (vgl. Tabelle 4-3).

Tabelle 4-3. Phasen des Politikzyklus am Beispiel der sozioökonomischen Problemsituation „Altersvorsorge“

Phase	Beschreibung und Beispiele
1) Problemstellung	- Es entsteht/besteht ein gesellschaftliches, ökologisches oder wirtschaftliches Problem. Politische Gruppierungen stellen ihre Forderungen zum Problem und verlangen nach einer Lösung des Problems. - Leitfragen: Worum geht es beim Problem? Wie ist das Problem entstanden (Ursache)? Wer ist vom Problem betroffen (Akteure)? <i>Bsp.: Die Finanzierung der AHV ist langfristig gefährdet. Wenn keine Finanzierungslösung gefunden wird, erhalten die heutigen Erwerbstätigen keine Rente mehr.</i>
2) Auseinandersetzung	- Politischen Gruppierungen und Einzelpersonen haben unterschiedliche Meinungen zum Problem. Es kommt zu einer öffentlichen Debatte bezüglich des Problems. - Leitfragen: Welche Meinungen werden vertreten („claims“)? Mit welchen Argumenten werden die Meinungen begründet („reasonings“)? Welche Lösungen werden vorgeschlagen? Was beinhalten die Lösungsvorschläge (Massnahmen)? Wie wirken sich die Lösungsvorschläge aus (Folgen, Interdependenzen)? <i>Bsp.: Wie wird die AHV finanziert? Wie ist der aktuelle Stand und wie sind die Prognosen für die Finanzierung der AHV? Wie entwickelt sich die Zahl der Arbeitnehmer*innen und Rentner*innen in den nächsten 20 Jahren? Wer verfolgt gleichgerichtete und entgegengesetzte Interessen bezüglich der Finanzierung der AHV? Welche anderen Interessen spielen bei der Finanzierung der AHV eine Rolle? Welche Konsequenzen hat eine Unterfinanzierung der AHV für welche Akteure? etc.</i>
3) Entscheid (und Umsetzung)	- Es wird über einen Lösungsvorschlag für das Problem entschieden. Je nach Entscheid werden Massnahmen ergriffen und umgesetzt. Der Entscheid hat Konsequenzen für den Einzelnen und die Gesellschaft. - Leitfragen: Wie wird über die Lösung entschieden (Abstimmung, Beschluss)? Wie wird die Lösung umgesetzt? Welche Schwierigkeiten ergeben sich bei der Umsetzung? <i>Bsp.: Erhöhung der Lohnbeiträge der Arbeitnehmer*innen und Arbeitgeber*innen, Senkung der Leistungen für die Renter*innen, Erhöhung der Bundesbeiträge (z.B. Mehrwertsteuer) zu Gunsten der AHV.</i>
4) Beurteilung und Reaktion	- Der Entscheid und seine Umsetzung (Massnahmen zur Lösung des Problems) werden von politischen Gruppierungen und Einzelpersonen beurteilt. - Leitfragen: Wie wird die Umsetzung beurteilt? Von wem wird die Umsetzung positiv oder negativ beurteilt? Inwiefern zeigt die Umsetzung die gewünschte Wirkung? <i>Bsp.: der Lösungsansatz „Lohnbeiträge erhöhen“ wird vom Stimmbürger*innen an der Urne angenommen, vom Parlament gesetzlich verankert, bei den AHV-Kassen und Unternehmen administrativ eingeführt und schliesslich durch die Behörden auf seine beabsichtigte Wirkung hin geprüft.</i>
1) Neue Problemstellung	- Aus dem Entscheid und seiner Umsetzung entstehen neue Probleme. - Leitfragen: Welche neuen Probleme entstehen?

Quelle: Eigene Darstellung i.A.a. Massing (1999), Gollob et al. (2007) und ZfD (2015a).

4.3.2.b *Gesellschaftliche Verantwortung und politische Teilhabe*

In einem modernen und liberalen Rechtsstaat wie der Schweiz übernimmt jeder einzelne Mensch die Verantwortung für sich selbst und trägt zur Bewältigung der Aufgaben in Staat und Gesellschaft bei (Art. 6 Bundesverfassung). Diese gesellschaftliche Verantwortung soll den jungen Menschen v.a. in den Bildungsinstitutionen bewusst gemacht werden, mehr noch, Bildungsinstitutionen sollen jungen Menschen auf ihre vielfältigen Aufgaben in der Gesellschaft vorbereiten. Sämtliche Bildungspläne auf der Sekundarstufe II in der Schweiz verweisen explizit auf diese gesellschaftliche Verantwortung:

- Gymnasiale Maturität «[...] auf anspruchsvolle Aufgaben in der Gesellschaft vorbereitet» (Art. 5 Abs. 1 MAR).
- Berufsmaturität: «[...] sich darin [im Fachhochschulstudium] auf eine anspruchsvolle Aufgabe in Wirtschaft und Gesellschaft vorzubereiten» (Art. 3 Abs. 1 lit. a Berufsmaturitätsverordnung).
- Kaufmännische Berufsbildung: «[Ich] übernehme Verantwortung für meine Arbeiten und mein Verhalten» (SKKAB, 2011a, S. 12) und «[Kaufleute] sind sich ihrer Verantwortung und Möglichkeiten als Wirtschafts- und Gesellschaftsbürger bewusst» (SKKAB, 2011b, S. 1).
- Berufsbildung: «[...] übernehmen Verantwortung für sich, für Mitmenschen und die Erhaltung der natürlichen Umwelt» (SBFI, 2006, S. 11).

Zur gesellschaftlichen Verantwortung gehört mitunter die Teilhabe am politischen und wirtschaftlichen Leben (Eberle, 2015, S. 26f.). Die politische Teilhabe hat unterschiedliche Facetten: Die Menschen engagieren sich freiwillig in staatlichen und nicht-staatlichen Organisationen (z.B. Vereine, Stiftungen), die einen politischen, ökologischen, kulturellen, sozialen, karitativen oder sportlichen Zweck verfolgen. Die Menschen werden durch Behörden und Medien mit gesellschaftlichen Frage-/Problemstellungen konfrontiert und im Rahmen einer öffentlichen Debatte in den (politischen) Meinungsbildungsprozess involviert. Die Menschen bzw. die Staatsbürger*innen werden mittels Sachabstimmungen aufgefordert, sich zu aktuellen wirtschafts- und gesellschaftspolitischen Problemstellungen bzw. zu deren Lösungsvorschlägen zu äussern und sind somit in den politischen Meinungsbildungs- und Entscheidungsprozess eingebunden.³⁴

³⁴ Beispielsweise stimmten die Schweizer Stimmbürger*innen in jüngster Zeit über folgende Vorlagen ab: Abschaffung der Pauschalbesteuerung, Einführung von gesetzlichen Mindestlöhnen, Beschränkung der Immigration, Liberalisierung der Ladenöffnungszeiten, die Begrenzung von Managergehältern, Ausstieg aus der Kernenergie, Vollgeld, bedingungsloses Grundeinkommen, Spekulation mit Nahrungsmitteln, Unternehmensbesteuerung.

Tabelle 4-4. Gesellschaftliche Verantwortung und politische Rechte

Art. 6 BV, Individuelle und gesellschaftliche Verantwortung

Jede Person nimmt Verantwortung für sich selber wahr und trägt nach ihren Kräften zur Bewältigung der Aufgaben in Staat und Gesellschaft bei.

Art. 34 BV: Politische Rechte

¹ Die politischen Rechte sind gewährleistet.

² Die Garantie der politischen Rechte schützt die [freie Willensbildung](#) und die [unverfälschte Stimmabgabe](#).

Art. 136 BV: Politische Rechte

¹ Die politischen Rechte in Bundessachen stehen allen [Schweizerinnen und Schweizern](#) zu, die [das 18. Altersjahr zurückgelegt](#) haben und die nicht wegen Geisteskrankheit oder Geistesschwäche entmündigt sind. Alle haben die gleichen politischen Rechte und Pflichten.

² Sie können an den Nationalratswahlen und [an den Abstimmungen des Bundes teilnehmen](#) sowie Volksinitiativen und Referenden in Bundesangelegenheiten ergreifen und unterzeichnen.

Quelle: Bundesverfassung der Schweizerischen Eidgenossenschaft vom 18. April 1999.

In einem demokratischen Staat wie der Schweiz haben die mündigen Staatsbürger*innen verfassungsmässig garantierte politische Rechte, wozu auch das Stimmrecht für Sachvorlagen zählt (Art. 34 Bundesverfassung i.V.m. Art. 136 Bundesverfassung) (vgl. Tabelle 4-4). Sachabstimmungen finden drei- bis viermal pro Jahr statt, am häufigsten auf Bundesebene, seltener auch auf kantonaler und kommunaler Ebene. Sie werden vielfach vom Parlament (parlamentarische Initiative) und von den Stimmbürgern (Volksinitiative), vereinzelt auch von den Kantonen (Standesinitiative) initiiert. In den Monaten und Wochen vor einer Abstimmung werden die Schweizer Staatsbürger*innen von Behörden, politischen Parteien, weiteren Interessengruppen und Medien über die Abstimmungsvorlage informiert und u.U. auch beeinflusst. Dies geschieht beispielsweise mittels der offiziellen Abstimmungsbroschüre des Bundes, Positionspapieren und Abstimmungsparolen von Parteien und Interessengruppen, Diskussionspodien, Zeitungsartikeln, Radio- und Fernsehsendungen etc.

Das Abstimmungsergebnis hat zum einen direkte Konsequenzen für die schweizerische Rechtsordnung, da der formelle Gegenstand einer Abstimmungsvorlage ein Verfassungsartikel oder ein Gesetz ist, der/das eingeführt, geändert oder abgeschafft werden soll. Zum anderen hat das Abstimmungsergebnis, auch wenn der materielle Gegenstand der Abstimmungsvorlage kein wirtschaftlicher, sondern ein gesellschaftlicher oder ökologischer ist, direkte oder indirekte Konsequenzen auf die Schweizer Wirtschaft. So wird beispielsweise mit der Forderung nach einer Immigrationsbeschränkung ein gesellschaftliches Anliegen aufgegriffen (Angst vor Überfremdung, Verlust der „nationalen Identität“), das aber kurz- und langfristige ökonomische (Mangel an qualifizierten Arbeitskräften) und politische Auswirkungen hat (Gefährdung der bilateralen Verträge mit der EU, insbesondere Personenfreizügigkeit). Schliesslich hat das Abstimmungsergebnis auch direkte Konsequenzen für das Individuum, das im Alltag vielfältige und wechselnde Rollen als

Familienmitglied, Arbeitnehmer, Konsument, Sparer/Anleger, Steuerzahler, Staatsbürger etc. innehat.

Im Rahmen des politischen Meinungsbildungs- und Entscheidungsprozesses besteht für den Einzelnen die anspruchsvolle gesellschaftliche Aufgabe darin, Informationen zu sammeln, zu sortieren und zu bewerten (d.h. relevante von irrelevanter Information zu unterscheiden sowie Tatsachen von Meinungen zu unterscheiden), um sich eine Meinung zu bilden und schliesslich eine informierte und begründete Entscheidung zu treffen. Des Weiteren erfordert der politische Meinungsbildungs- und Entscheidungsprozess grundlegendes, wenn nicht gar vertieftes Wissen über das gesellschaftliche System und die Interdependenzen seiner Subsysteme (Gesellschaft i.e.S., Politik, Wirtschaft) und anderer Systeme (z.B. Ökologie) sowie eine reflektierte Haltung gegenüber den Strukturen und Prozessen in diesem System (Dubs, 2011, S. 196; Eberle, 2006b, S. 19). Denn mit jeder Abstimmungsvorlage gehen individuelle und kollektive Interessenskonflikte, regionale und globale Interessenskonflikte sowie kurzfristige und langfristige Zielkonflikte einher. Die vorgelegten Abstimmungsalternativen (Annehmen der Volksinitiative bedeutet Aufgeben des *Status quo*, Ablehnen bedeutet Beibehalten des *Status quo*) sind komplex und kontrovers. Nichtsdestotrotz sind sie für das gegenwärtige und zukünftige (friedliche) Zusammenleben einer Gesellschaft richtungsweisend, insofern stellen sie grosse Herausforderungen dar.

4.3.2.c Politikbereiche und sozioökonomische Problemsituationen: Eine Domänenanalyse für die Schweiz, 1990 bis 2007

Um konkrete sozioökonomische Problemsituationen im gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereichs nicht intuitiv, sondern analytisch zu identifizieren, wird eine empirische Domänenanalyse vorgenommen. Die Domänenanalyse bezieht sich räumlich auf die Schweiz, zeitlich auf die Jahre 1990 bis 2007 und sachlich auf eidgenössische Abstimmungsvorlagen. Dabei wird unterstellt, dass eidgenössische Abstimmungsvorlagen komplexe und kontroverse gesellschaftliche Anliegen reflektieren, die in der breiten Öffentlichkeit auch als solche wahrgenommen werden durch die Berichterstattung in den Medien und die Debatten in den Politikarenen. Diese Domänenanalyse des gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereichs bildet die Grundlage für die Inhaltsspezifikationen des WBK-T2 (vgl. Kapitel 8.2).

Die vorliegende Domänenanalyse unterscheidet sich dahingehend von der bisherigen Domänenanalyse im Projekt CoBALIT (Eberle et al., 2016, S. 101f.), als dass sie im zeitlichen und sachlichen Anwendungsbereich umfassender sowie in der Methodik systematischer ist. Die Informationen zu den Abstimmungsvorlagen wurden der Datenbank des Bundesamts für Statistik entnommen (BFS, 2017). Im untersuchten Zeitraum gab es 260 eidgenössische Abstimmungen, davon 109 Volksinitiativen (42 %), 74 Bundesbeschlüsse (29 %) und 70 fakultative Referenden

(27 %).³⁵ 47 % der 260 Abstimmungsvorlagen wurden von den Stimmbürger*innen an der Urne angenommen.

Die Abstimmungsvorlagen wurden nach der Methode der strukturierenden qualitativen Inhaltsanalyse (vgl. Döring & Bortz, 2016, S. 553ff.; Mayring, 2015) ausgewertet. Die Kodierungen wurden von der Autorin und einem geschulten Kopierer getrennt vorgenommen und alle abweichenden Kodierungen in eine Konsenskodierung überführt (Döring & Bortz, 2016, S. 558). Zuerst wurde jede Abstimmungsvorlage mit eins bis drei Schlüsselwörtern aus ihrem Originaltitel kodiert. Danach wurde jede Vorlage mithilfe der Schlüsselwörter einem bekannten Politikfeld zugeordnet (vgl. Tabelle 4-5). Die Benennung und Granulierung der Politikfelder wurde deduktiv aus den bestehenden Kategorien hergeleitet (CAP, 2015; Gilg, Raymond, & Bindschedler-Robert, 1966). Wo keine oder keine passenden Politikfelder bekannt waren, wurden weitere Politikfelder induktiv anhand der Schlüsselwörter weiterentwickelt. Bei einer mehrfach möglichen Zuordnung wurde jeweils das in der Abstimmungsvorlage inhaltlich dominantere Politikfeld gewählt.

Insgesamt konnten 17 Politikfelder identifiziert werden (vgl. Abbildung 4-2). Die meisten Abstimmungsvorlagen häufen sich in den Politikfeldern soziale Sicherheit & Vorsorge (12.3 %), Verkehr (8.5 %), Gesundheit (7.7 %), Energie & Umwelt (7.3 %), Migration & Integration (7.3 %), Finanzen & Steuern (6.9 %) sowie Wettbewerb (5.4 %).

Für jedes Politikfeld wurden mögliche sozioökonomische Problemsituationen extrahiert (vgl. Tabelle 4-6 und Anhang Tabelle A- 1). So finden sich beispielsweise im Politikfeld „soziale Sicherheit & Vorsorge“ die Problemsituationen *Altersvorsorge* und *Arbeitslosigkeit*, im Politikfeld „Verkehr“ die Problemsituation Strassenverkehr (*Road Pricing*), im Politikfeld „Gesundheit“ die Problemsituation Krankenversicherung, im Politikfeld „Energie & Umwelt“ die Problemsituationen *Energieversorgung* und *Umwelt- und Klimaschutz* sowie im Politikfeld „Finanzen & Steuern“ die Problemsituationen *Staatsverschuldung* und *Steuerwettbewerb/Finanzausgleich*. Hingegen liessen sich für die Politikfelder „Bildung & Kultur“, „Gesellschaft & Ethik“, „öffentliche Sicherheit & Ordnung“, „Verfassung & Grundrechte“ sowie „Verwaltung“ keine *sozioökonomischen* Problemsituationen extrahieren, wohl aber andere Problemsituationen.

³⁵ Gegenstand von eidgenössische Abstimmungen sind Volksinitiativen (Revision der Bundesverfassung), obligatorische Referenden (Revision der Bundesverfassung, Beitritt zu internationalen Organisationen, dringliche Bundesgesetze ohne Verfassungsgrundlage) und fakultative Referenden (Revision von Bundesgesetzen, Abschluss von völkerrechtlichen Verträgen, Bundesbeschlüsse der Bundesversammlung) (Art. 138, Art. 139 und Art. 139b Bundesverfassung) (Tschannen, 2016, S. 648ff.). Eine Volksinitiative kommt durch ein Begehren von mindestens 100'000 Stimmbürger*innen innerhalb von 18 Monaten zustande (Art. 138 und 139 Bundesverfassung) (Tschannen, 2016, S. 603, S. 651ff.). Ein fakultatives Referendum kommt durch ein Begehren von 50'000 Stimmbürger*innen oder 8 Kantone zustande (Art. 141, Art. 141a und Art. 142 Bundesverfassung) (Tschannen, 2016, S. 584ff., S. 654f.). Ein Bundesbeschluss ist ein Rechtserlass der Bundesversammlung (Art. 163 Bundesversammlung) (Tschannen, 2016, S. 592ff.).

Tabelle 4-5. Exemplarische Kodierung von Abstimmungsvorlagen zu den Politikfeldern „Energie & Umwelt“ sowie „Soziale Sicherheit & Vorsorge“

Nr.	Datum	Abstimmungsvorlage (Originaltitel)	Schlüsselwörter	Politikfeld
4660	24.09.2000	Verfassungsartikel über eine Energielenkungsabgabe für die Umwelt (Gegenentwurf zur zurückgezogenen "Energie-Umwelt-Initiative")	Energielenkungsabgabe	Energie & Umwelt
5080	16.05.2004	Bundesbeschluss vom 03.10.2003 über die Finanzierung der AHV/IV durch Anhebung der Mehrwertsteuersätze	AHV/IV, Mehrwertsteuer-Erhöhung	Soziale Sicherheit & Vorsorge
5010	18.05.2003	Volksinitiative "Strom ohne Atom - Für eine Energiewende und schrittweise Stilllegung der Atomkraftwerke (Strom ohne Atom)"	Atomenergie-Ausstieg, Energiewende	Energie & Umwelt
5360	30.11.2008	Volksinitiative vom 28.03.2006 "Für ein flexibles AHV-Alter"	AHV, Rentenalter	Soziale Sicherheit & Vorsorge
6120	21.05.2017	Energiegesetz (EnG) vom 30.09.2016	Energiegesetz	Energie & Umwelt
6150	24.09.2017	Bundesgesetz vom 17.03.2017 über die Reform der Altersvorsorge 2020	Altersvorsorge, AHV	Soziale Sicherheit & Vorsorge

Quelle: BFS (2017); eigene Kategorisierung und Darstellung.

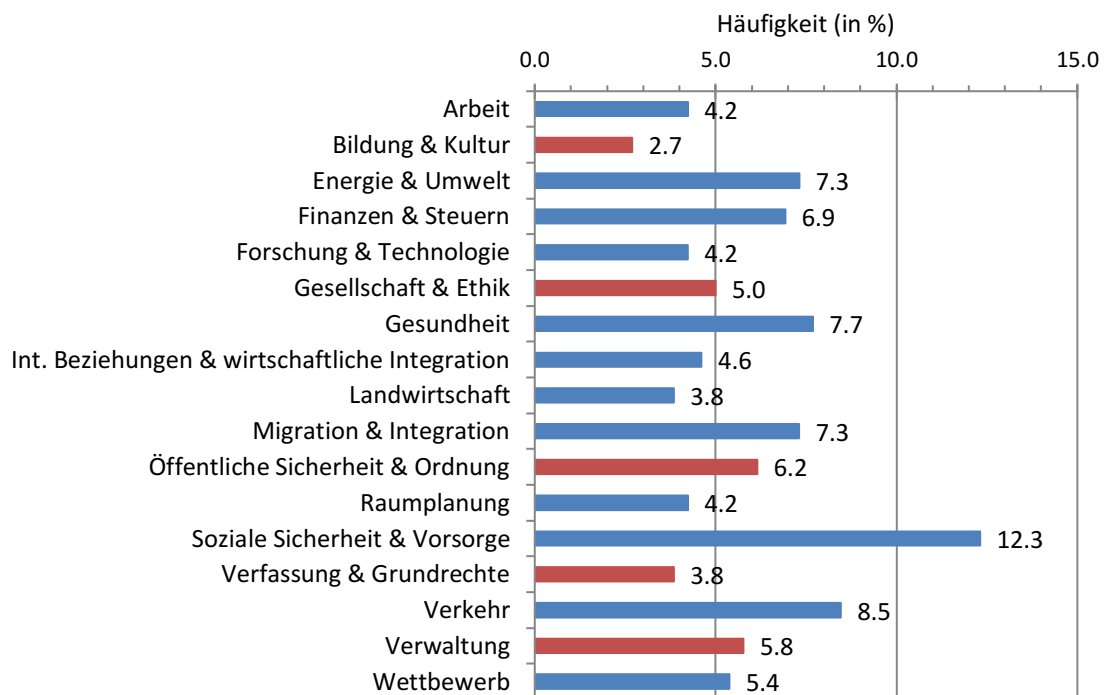


Abbildung 4-2. Eidgenössische Abstimmungsvorlagen nach Politikfeldern 1990-2017

Hinweis: $m = 260$ Abstimmungsvorlagen. Politikfelder in alphabetischer Reihenfolge. Blau = enthält sozioökonomische Problemsituationen, rot = enthält keine sozioökonomischen Problemsituationen.

Quelle: BFS (2017); eigene Kategorisierung und Darstellung.

Tabelle 4-6. Politikfelder und sozioökonomische Problemsituationen

Politikfeld	Sozioökonomische Problemsituation	Politikfeld	Sozioökonomische Problemsituation
Arbeit	▪ Gehaltsgerechtigkeit (Managervergütungen)	Landwirtschaft	▪ Agrarhandel ▪ Agrarproduktion
Energie & Umwelt	▪ Energieversorgung ▪ Umwelt- und Klimaschutz	Migration & Integration	▪ Flüchtlingsströme
Finanzen & Steuern	▪ Staatsverschuldung ▪ Steuerwettbewerb/ Finanzausgleich	Raumplanung	▪ Wohnungsmarkt und Mietpreise
Forschung & Technologie	▪ (Biotechnologie)	Soziale Sicherheit & Vorsorge	▪ Altersvorsorge ▪ Arbeitslosigkeit
Gesundheit	▪ Krankenversicherung	Verkehr	▪ Strassenverkehr ▪ öffentlicher Verkehr
Internationale Beziehungen & wirtschaftliche Integration	▪ Bilaterale Abkommen Schweiz/EU (Rahmenabkommen)	Wettbewerb	▪ Service Public ▪ Strukturpolitik

Hinweis: Politikfelder in alphabetischer Reihenfolge.

Quelle: BFS (2017); eigene Kategorisierung und Darstellung.

4.4 Prozessebene des WBK-Modells: Kognitionsprozesse

In diesem Kapitel werden die Kognitionsprozesse des WBK-Modells beschrieben und begründet, die für die Bewältigung von ökonomisch geprägter Anforderungssituationen notwendig sind, nämlich die individuelle Informationsverarbeitung (Kapitel 4.4.1) und das individuelle Problemlösen (Kapitel 4.4.2). Es wird auf die Kognitionsprozesse für sozioökonomische Problemsituationen fokussiert, da diese für das zu operationalisierende und zu validierende Konstrukt massgeblich sind.

4.4.1 Informationsverarbeitung bei sozioökonomische Problemsituationen

Kognitionen (lat. *cognitio*: Erkennen, auf Erkenntnis bezogen) sind all jene Prozesse, durch die der Mensch Kenntnis von seiner Umwelt erlangt, also Wissen erwirbt. Zu diesen Vorgängen zählen z.B. Wahrnehmung, Vorstellung, Denken, Urteilen. Von den kognitiven Prozessen sind die emotionalen und motivationalen Prozesse zu unterscheiden. Jedoch ist diese Unterscheidung eine analytische; sie lässt sich in der Realität nicht trennscharf vollziehen, denn kognitive Prozesse sind eng mit emotionalen und motivationalen Prozessen verbunden (Edelmann & Wittmann, 2012, S. 109).

Durch kognitive Prozesse baut der Mensch kognitive Strukturen (Wissensstrukturen) auf. Es sind also aktive, subjektive Strukturierungsprozesse. Häufig findet kein Neulernen, sondern ein

Umlernen bereits vorhandener Strukturen statt (Edelmann & Wittmann, 2012, S. 109f.). Nach dem *Modell der menschlichen Informationsverarbeitung* können kognitive Prozesse analytisch in einzelne Schritte zerlegt werden (Edelmann & Wittmann, 2012, S. 141f.). Jedoch hängen diese Schritte eng miteinander zusammen und beeinflussen sich gegenseitig: So beeinflusst das bereits gespeicherte Wissen die Aneignung neuer Information, die Enkodierung der Information beeinflusst die Dekodierung der Information.

- Aneignung (Enkodierung, Lernen i.e.S.): Aussenreiz (Input), aktive und selektive Aufnahme von Information, Verarbeitung von Information;
- Speicherung (Gedächtnis i.e.S.): mentale Repräsentation von deklarativem und prozeduralem Wissen;
- Abruf (Dekodierung): Erinnerung und aktive Rekonstruktion, Leistung (Output).

Repräsentationen sind innere, mentale Darstellungen von *Wissen* (Edelmann & Wittmann, 2012, S. 129ff.). Sie können abstrakt/aussagenartig (sprachlich-symbolisch, z.B. einen Berg als das Gegenteil eines Tals beschreiben), analog/bildhaft (konkret-anschaulich, Analogie zwischen äusserer Erscheinung und innerer Repräsentation, z.B. ein Bild vom einem Berg betrachten) und/oder handlungsmässig (unmittelbar, erfahrungsbasiert, handelnder Umgang mit Dingen, z.B. einen Berg besteigen) sein.

Wissen lässt sich in deklaratives und prozedurales unterscheiden (Edelmann & Wittmann, 2012, S. 111). *Deklaratives Wissen* (auch: Sachwissen) bezieht sich auf das *Was* etwas ist, z.B. was ein Begriff bedeutet oder wie ein Begriff verwendet wird. Es ist explizit, d.h. es kann sprachlich formuliert werden.³⁶ *Prozedurales Wissen* (auch: Handlungswissen) bezieht sich auf das *Wie* etwas zu tun ist, z.B. wie eine Aufgabe bewältigt oder wie ein Problem gelöst wird. Es ist häufig implizit, d.h. es sind automatisierte und unbewusste Routinen, die als allgemeine oder spezifische Fertigkeiten vorliegen können.³⁷ Mit zunehmendem Expertisegrad nimmt das Bewusstsein über das prozeduralen Wissens ab: Ein Novize muss sich für die Ausübung einer bestimmten Tätigkeit das prozedurale Wissen mehr bewusst machen, bei einem Experten ist das prozedurale Wissen unbewusst abrufbar. Doch auch diese Unterscheidung ist eine analytische, denn in der Realität sind die Grenzen zwischen deklarativem und prozeduralem Wissen fließend.

In der allgemein-didaktischen Lehr-Lernforschung werden für die Strukturierung von kognitiven Anforderungsniveaus von Lernzielen, Lernaufgaben und Prüfungsaufgaben seit den 1950er-Jahren Taxonomien entwickelt und weiterentwickelt. Die ursprüngliche Taxonomie

³⁶ Beispiele aus dem gesamtwirtschaftlichen Lebensbereich: das Phänomen Arbeitslosigkeit erklären, Massnahmen gegen Arbeitslosigkeit beschreiben.

³⁷ Beispiele aus dem beruflichen Lebensbereich: Stellenportale durchsuchen, eine Betriebsanlage bedienen.

(*Taxonomie of Educational Objectives: Cognitive Domain*) stammt von Bloom et al. (1956) und bezieht sich auf Lernziele in der kognitiven Domäne. Sie besteht aus sechs Kategorien (Bloom et al., 1956): (1) *knowledge*, (2) *comprehension*, (3) *application*, (4) *analysis*, (5) *synthesis*, (6) *evaluation*. Diese Taxonomie suggeriert eine kumulative Hierarchie der Kategorien: Das Bewältigen von Lernaufgaben mit höheren kognitiven Prozessen setzt das Bewältigen von jenen mit niedrigeren kognitiven Prozessen voraus. Oder anders ausgedrückt: Lernaufgaben der höheren Kategorie sind schwieriger als jene einer geringeren Kategorie. Gegeben dieser Annahme könnten kognitive Prozesse bei der Konstruktion von Testitems als Kriterien für Schwierigkeit und Komplexität berücksichtigt (vgl. Kapitel 6.2.2) und bei der Etablierung eines Kompetenzniveaumodells herangezogen werden (vgl. Kapitel 3.3). Allerdings kann die kumulative Hierarchie der Bloomschen Taxonomie empirisch nicht eindeutig bestätigt werden (vgl. dazu Winther, 2010, S. 105).

Die revidierte Taxonomie (*A Taxonomy for Learning, Teaching, and Assessing*) von Anderson & Krathwohl (2001) bezieht sich auf Lernen, Lehren und Prüfen. Sie verbindet in einer zweidimensionalen Matrix Kognitionsprozesse (*cognitive process dimensions*) und Wissensarten (*knowledge dimensions*). Für kognitive Prozesse werden sechs Kategorien definiert (L. W. Anderson & Krathwohl, 2001): (1) *remember* (erinnern), (2) *understand* (verstehen), (3) *apply* (anwenden), (4) *analyze* (analysieren), (5) *evaluate* (beurteilen), (6) *create* (entwerfen, entwickeln). Bei den Wissensarten werden vier Kategorien unterschieden: *factual knowledge* (Faktenwissen), *conceptual knowledge* (Konzeptwissen), *procedural knowledge* (Handlungswissen), *metacognitive knowledge* (Metakognition). Bei der revidierten Taxonomie der kognitiven Prozesse wird eingeräumt, dass ihre kumulative Hierarchie nicht streng angelegt ist und Überschneidungen zwischen den Kategorien zulässt. Nichtsdestotrotz beschreiben diese Taxonomien generelle kognitionspsychologische Kategorien der Informationsverarbeitung und eignen sich deshalb nur bedingt, um kognitive Prozesse für domänenspezifische Anforderungssituationen vorherzusagen (vgl. dazu Winther, 2010, S. 105f.).

Eine neue Taxonomie (*The New Taxonomy of Educational Objectives*) wird von Marzano & Kendall (2007) vorgeschlagen (Marzano & Kendall, 2007). Diese Taxonomie spannt zwei Dimensionen auf, die Prozessdimension (*levels-of-processing dimension*) und die Wissensdimension (*domain-of-knowledge dimension*). Die Wissensdimension definiert drei Kategorien: *information* (verstanden als deklaratives Wissen), *mental procedures* (verstanden als prozedurales Wissen) und *psychomotor procedures* (verstanden als physische Aktivitäten). Für die Prozessdimension werden drei mentale Systeme definiert: *self-system* (das Selbst), *metacognitive system* (Metakognition) und *cognitive system* (Kognition). Das kognitive System wiederum bezieht sich auf kognitive Prozesse für Lern- bzw. Arbeitsanforderungen: (1) *retrieval* (*recognizing, recalling, executing*), (2) *comprehension* (*integrating, symbolizing*), (3) *analysis* (*matching/classifying, analyzing, generalizing/specifying*) (4) *knowledge utilization* (*decision-making, problem-solving, experimenting, investigating*).

Tabelle 4-7. Prozessebene des WBK-Modells: Informationsverarbeitung am Beispiel der sozioökonomischen Problemsituation „Altersvorsorge“

Informations- verarbeitungsstufen	Beschreibung und Beispiele
(k1) wiedererkennen, wiedergeben	Informationen unverändert reproduzieren (z.B. benennen, zuordnen, vervollständigen) <i>Bsp.: Leistungsempfänger, Beitragszahler, Generationenvertrag.</i>
(k2) verstehen, anwenden	- Informationen transferieren: vom Allgemeinen zum Speziellen, vom Abstrakten zum Konkreten, von Situation A nach Situation B - Informationen unterscheiden: belegbare (Tatsachen) und nicht-belegbare Aussagen (Meinungen, Positionen), deskriptive und normative Aussagen - Informationen sinngemäss reproduzieren und interpretieren: Text, Tabelle, Abbildung (beschreiben, erklären, begründen) <i>Bsp.: Ursachen der Verschuldung der Altersversicherung, Auswirkungen der Erhöhung des Rentenalters auf Arbeitsmarkt und Altersversicherung.</i>
(k3) vergleichen, beurteilen, entscheiden	- Bestehende Informationen verknüpfen und daraus neuartige Informationen generieren - Informationen anhand vorgegebener oder eigener Kriterien evaluieren - Informationen umfassend und systematisch analysieren <i>Bsp.: Lösungsvorschläge zur Finanzierung der AHV.</i>

Quelle: Eigene Entwicklung und Darstellung i.A.a. Marzano & Kendall (2007).

Im WBK-Modell wird für die Informationsverarbeitung eine reduzierte kognitive Taxonomie mit drei Stufen definiert (i.A.a. Marzano & Kendall, 2007). Diese reduzierte Taxonomie ermöglicht m.E. eine trennschärfere Operationalisierung des Konstrukts WBK-soek (vgl. Kapitel 8.3.3) und eine einfachere empirische Überprüfung (vgl. Kapitel 11.1). Nachfolgend werden die Kognitionsstufen allgemein beschrieben und exemplarisch anhand der sozioökonomischen Problemsituation „Altersvorsorge“ illustriert (vgl. Tabelle 4-7).

- Auf der *unteren Kognitionsstufe (k1)* „wiedererkennen/wiedergeben“ werden Informationen unverändert reproduziert. Beispiel: ökonomische/politische Fakten wie z.B. „Leistungsempfänger“ oder „Beitragssätze“ sowie ökonomische/politische Konzepte wie z.B. „3-Säulen-Prinzip“ und „Generationsvertrag“ benennen oder zuordnen.
- Auf der *mittleren Kognitionsstufe (k2)* „verstehen/anwenden“ werden Informationen sinngemäss reproduziert, interpretiert und transferiert. Beispiel: ökonomische/politische Konzepte beschreiben, erläutern oder begründen wie z.B. „Ursachen der Verschuldung der Altersversicherung“ oder „Auswirkungen der Erhöhung des Rentenalters auf Arbeitsmarkt“.
- Auf der *oberen Kognitionsstufe (k3)* „vergleichen/beurteilen/entscheiden“ werden Informationen systematisch verglichen, kriteriengeleitet beurteilt, aufgrund dessen stringent entschieden und differenziert und reflektiert begründet. Beispiel: konkrete Lösungsansätze für Problemstellungen vergleichen und beurteilen, so z.B. „Lösungsansätze für die Finanzierung der Altersversicherung“.

4.4.2 Problemlösen bei sozioökonomischen Problemsituationen

Handeln wird über kognitive Strukturen gesteuert; grösstenteils über Wissensstrukturen, kleinernteils über Problemlösestrukturen (Edelmann & Wittmann, 2012, S. 179f.). Erst wenn die Wissensstrukturen nicht ausreichen, um das intendierte Ziel zu erreichen, werden Problemlösestrukturen angewendet. Wissensstrukturen (auch: epistemische Strukturen) sind die durch Informationsverarbeitung erlernten Begriffe und Regeln zur Bewältigung von Aufgaben (vgl. Kapitel 4.4.1); dem gegenüber sind Problemlösestrukturen (auch: heuristische Strukturen) alle verfügbaren Verfahren zur Lösung von Problemen. Ob eine bestimmte Situation eine Aufgabe oder ein Problem darstellt, hängt vom Vorwissen bzw. von den verfügbaren Wissensstrukturen des Individuums ab. Ein Problem ist also nicht *per se* ein Problem, sondern nur ein Problem für bestimmte Personen bzw. Personengruppen.

Handeln kann spontan, intuitiv oder planvoll erfolgen. *Planvolles Handeln* charakterisiert sich durch die Merkmale des «idealen Handelnden» (Edelmann & Wittmann, 2012, S. 168ff.): Innensteuerung durch das autonome Subjekt,³⁸ Entscheidung über Alternativen, subjektive Sinnhaftigkeit, Intentionalität (Zielgerichtetheit), Bewusstheit, flexibles und antizipatorisches Handlungskonzept, Verantwortlichkeit, Wissenserwerb.

Problemlösen ist also ein Sonderfall des planvollen Handelns, bei dem das Ziel des Handelns wegen eines „Hindernisses“ nicht direkt erreichbar ist (Edelmann & Wittmann, 2012, S. 178f.). Ein „Problem“ besteht also aus einem unerwünschten Anfangszustand, einem erwünschten Zielzustand sowie einer Barriere zwischen Anfangszustand und Zielzustand.

«Ein Individuum steht einem Problem gegenüber, wenn es sich in einem inneren oder äußeren Zustand befindet, den es aus irgendwelchen Gründen nicht für wünschenswert hält, aber im Moment nicht über die Mittel verfügt, um den unerwünschten Zustand in den wünschenswerten Zustand zu überführen» (Dörner, 1979, S. 10).

Probleme haben unterschiedliche Komplexitätsgrade. Sozioökonomische Problemsituationen (wie z.B. Arbeitslosigkeit, Umweltbelastung) zeichnen sich durch eine hohe Komplexität aus. Für ihre Lösung muss die Struktur und Dynamik des gesamten Systems analysiert werden; sie erfordert also Problemlösen durch Systemdenken (vgl. Dörner, Kreuzig, Reither, & Ständel, 1983). Oder anders ausgedrückt, isolierte und kurzfristige Massnahmen lösen das Problem nicht (nachhaltig). Nach der Problemlösetheorie des Systemdenkens charakterisieren sich komplexe Probleme wie folgt (Dörner et al., 1983, zit. nach Edelmann & Wittmann, 2012, S. 189):

³⁸ Das Gegenmodell zur Innensteuerung der Handlung ist die Aussensteuerung des *Verhaltens* beim Reiz-Reaktions-Lernen und beim instrumentellen Lernen (Edelmann & Wittmann, 2012, S. 172).

- *Komplexität*: Es bestehen viele verschiedene Einflussfaktoren; dies erfordert Informationsselektion und Informationsreduktion.
- *Vernetztheit*: Die einzelnen Faktoren beeinflussen sich gegenseitig; dies führt zu Nebenfolgen und Spätfolgen.
- *Intransparenz*: Es sind nicht alle relevanten Informationen für die Problemlösung vorhanden; Zustände von und Beziehungen zwischen Faktoren sind nur teilweise, nur indirekt und/oder zeitverzögert erfassbar.
- *Eigendynamik*: Das Problem oder Elemente davon entwickeln sich *während* des Problemlöseprozesses weiter, auch ohne Zutun der Problemlöser*in; dies erhöht den Druck für „rasche“ Lösungen.
- *Offenheit* des Zielzustandes: Das Ziel ist häufig nur vage formuliert (z.B. „Die Arbeitslosigkeit soll gesenkt werden“) und muss präzisiert werden; die Präzisierung des Ziels ist Teil der Problemlösung.
- *Polytelie*: Es sind mehrere und ggf. auch widersprüchliche Ziele zu berücksichtigen; dies führt zu Zielkonflikten und erfordert ein multiperspektivisches Vorgehen.

Das allgemeine *Modell des Problemlöseprozesses* beschreibt vier Schritte (Edelmann & Wittmann, 2012, S. 190f.). Die ersten zwei Phasen des Problemlöseprozesses beziehen sich auf das Problem, die letzten zwei auf die Lösung.

- *Problemraum*: Die Situation wird aufgrund der individuellen Wissensstrukturen als Problem repräsentiert. Zwischen Anfangszustand und Zielzustand besteht eine Barriere („Es geht nicht“).
- *Situationsanalyse*: Für das Problem wird eine Zielanalyse („Was soll gehen?“) und eine Konfliktanalyse („Warum geht es nicht?“) durchgeführt. Ein schlecht definiertes Problem (*ill-structured problem*) (unscharfer Anfangs- und Zielzustand) wird in ein wohldefiniertes Problem (präziser Anfangs- und Zielzustand) umstrukturiert.
- *Suchraum*: Die Lösung des Problems wird abhängig von den Merkmalen der Problemsituation und von den Handlungsmöglichkeiten der Person gesucht („Wie könnte es gehen?“). Dabei können verschiedene problemlösende Denkverfahren angewendet werden: Versuch und Irrtum (die Lösung wird zufällig gefunden), Umstrukturierung (die Lösung des Problem ergibt sich aus der präzisen Zielformulierung), Kreativität (die Lösung kommt durch Eingebung bzw. Erleuchtung), Systemdenken (die Lösung eines ähnlichen Problems wird transferiert).
- *Lösung und Evaluation*: Die gefundene Lösung („So geht es!“) wird hinsichtlich des Problems bewertet.

Tabelle 4-8. Prozessebene des WBK-Modells: Problemlösen für sozioökonomische Problemsituationen

Problemlösephasen	Beschreibung und Beispiele
(p1) Problem identifizieren und formulieren	Die politischen Akteure oder die Medien greifen eine sozioökonomische Problemsituation auf, z.B. Finanzierung der Altersvorsorge.
(p2) Problem analysieren	Das Individuum setzt sich mit der sozioökonomischen Problemsituation auseinander. ▪ Informationen recherchieren, verifizieren und selektieren ▪ Fakten und Meinungen differenzieren ▪ involvierte Akteure, ihre Interessen und Argumente identifizieren ▪ Zielkonflikte und Interessenskonflikte erkennen ▪ tangierte Systeme und Interdependenzen erkennen
(p3) Alternative Lösungen suchen, vergleichen und beurteilen	Die Behörden, die politischen Akteure und/oder die Medien schlagen diverse Lösungsansätze für die sozioökonomische Problemsituation vor. Das Individuum setzt sich mit den Lösungsvorschlägen auseinander: ▪ anhand Kriterien vergleichen ▪ Vorteile und Nachteile abwägen, Chancen und Gefahren abwägen ▪ eigenen Standpunkt finden und entscheiden (Stimmzettel ausfüllen)
(p4) Eine Lösung implementieren, evaluieren, ggf. modifizieren	Die Individuen entscheiden über die Lösungsansätze durch eine politische Abstimmung. Die Behörden setzen den Lösungsansatz um und überprüfen ihn in regelmässigen Abständen.

Quelle: Eigene Entwicklung und Darstellung i.A.a. Edelmann & Wittmann (2012, S. 190f.), Massing (1999), Gollob et al. (2007).

Im WBK-Modell wird für das Problemlösen auf *individueller Ebene* ein zirkulärer und iterativer Prozess mit vier Phasen definiert (vgl. Edelmann & Wittmann, 2012, S. 190f.). Der Problemlöseprozess ist iterativ, da aus der letzten Phase wiederum ein neues Problem entstehen kann, für das eine Lösung gefunden werden muss. Nachfolgend werden die Prozessphasen für sozioökonomische Problemsituationen beschrieben (vgl. Tabelle 4-8).

- In Phase 1 (p1) wird eine Problemsituation identifiziert und formuliert. Bei sozioökonomischen Problemsituationen wird sie i.d.R. durch die politischen Akteure oder die Medien aufgegriffen, z.B. die Finanzierung der Altersvorsorge.
- In Phase 2 (p2) wird die Problemsituation analysiert. Das Individuum setzt sich mit der sozioökonomischen Problemsituation auseinander: Informationen recherchieren, verifizieren und selektieren; Fakten und Meinungen differenzieren; involvierte Akteure, ihre Interessen und Argumente identifizieren; Zielkonflikte und Interessenskonflikte erkennen; tangierte Systeme und Interdependenzen erkennen.
- In Phase 3 (p3) werden alternative Lösungen für die Problemsituation gesucht, verglichen und beurteilt. Bei sozioökonomischen Problemsituationen werden Lösungsansätze durch die Behörden, die politischen Akteure und/oder die Medien vorgeschlagen. Das Individuum setzt

sich mit den verschiedenen Lösungsansätzen für die sozioökonomische Problemsituation auseinander: anhand Kriterien vergleichen, Vorteile und Nachteile abwägen, Chancen und Gefahren abwägen, eigenen Standpunkt finden und entscheiden (Stimmzettel ausfüllen).

- In Phase 4 (p4) wird ein Lösungsansatz für die Problemsituation implementiert, evaluiert und ggf. modifiziert. Bei sozioökonomischen Problemsituationen erfolgt die Entscheidung für einen Lösungsansatz durch politische Abstimmungen, die Umsetzung und Prüfung des Lösungsansatzes durch die Behörden.

4.5 Zwischenfazit

In diesem Kapitel wurde die bisherige Definition der wirtschaftsbürgerlichen Kompetenz (WBK) kritisch beleuchtet. Es wurde auf die mehrdeutige Bezeichnung «wirtschaftsbürgerlich» und auf die vage und breite Beschreibung der Domäne «wirtschaftliche Problemstellungen» eingegangen. Ausgehend von den reklamierten Punkten hat die Autorin die Definition der wirtschaftsbürgerlichen Kompetenz modifiziert, die Domäne der WBK präzisiert und darauf aufbauend ein neues Rahmenmodell für die wirtschaftsbürgerliche Kompetenz (WBK-Modell) entworfen. Die modifizierte Definition der WBK fokussiert auf *ökonomisch geprägte Anforderungssituationen*, sie orientiert sich am lebenssituationalen Prinzip der allgemeinen ökonomischen Bildung und bezieht kognitive Prozesse der Informationsverarbeitung und des Problemlösens ein.

Im neuen WBK-Modell werden eine Strukturebene mit Lebensbereichen und eine Prozessebene mit Kognitionsprozessen aufgespannt. In den drei Lebensbereichen (s1) persönlich-finanziell, (s2) beruflich-unternehmerisch und (s3) gesamtgesellschaftlich/gesamtwirtschaftlich hat der Mensch unterschiedliche Rollen inne (z.B. Verbraucher*in, Erwerbstätige*r, Staatsbürger*in) und wird mit verschiedenen ökonomisch geprägten Anforderungssituationen konfrontiert. Zu diesen Anforderungssituationen gehören bspw. Geld ausgeben und ansparen, eine Versicherung abschliessen, eine Steuererklärung ausfüllen, sich für eine Berufsausbildung entscheiden, ein eigenes Unternehmen gründen, über wirtschafts-, sozial- und umweltpolitische Fragen entscheiden.

Die informationsverarbeitenden Kognitionsprozesse werden in drei Stufen zerlegt: (k1) Informationserinnerung (wiedererkennen/wiedergeben), (k2) Informationsverarbeitung (verstehen/anwenden) und (k3) Informationsnutzung (vergleichen/evaluieren/entscheiden). Die problemlösenden Kognitionsprozesse laufen in vier Phasen ab: (p1) Problem identifizieren und formulieren, (p2) Problemsituation multiperspektiv analysieren, (p3) alternative Lösungsansätze suchen, vergleichen und beurteilen, (p4) eine Lösung implementieren, evaluieren und ggf. modifizieren.

Der gesamtgesellschaftliche/gesamtwirtschaftliche Lebensbereich ist fachwissenschaftlich und fachdidaktisch als Überschneidungsbereich von Ökonomie und Politik zu verstehen, in dem sich viele Politikfelder verorten lassen, z.B. Sozialpolitik, Umweltpolitik, Finanzpolitik. Der *politische*

Problemlöseprozess für gesamtgesellschaftliche/gesamtwirtschaftliche Problemsituationen lässt sich mit dem Modell des Politikzyklus abbilden. In der Schweiz ist die gesellschaftliche Verantwortung des Einzelnen sowohl in der Bundesverfassung wie auch in den Lehrplänen auf der Sekundarstufe II verankert und hat eine erhebliche praktische Bedeutung, was sich z.B. durch politische Teilhabe äussert. Für den Schweizer Kontext wurden mittels einer umfassenden empirischen Domänenanalyse reale Politikfelder identifiziert und konkrete sozioökonomische Problemsituationen extrahiert, so z.B. Altersvorsorge, Agrarhandel, Energieversorgung, Gehaltsgerechtigkeit, Krankenversicherung, Service Public, Staatsverschuldung, Strassenverkehr. Damit liegt nun eine theoretisch und empirisch begründete Beschreibung des gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereichs und der sozioökonomischen Problemsituationen vor, die bei der Konstruktion bzw. Revision des wirtschaftsbürgerlichen Kompetenztests (WBK-Test) verwendet werden kann, insbesondere für die Inhaltsspezifikationen des WBK-T2.

Diese Dissertation fokussiert auf sozioökonomische Problemsituationen im gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereich und definiert sie als die *sozioökonomische Facette der wirtschaftsbürgerlichen Kompetenz* (WBK-soek). Die WBK-soek bildet das zu beschreibende, zu operationalisierende und zu validierende Konstrukt. Die Operationalisierung des Konstrukts WBK-soek wird mit dem revidierten Test zur wirtschaftsbürgerlichen Kompetenz erreicht (WBK-T2) (vgl. Kapitel 8). Die evidenzbasierte Validierung des Konstrukts WBK-soek wird mittels drei Ansätze verfolgt (vgl. Kapitel 9, 11 und 12).

TEIL B: TESTENTWICKLUNG

5 Psychologische Tests und Testtheorie

In diesem Kapitel werden psychologische Tests (Kapitel 5.1) sowie die klassische und die probabilistische Testtheorie (Kapitel 5.2) behandelt. Aus der probabilistischen Testtheorie werden insbesondere die Rasch-Modelle vorgestellt.

5.1 Psychologische Tests

Dieses Kapitel erläutert die Definition und die Merkmale psychologischer Tests (Kapitel 5.1.1) sowie die Gütekriterien psychologischer Test (Kapitel 5.1.2).

5.1.1 Definition und Merkmale psychologischer Tests

Ein psychologischer Test ist «[...] ein wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Persönlichkeitsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad der individuellen Merkmalsausprägung» (Lienert, 1969, S. 7, zit. nach Rost, 2004, S. 17). Diese klassische Definition eines psychologischen Tests muss nun wie folgt präzisiert werden:

- Mit diesem Verfahren sind im engeren Sinne nur Tests gemeint, die von den Testpersonen nicht willentlich in eine bestimmte Richtung verfälscht werden können. Im weiteren Sinne gehören dazu nicht nur Tests, sondern auch standardisierte Fragebögen, standardisierte Beobachtungen und strukturierte Interviews (Rost, 2004, S. 17).
- Das Verfahren muss hinsichtlich Durchführung und Auswertung an einer grösseren Stichprobe erprobt und für Dritte nachvollziehbar beschrieben sein («Routineverfahren») (Rost, 2004, S. 17). Dazu braucht es u.a. ein Testheft mit Items für die Testpersonen und ein Testmanual zur für die Testanwender*innen (Döring & Bortz, 2016, S. 431).
- Für das Verfahren muss es eine Theorie geben, ob und welche Aussagen aus den Testergebnissen über die Testpersonen abgeleitet werden können («wissenschaftlich») (Rost, 2004, S. 17). Diese „Theorie“ ist dreifach ausgerichtet (Döring & Bortz, 2016, S. 431): *inhaltsbezogen* (Annahmen über das zu messende Konstrukt), *methodenbezogen* (sog. Testtheorie: Aussagen über Eigenschaften des Tests und der Testitems) und *evidenzbezogen* (empirische Belege zur Stützung der Theorie).
- Das Verfahren kann schriftlich als *Paper-and-Pencil-Test* und *elektronischer Test* durchgeführt werden oder mündlich als strukturiertes oder semi-strukturiertes Interview (Döring & Bortz, 2016, S. 431). Beim *Paper-and-Pencil-Verfahren* bearbeiten die Testpersonen den Test mit Stift und Papier in Anwesenheit der Testanwender*innen. Das elektronische Verfahren verwendet

Computer bzw. Computer-Software, wobei die Testpersonen den Test entweder offline oder online bearbeiten.

- Das Verfahren kann als *psychometrischer Test* vollstrukturiert oder als *projektiver Test* nicht strukturiert oder teilstrukturiert sein (Döring & Bortz, 2016, S. 434). Bei psychometrischen Tests werden den Testpersonen standardisierte Items (Frage/Aussage/Aufgabe mit festem Antwortformat) vorgelegt, die Antworten werden nach präzisen Schemata bewertet und quantitativ psychometrisch ausgewertet. Bei projektiven Tests verwenden die Testanwender*innen offene und unstrukturierte Aufgaben (z.B. ein Bild interpretieren, ein Bild malen), die Antworten werden i.d.R. qualitativ interpretativ ausgewertet.
- Das Verfahren muss ein relativ stabiles und konsistentes Merkmal von Personen erfassen, z.B. Intelligenz, Kompetenz, Einstellung («Persönlichkeitsmerkmal», psychologisches Merkmal) (Rost, 2004, S. 17). Je nach Art des zu messenden psychologischen Merkmals unterscheidet man zwischen Leistungstests und Persönlichkeitstest. Beim *Leistungstest* bearbeiten die Testpersonen Items, für die es objektiv richtige und falsche Lösungen gibt. Es werden zwei Formen von Leistungstests unterschieden (Döring & Bortz, 2016, S. 434ff.; Jonkisz, Moosbrugger, & Brandt, 2012, S. 29f.; Rost, 2004, S. 43ff.). Der *Geschwindigkeitstest* (*speed test*) verwendet „einfache“ Items, die von den allen Probanden gelöst werden können und begrenzt die Bearbeitungszeit. Der *Niveautest* (*power test*) verwendet „schwierige“ Items, die nicht von allen Probanden korrekt gelöst werden können und erhöht das Niveau der Items sukzessive über die Testlänge hinweg. Beim *Persönlichkeitstest* beantworten die Testpersonen im Sinne einer Selbstauskunft Items über ihr emotionales, motivationales oder soziales Erleben und Verhalten, das situations- und zeitübergreifend sein soll. Die Selbstauskunft kann sich auf Bewertungen (Einstellungstest), Interessen (Motivations- und Interessensfragebogen), Verhaltensweisen (Verhaltensfragebogen) oder situative Reaktionen (Situationsfragebogen) beziehen (Döring & Bortz, 2016, S. 434ff., S. 457f.; Jonkisz et al., 2012, S. 30f.).
- Das Verfahren kann darauf abzielen, quantitative oder qualitative Aussagen über die Ausprägung des individuellen Personenmerkmals zu machen (Langeheine & Rost, 1988; Rost, 2004, S. 17). Für quantitative Aussagen wird aus den Antwortsummen (wie viele Fragen wurden bejaht bzw. korrekt gelöst?) ein kontinuierliches Personenmerkmal gebildet (*Latent-Trait-Modelle*, vgl. Kapitel 5.2.2.b). Für qualitative Aussagen wird das Antwortmuster (welche Fragen wurden mit Ja bzw. Nein beantwortet?) berücksichtigt, um daraus ein kategoriales Personenmerkmal zu bilden (*Latent-Class-Modelle*, vgl. Kapitel 5.2.2.b).
- Mit dem Verfahren kann ein unidimensionales oder multidimensionales theoretisches Konstrukt repräsentiert werden (Döring & Bortz, 2016, S. 436; Jonkisz et al., 2012, S. 28f.). Bei einem *eindimensionalen Test* werden alle Items zu einem einzigen Testwert verrechnet; dieser

Testwert informiert über die Ausprägung des Konstrukts (z.B. allgemeine Intelligenz). Bei einem *mehrdimensionalen Test* werden die Items zu Subtestwerten verrechnet, wobei jeder Subtest eine Dimension des Konstrukts darstellt (z.B. sprachliches, numerisches und räumliches Denken).

Von psychologischen Tests abzugrenzen sind nicht-wissenschaftliche Varianten von Leistungstests im Bildungsbereich (z.B. Klausuren, Prüfungen, Klassenarbeiten), die von Lehrer*innen *ad hoc* erstellt werden und keinem wissenschaftlichen Konstruktions- und Evaluationsprozess durchlaufen. Ebenfalls nicht zu den psychologischen Tests zählen nicht-wissenschaftliche Varianten von Persönlichkeitstests (sog. Psychotests in Zeitschriften und auf Websites), die nicht nach wissenschaftlichen Kriterien, sondern nach „gesundem Menschenverstand“ konstruiert werden und primär der Unterhaltung des Publikums dienen (Döring & Bortz, 2016, S. 434).

5.1.2 Gütekriterien psychologischer Tests

Mit einem psychologischen Test soll ein bestimmtes latentes, d.h. nicht direkt beobachtbares Persönlichkeitsmerkmal gemessen werden, um wissenschaftliche Aussagen zu machen und praktische Entscheidungen zu treffen. Daher müssen bei der Entwicklung und Anwendung eines psychologischen Tests – im Gegensatz zu einem „unwissenschaftlichen“ Test (z.B. Klassenarbeit, Fragekatalog) – wissenschaftliche Standards und Kriterien eingehalten und überprüft werden (Döring & Bortz, 2016, S. 440; Moosbrugger & Kelava, 2012a, S. 8).

Die offiziellen *Standards for Educational and Psychological Testing*, kurz *Standards*, folgen einem modernen Verständnis der „Testqualität“ und insbesondere der Validität (AERA, APA, & NCME, 2014; Messick, 1980, 1988, 1995). Bei diesem muss anhand theoretischer Begründungen und empirischer Befunde differenziert aufgezeigt werden, inwiefern die vorgenommene Interpretation der Testergebnisse und die beabsichtigte Testanwendung gültig sein soll (Döring & Bortz, 2016, S. 440f.; Hartig, Frey, & Jude, 2012, S. 144). Auf dieses moderne Verständnis der „Testqualität“ und das neue Konzept der Validität wird weiter unten eingegangen (vgl. Kapitel 5.2).

Im historischen Verständnis ist „Testqualität“ ein Merkmal des Tests und wird in einen Katalog von Testgütekriterien aufgeteilt. Diese Typologie der Testgütekriterien wird – trotz des Paradigmenwechsels in den *Standards* – nach wie vor in Methodenlehrbüchern und einschlägigen Zeitschriften verwendet (Newton & Shaw, 2013). Ein psychologischer Test muss demnach gewisse Kriterien, sog. *Gütekriterien*, erfüllen, um wissenschaftlichen Ansprüchen zu genügen.

Zu den wichtigsten Gütekriterien, sog. *Hauptgütekriterien* psychologischer Tests zählen *Objektivität*, *Reliabilität* (Zuverlässigkeit) und *Validität* (Gültigkeit). Die Hauptgütekriterien stehen in einer hierarchischen Beziehung zueinander (Rost, 2004, S. 33): Objektivität ist eine logische Voraussetzung für Reliabilität, Reliabilität wiederum eine logische Voraussetzung für Validität. Ein

Test, der bei einem anderen Testleiter und in einem anderen Testraum bei denselben Testpersonen andere Ergebnisse hervorbringt, also nicht objektiv ist, kann auch nicht reliabel sein. Ein Test, der bei denselben Testpersonen wiederholt eingesetzt wird und andere Ergebnisse erzeugt, also nicht reliabel ist, kann auch nicht valide sein. Weitere Gütekriterien, sog. *Nebengütekriterien* psychologischer Tests sind Skalierung, Normierung (Eichung), Testökonomie, Nützlichkeit, Zumutbarkeit, Unverfälschbarkeit und Fairness (vgl. Kapitel 5.1.2.d).

5.1.2.a Objektivität

Mit *Objektivität* ist gemeint, inwiefern das Testergebnis einer Testperson unabhängig von äusseren Einflüssen ist, insbesondere von anderen Personen (z.B. Testleiter*innen, Itemkodierer*innen) und anderen „Dingen“ (z.B. Testsituation, Itemauswahl) (Rost, 2004, S. 33, S. 39f.). Objektiv ist ein Test, wenn unterschiedliche Testanwender*innen bei denselben Testpersonen zu denselben Testergebnissen kommen. Die Objektivität des Tests gilt i.d.R. als erfüllt, wenn das Testmanual zur Durchführung, Auswertung und Interpretation präzise Beschreibungen und Beispiele gibt und sich die Testanwender*innen seriös daranhalten. Insofern ist Objektivität bei vollstrukturierten Tests, d.h. psychometrischen Tests, relativ unkritisch, wenn nicht gar obsolet, da die Testwender*innen durch die Standardisierung des Tests wenig bis keinen Spielraum für subjektive Interpretationen haben (Döring & Bortz, 2016, S. 442).

Bei der Objektivität i.S.e. Unabhängigkeit von anderen *Personen* lassen sich verschiedene Elemente unterscheiden (Döring & Bortz, 2016, S. 443f.; Moosbrugger & Kelava, 2012a, S. 8ff.; Rost, 2004, S. 39):

- *Durchführungsobjektivität*: Die Testergebnisse sind unabhängig von der Person, die den Test den Testpersonen vorlegt. Dazu bedarf es einer sorgfältigen Auswahl und ausführlichen Schulung der Testleiter*innen und eines Manuals für die Testdurchführung (Testmanual/Erhebungsmanual: Testmaterial, Instruktionen an Testpersonen, Umgang mit Rückfragen der Testpersonen, standardisierter Testablauf, festgelegte Testdauer, Vorlage für Testprotokolle).
- *Auswertungsobjektivität*: Die Testergebnisse sind unabhängig von der Person, die den Test auswertet. Bei Testitems mit gebundenem Antwortformat (z.B. Multiple-Choice) ist die Auswertungsobjektivität i.d.R. unkritisch, da sie semi-automatisch oder vollautomatisch erfolgen kann. Hingegen braucht es bei Testitems mit freiem Antwortformat eine sorgfältige Auswahl und ausführliche Schulung der Kodierer*innen und ein Manual für die Testauswertung (auch Testmanual/Kodierungsmanual: Kodierungs-/Bewertungsschema für jedes Item, standardisiertes elaboriertes Verfahren für die Überprüfung der einheitlichen Kodierung, d.h. Doppelkodierung und Konsensdiskussion, Kennwerte für Inter-Kodierer-Reliabilität).

- *Interpretationsobjektivität*: Die Testergebnisse sind unabhängig von der Person, die den Testwert interpretiert. Dazu braucht es ein Manual, mit dessen Hilfe die Testergebnisse norm- oder kriteriumsorientiert interpretiert werden.

Bei der Objektivität i.S.e. Unabhängigkeit von anderen *Dingen* wird im Wesentlichen zwischen zwei Elementen unterschieden, nämlich Testsituation und Itemauswahl (Rost, 2004, S. 39f.). Beim Objektivitätselement der *Testsituation* sollte das Testergebnis unabhängig davon sein, in welcher Situation der Test durchgeführt wurde. Beim Objektivitätselement der *Itemauswahl* sollte das Testergebnis unabhängig davon sein, welche Items aus einem hypothetischen Itemuniversum ausgewählt wurden. Mit anderen Worten, das Testergebnis sollte für jede beliebige Itemauswahl aus dem Itemuniversum dasselbe sein, da jeder Test nur eine begrenzte Anzahl vom Items umfassen kann. Die Interpretation des Testwerts hinsichtlich des Personenmerkmals bezieht sich also nicht nur auf die ausgewählten Items, sondern auf ein ganzes Itemuniversum mit unendlich vielen Items. Der Objektivitätsaspekt der *Itemauswahl* wird auch als *spezifische Objektivität* bezeichnet; sie ist sowohl eine Eigenschaft des Tests als auch eine Eigenschaft des jeweiligen Testmodells, namentlich des Rasch-Modells (vgl. Kapitel 5.2.3).

5.1.2.b Reliabilität (Zuverlässigkeit)

Mit *Reliabilität* (Zuverlässigkeit, Präzision, Messgenauigkeit) ist gemeint, wie genau der Test ein Personenmerkmal misst. Reliabilität bezieht sich auf die numerische Präzision der Messung, unabhängig davon, was der Test misst oder was er messen soll. Um die Reliabilität eines psychometrischen Tests empirisch zu prüfen, wurden im Rahmen der klassischen Testtheorie vier Methoden entwickelt (vgl. Schermelleh-Engel & Werner, 2012): Test-Retest-Reliabilität, Paralleltest-Reliabilität, Split-Half-Reliabilität, interne Konsistenz. Das geläufigste Verfahren ist die interne Konsistenz, bei welcher der Reliabilitätskoeffizient Cronbachs α geschätzt wird (Cronbach, 1951).

Gemäss der klassischen Testtheorie ist ein Test reliabel, wenn der Testwert (*test score*) die wahre Ausprägung des Persönlichkeitsmerkmals (*true score*) genau abbildet und nicht oder nur geringfügig durch einen Messfehler (*error score*) beeinträchtigt ist (Döring & Bortz, 2016, S. 442f.). Formal ist die Reliabilität eines Tests definiert als das Verhältnis von wahrer Varianz (Varianz der *true scores*) zu beobachteter Varianz (Gesamtvarianz der *test scores*). Eine hohe Reliabilität bedeutet, dass ein grosser Anteil der Varianz der Testwerte aufgrund der Varianz der wahren Personenwerte entstanden ist (Rost, 2004, S. 38f.). Der Reliabilitätskoeffizient (Rel) bewegt sich im Wertebereich zwischen $0 \leq \text{Rel} \leq 1$, wobei 1 einen perfekten messfehlerfreien Testwert bedeutet und 0 einen ausschliesslich durch Messfehler generierten Testwert (Moosbrugger & Kelava, 2012a, S. 11).

Die Reliabilität eines Tests soll möglichst hoch sein, aber perfekte Reliabilität ist ein theoretischer Idealfall. In der Praxis hängt die Reliabilität nämlich von verschiedenen Faktoren ab, so

dass Fehlereinflüsse (z.B. durch Müdigkeit der Testpersonen, Missverständnisse bzgl. der Items) nie ganz auszuschliessen sind (Döring & Bortz, 2016, S. 465). Nach der konventionellen Nomenklatur wird ein Reliabilitätskoeffizient über .90 als „hoch“ und einer über .80 als „ausreichend“ eingestuft, er sollte jedoch .70 nicht unterschreiten (Bühner, 2011, S. 81). Allerdings sollten diese Faustregeln bei der Beurteilung der Reliabilität eines Tests nicht mechanisch angewendet, sondern kontextspezifisch interpretiert werden. Im Sinne eines modernen Verständnisses von Testqualität (vgl. AERA et al., 2014) sollten insbesondere weitere inhaltliche Aspekte berücksichtigt werden (Schermelleh-Engel & Werner, 2012, S. 135ff.), z.B.: Art des gemessenen Merkmals und Vergleich mit konkurrierenden Tests, Individual- versus Kollektivdiagnostik, Kosten-Nutzen-Abwägung, Objektivität als Voraussetzung der Reliabilität, Passung von Homogenität/Heterogenität des Verfahrens.

Basiert der Test auf der klassischen Testtheorie (KTT), so liefert der geschätzte Reliabilitäts-Koeffizient Cronbachs α eine *pauschale Genauigkeitsschätzung* der Personen-Testwerte, weil der wahre Wert und der Messfehler nicht für jede Testperson bestimmbar sind. Basiert der Test auf der probabilistischen Testtheorie (auch: Item-Response-Theorie, IRT), können die Item- und Personenparameter mit der Maximum-Likelihood-Methode und/oder der bayesianischen Methode geschätzt werden (vgl. Kapitel 5.2.3.c). Daraus resultieren die beiden ML-Schätzer *maximum likelihood estimate* (MLE) und *weighted likelihood estimate* (WLE) sowie der bayesianische Schätzer *expected a posteriori* (EAP), deren Reliabilität – zusätzlich zum klassischen Reliabilitäts-Koeffizienten Cronbachs α – für die Beurteilung der Schätzgenauigkeit der Personenparameter zur Verfügung stehen (Frey, 2012, S. 284f.). Zudem ist in der probabilistischen Testtheorie mithilfe der Informationsfunktion eine *spezifische testwertabhängige Genauigkeitsschätzung* der einzelnen Personen-Testwerte möglich (Döring & Bortz, 2016, S. 443; Moosbrugger & Kelava, 2012a, S. 13ff.). Aus der Informationsfunktion lässt sich die Varianz der Parameterschätzwerte (sog. Schätzfehlervarianz) berechnen. Eine hohe Reliabilität zeigt sich darin, dass die Schätzfehlervarianz der Personenparameter sehr gering ist (Rost, 2004, S. 356ff.).

5.1.2.c Validität (Gültigkeit)

Das Konzept der Validität hat eine Metamorphose durchlebt (Geisinger, 1992). Nach der traditionellen Auffassung bezieht sich Validität auf den *Test per se*, nach der modernen Auffassung auf die *Interpretation der Testergebnisse*. Nachfolgend werden zum besseren Verständnis beide Konzepte erläutert. Jedoch vertritt die Autorin die moderne Auffassung der Validität und verfolgt in dieser Dissertation eine evidenzbasierte Validierung (vgl. Kapitel 9, 11 und 12).

Nach der *traditionellen* Auffassung ist mit Validität (Gültigkeit) gemeint, inwiefern der Test das Merkmal misst, das er messen soll und nicht ein anderes (Rost, 2004, S. 33). Ein Test ist demnach valide, wenn das vom Test gemessene Merkmal mit dem zu messenden Merkmal des Konstrukts theoretisch und empirisch übereinstimmt. Eine hohe Validität erlaubt es, das Verhalten in der

Testsituation für das Verhalten ausserhalb der Testsituation, d.h. in realen Situationen zu generalisieren (Moosbrugger & Kelava, 2012a, S. 13f.). Insofern ist Validität das umfassendste und komplexeste Gütekriterium; ihr kommt denn auch bei der Testentwicklung und Testanwendung die wichtigste Bedeutung zu (Döring & Bortz, 2016, S. 445; Hartig et al., 2012, S. 144).

Die historische Nomenklatur unterscheidet verschiedene *Typen* von Validität (vgl. Hartig et al., 2012). So werden in der ersten Ausgabe der *Standards* vier Typen von Validität genannt: *predictive validity*, *concurrent validity*, *content validity* und *construct validity* (APA, 1954). In der zweiten und dritten Ausgabe der *Standards* sind es drei Typen: *criterion validity* (Kriteriumsvalidität), *content validity* (Inhaltsvalidität) und *construct validity* (Konstruktvalidität) (APA, AERA, & NCME, 1966, 1974). Diesen Typen von Validität ist gemein, dass sie als Eigenschaften des psychologischen *Tests* zu verstehen sind und dass je nach Testeinsatz ein anderer Typ herangezogen werden kann.

- *Inhaltsvalidität (content validity)*: Die einzelnen Testitems repräsentieren das Konstrukt inhaltlich vollständig und zutreffend. Für die Inhaltsvalidierung braucht es genaue Test- bzw. Inhaltspezifikationen basierend auf einer Domänenanalyse sowie eine Beurteilung des Testinhalts durch Experten. Die Inhaltsvalidierung erfolgt theoretisch-argumentativ, nicht numerisch (Döring & Bortz, 2016, S. 446; Hartig et al., 2012, S. 148ff.).
- *Konstruktvalidität*: Die Testergebnisse sind im Sinne des theoretischen Konstrukts interpretierbar. Für die Konstruktvalidierung ist ein theoretisch und/oder empirisch fundiertes Netz an *a priori* formulierten Hypothesen bezüglich des zu messenden Konstrukts – idealerweise ein «nomologisches Netz» (Cronbach & Meehl, 1955) – erforderlich. Anhand der Testergebnisse können die theoriegeleiteten Hypothesen zur inneren Struktur des Konstrukts und/oder zur Beziehung des Konstrukts mit anderen Konstrukten bestätigt werden (Hartig et al., 2012, S. 153ff.; Moosbrugger & Kelava, 2012a, S. 16ff.). Es sind zwei Typen der Konstruktvalidität zu unterscheiden:
 - Die *faktorielle Validität (factorial validity)* bezieht sich auf die faktorielle Struktur bzw. Dimensionalität des Tests. Bei einem *eindimensionalen* Test müssen alle Items auf einen einzigen Faktor laden. Bei einem *mehrdimensionalen* Test müssen die jeweiligen Items der Subtests, also die inhaltlich zusammengehörigen Items, auf ein und denselben Faktor laden (Döring & Bortz, 2016, S. 446; Jonkisz et al., 2012, S. 28f.). Die faktorielle Validität kann durch zwei Verfahren der KTT untersucht werden (vgl. Moosbrugger & Schermelleh-Engel, 2012): hypothesengenerierend mittels explorativer Faktorenanalyse (EFA) oder hypothesenprüfend mittels konfirmatorischer Faktorenanalyse (CFA). Zudem kann die faktorielle Validität mit Testmodellen der IRT untersucht werden (Hartig et al., 2012, S. 162).

- Bei der *konvergenten Validität* (*convergent validity*) soll der zu validierende Test relativ hoch mit eng verwandten Konstrukten bzw. konstrukt nahen Tests korrelieren. Bei der *divergenten Validität* (*divergent validity*, *discriminant validity*) soll der zu validierende Test gar nicht oder nur gering mit entfernt verwandten Konstrukten bzw. konstrukt fremden Tests korrelieren (Döring & Bortz, 2016, S. 446; Hartig et al., 2012, S. 152ff.). Die Konstruktvalidität kann mittels experimenteller, bivariater (Korrelationsanalyse) und multivariater (z.B. Strukturgleichungsmodelle) Analyseverfahren untersucht werden (Döring & Bortz, 2016, S. 446; Hartig et al., 2012, S. 156ff.).
- **Kriteriumsvalidität:** Vom Verhalten der Testpersonen innerhalb der Testsituation kann auf das Verhalten ausserhalb der Testsituation (Kriterium, Aussenkriterium) geschlossen werden. Je grösser der Zusammenhang zwischen Testwerten und Kriterium, desto bedeutsamer ist es für diagnostische Entscheidungen (Moosbrugger & Kelava, 2012a, S. 18). Es sind drei Typen der Kriteriumsvalidität zu unterscheiden (Döring & Bortz, 2016, S. 446; Hartig et al., 2012, S. 162ff.):
 - *Retrospektive Validität* (*retrospective validity*): Der Kriteriumswert wird zeitlich vor dem zu validierenden Testwert erhoben (z.B. ein Schulleistungstest wird mit vergangenen Schulnoten validiert).
 - *Konkurrente Validität* (Übereinstimmungsvalidität, *concurrent validity*): Der Kriteriumswert wird zeitgleich mit dem zu validierenden Testwert erhoben (z.B. ein Schulleistungstest wird mit aktuellen Schulnoten validiert).
 - *Prognostische Validität* (Vorhersagevalidität, *predictive validity*): Der Kriteriumswert wird zeitlich nach dem zu validierenden Testwert erhoben (z.B. ein Berufseignungstest wird über den zukünftigen Berufserfolg validiert).

Eine alternative und ebenfalls traditionelle Unterscheidung bieten zwei *quantitative* Ansätze der Validität für psychologische Tests, die interne und die externe Validität (Rost, 2004, S. 34ff.):

- **Interne Validität:** Für die Prüfung der internen Validität der Testergebnisse braucht es die Datenmatrix des Tests, von der ein systematischer Zusammenhang zwischen den Testantworten angenommen wird. Die Testergebnisse gelten als intern valide, wenn sich die Annahmen über die Item-Antworten anhand der Datenmatrix bestätigen lassen. Interne Validität der Testergebnisse ist also gleichzusetzen mit der Gültigkeit des zugrunde gelegten Testmodells und ist eine zentrale Frage der Testtheorie. Allerdings ist die Frage nach der Gültigkeit eines Testmodells oft nicht absolut (ja/nein), sondern relativ zu alternativen Testmodellen (besser/schlechter) zu beantworten.

- *Externe Validität*: Für die Prüfung der externen Validität der Testergebnisse braucht es eine externe Variable (Kriteriumsvariable, Aussenkriterium), d.h. eine Variable, die ausserhalb des Tests gebildet wurde, jedoch dasselbe Personenmerkmal wie der Test erfasst. Die externe Validität der Testergebnisse widerspiegelt sich durch die *Korrelation* der internen Variablen mit der externen Variablen: Ein hoher positiver Korrelationskoeffizient bedeutet, dass die Testergebnisse extern valide sind. Externe Validität kann also mit konventionellen statistischen Analyseverfahren untersucht werden, welche nicht zur Testtheorie gehören.

Nach der *modernen* Auffassung bezieht sich Validität nicht auf den Test, sondern auf die *Interpretation der Testergebnisse*. Validität wird nicht in verschiedene Typen unterteilt, sondern als einheitliches Konzept verstanden, das theoretisch und empirisch eng an das zu messende Konstrukt geknüpft ist (Cronbach, 1988; Cronbach & Meehl, 1955; Messick, 1988, 1989, 1995). Seit der vierten Ausgabe der *Standards* bezieht sich Validität ebenfalls auf die gültige Interpretation der Testwerte und auf den gültigen Testeinsatz (AERA, APA, & NCME, 1985, 1999; AERA et al., 2014). Validität bezieht sich auf den Grad, nach welchem Theorie und Empirie, also alle gesammelte Evidenz, die vorgeschlagene Interpretation der Testergebnisse und die vorgeschlagene Testanwendung stützen (Hartig et al., 2012, S. 144). Für die Validität werden in den *Standards* fünf *Evidenzquellen* benannt (AERA et al., 2014) (vgl. Kapitel 6.2).

«Validity refers to the degree to which evidence and theory supports the interpretation of test scores for proposed uses of tests. [...] [It] is, therefore, the most fundamental consideration in developing tests and evaluating tests» (AERA et al., 2014, S. 11).

5.1.2.d Nebengütekriterien

Nachfolgend werden die sieben *Nebengütekriterien* psychologischer Tests erläutert und gängige Verfahren für ihre Überprüfung beschrieben.

- *Skalierung/Skalierbarkeit (scalability)*: Der Testwert wird durch eine gültige Verrechnungsvorschrift aus den Testitems gebildet. In der Klassischen Testtheorie (KTT) werden die Items i.d.R. zu einem ungewichteten Summen- oder Mittelwert verrechnet, ohne Prüfung der messtheoretischen Eigenschaften. In der Probabilistischen Testtheorie (IRT) werden nur testmodellkonforme Items für den Testwert verwendet, nachdem ihre Skalierbarkeit einzeln empirisch geprüft wurde (Döring & Bortz, 2016, S. 449f.; Moosbrugger & Kelava, 2012a, S. 18f.).
- *Normierung/Eichung (standardization)*: Für den Test liegen aktuelle Normwerte, d.h. durchschnittliche Testergebnisse repräsentativer Vergleichsstichproben vor, die eine normorientierte Interpretation der individuellen Testergebnisse ermöglichen (sog. normorientiertes Testen). Die Testrohwerte werden anhand der Normwerte in standardisierte

Testwerte überführt (z.B. z-Wert, Prozentrang). Für die Normierung des Tests braucht es eine repräsentative Eichstichprobe ($N > 300$) mit unterschiedlicher soziodemografischer Zusammensetzung (i.d.R. Geschlecht, Alter, Schulbildung); sie sollte *spätestens* nach acht Jahren aktualisiert werden (Döring & Bortz, 2016, S. 449f.; Moosbrugger & Kelava, 2012a, S. 19f.; Rost, 2004, S. 41f.).

- *Testökonomie (economical testing)*: Der Erkenntnisgewinn aus dem Test steht in einem vertretbaren Verhältnis zum Durchführungs- und Auswertungsaufwand für die Testanwender*innen: kurze Testzeit, wenig Materialien, einfache Handhabung, als Gruppentest durchführbar, schnelle und einfache Auswertung (Döring & Bortz, 2016, S. 449f.; Moosbrugger & Kelava, 2012a, S. 21).
- *Nützlichkeit (utility)*: Der Test stellt für Wissenschaft und/oder Praxis neue Erkenntnisse zu einem relevanten Persönlichkeitsmerkmal zur Verfügung, zu dem es bislang keinen Test oder nur einen „unguten“ Test gab. Hierfür muss die Relevanz des zu messenden Konstrukts inhaltlich begründet werden (Döring & Bortz, 2016, S. 449f.; Moosbrugger & Kelava, 2012a, S. 22).
- *Zumutbarkeit (reasonableness)*: Der Erkenntnisgewinn aus dem Test steht in einem vertretbaren Verhältnis zur zeitlichen, psychischen und/oder physischen Belastung für die Testpersonen. Für die Beurteilung der Zumutbarkeit kann der Test mit ähnlichen Tests verglichen werden. Die Zumutbarkeit ist eng mit den Anforderungen der Forschungsethik verknüpft, nach der Proband*innen nicht ungebührlich belastet oder beeinträchtigt werden dürfen (Döring & Bortz, 2016, S. 449f.; Moosbrugger & Kelava, 2012a, S. 22f.).
- *Unverfälschbarkeit (non-fakeability)*: Den Testpersonen gelingt es nicht oder kaum, absichtlich ein besonders gutes oder besonders schlechtes Testergebnis zu erzielen, ohne dass es den Testanwender*innen als unplausibel oder gefälscht auffällt. Der Grad der Verfälschbarkeit wird im Pretesting empirisch untersucht, indem Testpersonen Instruktionen für sog. *Faking-Good* und *Faking-Bad* erhalten, und im Testmanual dokumentiert (Döring & Bortz, 2016, S. 449f.; Moosbrugger & Kelava, 2012a, S. 23).
- *Fairness (test fairness)*: Der Test bevorzugt bzw. benachteiligt keine Personen, für die die Testanwendung bestimmt ist (sog. Zielgruppe), aufgrund ihrer Gruppenzugehörigkeit (z.B. Geschlecht, Ethnie, Schulbildung). Die Testfairness ist ebenfalls eng mit der Forschungsethik verbunden. Die Fairness wird für den gesamten Testwert mittels Gruppen-Vergleiche empirisch untersucht (vgl. Kapitel 12.2). In der Probabilistischen Testtheorie (IRT) wird die Fairness einzelner Items mittels einer sog. DIF-Analyse untersucht (Döring & Bortz, 2016, S. 449f.; Moosbrugger & Kelava, 2012a, S. 24f.) (vgl. Kapitel 5.2.3.d).

In dieser Dissertation wird das Gütekriterium der Objektivität entlang der Testentwicklung (vgl. Kapitel 8) sowie der Testdurchführung und -auswertung (vgl. Kapitel 10) berücksichtigt. Das Gütekriterium der Reliabilität wird bei der Testauswertung untersucht (vgl. Kapitel 10.2.1 und 11.3). Das Gütekriterium der Validität schliesslich wird bei der Testauswertung hinsichtlich qualitativer Aspekte (vgl. Kapitel 9) und quantitativer Aspekte (vgl. Kapitel 11 und 12) evaluiert.

5.2 Testtheorie: Formale Modelle psychologischer Tests

Dieses Kapitel gibt einen Überblick über die formalen Modelle der klassischen Testtheorie (KTT) (Kapitel 5.2.1) und der probabilistischen Testtheorie (IRT) (Kapitel 5.2.2). Schliesslich werden aus der IRT das dichotome Rasch-Modell (Kapitel 5.2.3) und das ordinale Partial-Credit-Rasch-Modell erläutert (Kapitel 5.2.4).

Testtheorie beschreibt in einem formalen Modell den systematischen *Zusammenhang* zwischen dem nicht beobachtbaren Personenmerkmal (latente Variable) und dem beobachteten Testverhalten (manifeste Variablen) (Rost, 2004, S. 21). Das Personenmerkmal beeinflusst in der Testsituation die Item-Antworten. Umgekehrt wird bei der Testauswertung von den Item-Antworten bzw. vom Testwert auf das Personenmerkmal geschlossen (Rost, 2004, S. 21) (vgl. Abbildung 5-1). Gegenstand der Testtheorie sind zwei wesentliche Punkte (Döring & Bortz, 2016, S. 461):

- Es werden *Annahmen* darüber getroffen, *wie* aus den empirischen Daten (Item-Antworten, Testwerte), die mit dem psychologischen Test gewonnenen werden, Aussagen über die tatsächliche Ausprägung des Personenmerkmals gemacht werden können (Rückschluss). Solche Aussagen können entweder deterministisch (vgl. KTT, Kapitel 5.2.1) oder probabilistisch (vgl. IRT, Kapitel 5.2.2) sein.
- Es werden *Anforderungen* an einen Test formuliert, damit der Rückschluss vom Testverhalten auf das latente Merkmal „gültig“ ist. Solche Anforderungen sind z.B. (Döring & Bortz, 2016, S. 462, S. 479f.): Der Test besteht aus mehreren Items, die das gemeinsame latente Merkmal erfassen (keine Einzelindikatoren); der Test ist eindimensional, d.h. die Items sind inhaltlich ähnlich und untereinander hoch korreliert (Test-Dimensionalität, Item-Dimensionalität); die Itemwerte dürfen additiv zu einem Testwert verrechnet werden, entweder als Summenwert oder Mittelwert (Skalierbarkeit).

In der Literatur werden zwei testtheoretische Ansätze unterschieden, die unterschiedliche Voraussetzungen an die Variablen stellen und die unterschiedliche Annahmen über die Zusammenhänge der Variablen treffen: die klassische Testtheorie (KTT) und die probabilistische Testtheorie (PTT, auch IRT). Diese beiden Testtheorien sind jedoch nicht als substituierende, sondern als komplementäre Ansätze aufzufassen (Döring & Bortz, 2016, S. 282; Moosbrugger, 2012a, S. 228; Rost, 2004, S. 12).

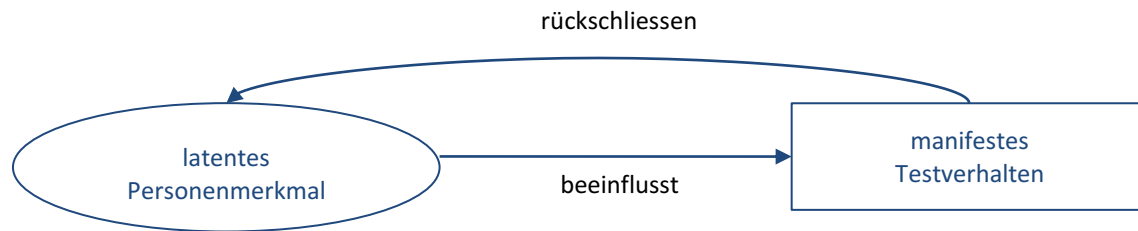


Abbildung 5-1. Gegenstand der Testtheorie

Quelle: Rost (2004, S. 21).

Die Vorteile der KTT liegen in der Überprüfung der *Reliabilität* eines psychologischer Tests mittels Reliabilitätskoeffizienten, z.B. Cronbachs α (Döring & Bortz, 2016, S. 464ff.; Moosbrugger, 2012b, S. 109f.; Moosbrugger & Kelava, 2012a, S. 11ff.; Schermelleh-Engel & Werner, 2012) (vgl. Kapitel 5.1.2.b). Für die Überprüfung der theoriebasierten *Konstruktvalidität* (vgl. Kapitel 5.1.2.c) kann mittlerweile auf den Multitrait-Multimethod-Ansatz (MTMM-Ansatz) zurückgegriffen werden (Döring & Bortz, 2016, S. 472ff.; Schermelleh-Engel & Schweizer, 2012).³⁹ Für die Testentwicklung und Testrevison sind die Kennzahlen der *klassischen Itemanalyse* (vgl. Kapitel 5.2.1.b) hilfreich, um „schlechte“ Items gezielt zu modifizieren (vgl. Döring & Bortz, 2016, S. 475ff.; Kelava & Moosbrugger, 2012).

Zu den Vorteilen der IRT zählen eine inferenzstatistische Prüfung der Modellannahmen (v.a. Item-Homogenität), eine stringendere Prüfung der Testgütekriterien (v.a. Skalierung und Skalenniveau) sowie eine Generalisierbarkeit bzw. Stichprobenabhängigkeit der geschätzten Itemparameter (spezifische Objektivität). Die IRT stellt also deutlich höhere Anforderungen an einen Test bzw. an seine Items, damit sie als „modellgültig“ gelten. Entsprechend können die Item-Antworten und Testwerte basierend auf der IRT valider hinsichtlich des Konstrukts interpretiert werden. Insofern ist die IRT methodisch überzeugender, aber auch statistisch anspruchsvoller als die KTT (Döring & Bortz, 2016, S. 461f., S. 482f., S. 486; Moosbrugger, 2012a, S. 229).

³⁹ Der Multitrait-Multimethod-Ansatz (MTMM-Ansatz, *multitrait multimethod approach*) überprüft systematisch, wie übereinstimmend verschiedene Methoden dasselbe Konstrukt messen (konvergente Validität) und wie abweichend verschiedene Konstrukte durch eine Methode gemessen werden (divergente Validität) (Döring & Bortz, 2016, S. 472).

5.2.1 Klassische Testtheorie (KTT)

Die *klassische Testtheorie* (KTT, *classical test theory*) wird auch als Messfehlertheorie (*true score theory*) bezeichnet. Die Bezeichnung „klassisch“ verweist darauf, dass die KTT wesentlich älter als die IRT ist.

5.2.1.a Annahmen der KTT

Die KTT postuliert einen *deterministischen* Zusammenhang zwischen dem Personenmerkmal und dem Testwert (vgl. Abbildung 5-2). Sie basiert auf der *Annahme*, dass der beobachtete Wert einer Person (Messwert, *test score*) dem wahren Wert dieser Person (*true score*) entspricht, aber durch einen Messfehler (*error score*) verzerrt ist, wobei der Messfehler den wahren Wert vergrößern oder verkleinern kann (Döring & Bortz, 2016, S. 461). Sie orientiert sich also an einem naturwissenschaftlichen Messmodell.

Die KTT formuliert Annahmen über die Messfehler der Messwerte, daher Messfehlertheorie; diese Annahmen werden als Axiome bezeichnet (Döring & Bortz, 2016, S. 463f.). Allerdings werden die Axiome in der Literatur nicht einheitlich gegliedert und nicht einheitlich notiert (vgl. Bühner, 2011; Lienert & Raatz, 1998; Moosbrugger, 2012b). Die KTT setzt kontinuierliche manifeste und kontinuierliche latente Variablen voraus. Da bei psychometrischen Tests die Item-Antworten selten als kontinuierliche manifeste Variablen vorliegen, wird stattdessen meist der gesamte Testwert für die testtheoretischen Analysen verwendet (Eid & Schmidt, 2014, S. 39).

Die *Reliabilitätsschätzung* der klassischen Testtheorie wurde bereits weiter oben skizziert (vgl. Kapitel 5.1.2.b); für eine ausführliche Darstellung sei auf die einschlägige Literatur verwiesen (Döring & Bortz, 2016, S. 464ff.; Moosbrugger, 2012b, S. 109f.; Moosbrugger & Kelava, 2012a, S. 11ff.; Schermelleh-Engel & Werner, 2012). Nachfolgend wird auf die *klassische Itemanalyse* eingegangen, da sie im Rahmen dieser Dissertation als Ergänzung zur probabilistischen Itemanalyse bedeutsam ist (vgl. Kapitel 11.5).

5.2.1.b Klassische Item- und Faktorenanalyse

Die Itemanalyse (*item analysis*) ist ein zentrales Instrument für die Testentwicklung und -evaluation. Das Ziel der Itemanalyse ist es, „problematische“ Items zu identifizieren, sie zu modifizieren oder zu eliminieren und dadurch die Qualität des Tests, v.a. Reliabilität und Validität (vgl. Kapitel 5.1.2.b und 5.1.2.c) zu verbessern. Grundlage für die Itemanalyse sollte eine sog. Eichstichprobe sein, d.h. eine für die Zielpopulation des Tests repräsentative Stichprobe. Gegenstand der Itemanalyse sind die Verteilung der Testrohwerte, die klassischen Item-Schwierigkeiten, die klassischen Item-Trennschärfen sowie die Dimensionalität des Tests (Döring & Bortz, 2016S. 475f.).

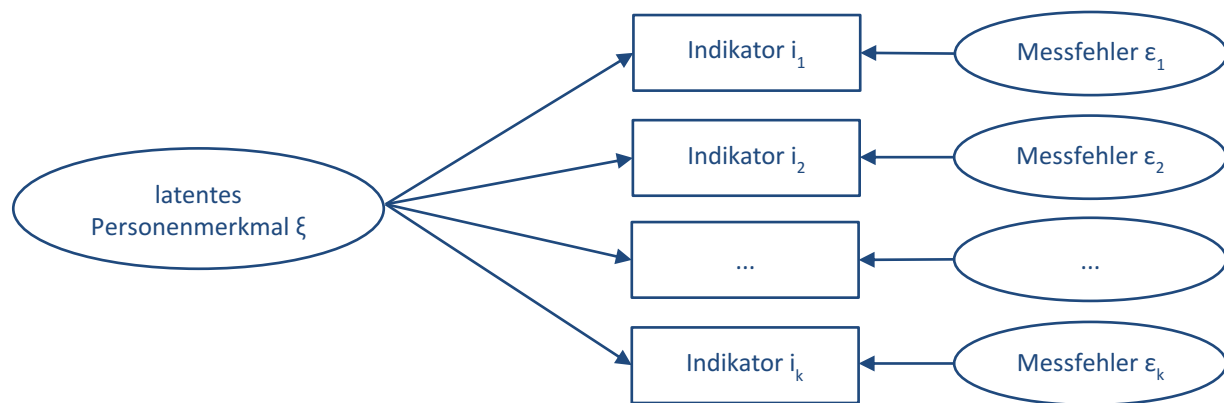


Abbildung 5-2. Annahmen der KTT

Hinweis: Der Messwert eines Indikators (Item, manifeste Variable) wird vom wahren Wert (Personenmerkmal, latente Variable) und einem Messfehler beeinflusst.

Quelle: Eigene Darstellung i.A.a. Embretson (1999, S. 6).

Rohwerte-Verteilung. An der Häufigkeitsverteilung bzw. am Histogramm kann abgelesen werden, wie stark die Testwerte über den gesamten Wertebereich streuen. Normalverteilte Testwerte sind wünschenswert, da viele inferenzstatistische Verfahren (z.B. Regressionsanalyse) die Normalverteilung der metrisch-skalierten Variablen voraussetzen (Döring & Bortz, 2016, S. 476).

Ob eine Stichprobe aus einer (annähernd) normalverteilten Grundgesamtheit stammt, kann empirisch mit dem χ^2 -Test, dem Kolmogoroff-Smirnov-Test, dem Shapiro-Wilk-Test oder dem *Critical-Ratio*-Test überprüft werden (Bortz, Lienert, & Boehnke, 2008, S. 299ff.; Bortz & Schuster, 2010, S. 144ff.). Bei all diesen Tests ist die Nullhypothese jeweils die sog. „Wunschhypothese“: Die Stichprobe stammt aus einer normalverteilten Grundgesamtheit bzw. die empirische Verteilung ist nur zufällig von einer Normalverteilung verschieden. Eine *signifikante* Statistik bei diesen Tests bedeutet, dass die Daten *nicht* normalverteilt sind. Allerdings hängt diese Statistik – wie bei allen Signifikanztests – von der Stichprobengröße ab: Je grösser die Stichprobe, desto grösser ist die Wahrscheinlichkeit, dass die Stichprobe normalverteilt ist (Bortz & Schuster, 2010, S. 146).

Der Kolmogoroff-Smirnov-Test wird für eine Stichprobengröße $N < 120$ empfohlen, da er für grössere Stichproben zu streng ist. Der *Critical-Ratio*-Test kann bei einer Stichprobengröße von $N < 400$ verwendet werden; auch er wird für grössere Stichproben zu streng. Beim *Critical-Ratio*-Test werden Schiefe und Steilheit im Verhältnis zu ihrem Standardfehler gesetzt: $\frac{\text{Schiefe}}{\text{Standardfehler der Schiefe}}$

und $\frac{\text{Steilheit}}{\text{Standardfehler der Steilheit}}$ und der Quotient (*critical ratio*) an der z-Statistik (für $N > 120$) bzw. der t-Statistik (für $N < 120$) geprüft. Eine (annähernde) Normalverteilung wird angenommen, wenn $|z| < 1.96$ bzw. $p \geq .05$ ist (Hopkins & Weeks, 1990). Auf „sinnvolle“ Grenzwerte für Schiefe und Steilheit der empirischen Verteilung wird bei einer Stichprobengröße $N > 400$ zurückgegriffen und folgendes empfohlen: ($| \text{Schiefe} | < 0.5$, $| \text{Steilheit} | < 1.0$) (Lienert & Raatz, 1998).

Eine normalverteilte Häufigkeitsverteilung der Testwerte impliziert, dass der Test angemessene Anforderungen an die Probanden richtet. Eine Abweichung von der Normalverteilung kann unterschiedliche Ursachen haben (Kelava & Moosbrugger, 2012, S. 94f.): Erstens, die Testkonstruktion ist mangelhaft; bei einer *linksschiefen* Verteilung scheint der Test für die Stichprobe „zu einfach“ zu sein ($M_{adj} > 0.50$, $M_{adj} < \text{Median}$, Schiefe < 0), bei einer *rechtsschiefen* Verteilung „zu schwierig“ ($M_{adj} < 0.50$, $M_{adj} > \text{Median}$, Schiefe > 0). Zweitens, die Stichprobe ist heterogen und besteht aus Substichproben, die jeweils normalverteilt sind, aber zusammengenommen eine Mischverteilung bilden. Dieser Umstand ist bei der Testnormierung in Form von verschiedenen Testnormen zu berücksichtigen. Drittens, das gemessene Persönlichkeitsmerkmal ist auch in der Grundgesamtheit nicht normalverteilt.

Item-Schwierigkeit. Die Schwierigkeit eines Items (*item difficulty*) zeigt an, wie gut das Item die Differenzierung im Antwortverhalten der Testpersonen erfassen kann. Ein Item muss also derart konstruiert sein, dass nicht alle Testpersonen dieselbe Antwort geben, d.h. es darf weder „zu einfach“ noch „zu schwierig“ sein (Kelava & Moosbrugger, 2012, S. 76).

Für die klassische Item-Schwierigkeit wird für jedes Item der Schwierigkeitsindex berechnet, der *de facto* ein Leichtigkeitsindex ist: Eine hohe inhaltliche Schwierigkeit des Items korrespondiert mit einem geringen numerischen Wert des Indexes (Döring & Bortz, 2016, S. 476f.). In einem Leistungstest lässt sich die Item-Schwierigkeit vereinfacht als Mittelwert des Items (M_i) ausdrücken (Kelava & Moosbrugger, 2012, S. 77f., S. 79ff.). Der Wertebereich des Item-Mittelwerts liegt zwischen 0 und 1, wobei 0 ein extrem schwieriges und 1 ein extrem einfaches Items anzeigt. Je höher der Item-Mittelwert, desto mehr Testpersonen konnten das Item (korrekt) lösen; je geringer der Mittelwert, desto weniger Personen konnten das Item (korrekt) lösen. Items mit mittlerer Schwierigkeit ($M_i \approx 0.5$) leisten viel Differenzierung, Items mit extremer Schwierigkeit ($M_i \approx 0.0$ oder $M_i \approx 1.0$) leisten wenig bis keine Differenzierung (Kelava & Moosbrugger, 2012, S. 81ff.).

Die konventionellen Toleranzwerte für die klassische Item-Schwierigkeit liegen im Bereich $0.20 \leq M_i \leq 0.80$. Ein Item-Mittelwert von $M_i \leq 0.20$ bedeutet, dass das Item von fast keiner Testperson korrekt gelöst wurde, einer von $M_i \geq 0.80$ bedeutet, dass das Item von fast allen Testpersonen korrekt gelöst wurde (Döring & Bortz, 2016, S. 476f.). Falls mit dem Test auch extreme Merkmalsausprägungen erfasst werden sollen und falls die Items gleichzeitig auch genügend hohe Trennschärfen zeigen, ist ein Bereich von $0.05 \leq M_i \leq 0.95$ tolerierbar (Kelava & Moosbrugger, 2012, S. 87).

Um einen direkten Mittelwertvergleich von dichotomen und polytomen Items sowie von polytomen Items mit unterschiedlichen Antwortkategorien zu erreichen, kann für dichotome Items ein „normaler“ Mittelwert und für polytome Items ein adjustierter Mittelwert berechnet werden (Eid

& Schmidt, 2014, S. 131, S. 226f.). Die Antwortkategorien (k) der dichotomen Items müssen mit 0 und 1 kodiert sein, jene der polytomen Items mit 0 bis k . Der „normale“ Mittelwert (M) berechnet sich als Quotient der Anzahl korrekt gelöster Items (m_R) und der Anzahl vorgelegter Items (m); er entspricht der relativen Häufigkeit der Kategorie 1: $M = \frac{m_R}{m}$. Der adjustierte Mittelwert (M_{adj}) berechnet sich als Quotient aus dem „normalen“ Mittelwert und der höchsten Antwortkategorie ($k-1$):

$$M_{adj} = \frac{M}{k-1} = \frac{\frac{m_R}{m}}{k-1}.$$

Um in den Antworten der Selected-Response-Items das Wissen vom Raten zu isolieren, kann ein korrigierter Mittelwert berechnet werden (Döring & Bortz, 2016, S. 441; Eid & Schmidt, 2014, S. 215f.). Der korrigierte Mittelwert (M_{cor}) reduziert den Zähler des „normalen“ Mittelwerts um eine pauschale Ratekorrektur, die als Quotient der falsch gelösten Items (m_F) und der Anzahl Distraktoren ($k-1$) gebildet wird: $M_{cor} = \frac{m_R - \frac{m_F}{k-1}}{m}$. Durch die Ratekorrektur ergeben sich gelegentlich negative Mittelwerte, die «die naturgemäß nicht zu interpretieren sind; sie geben aber einen Hinweis in dem Sinne, dass es sich um eine sehr schwierige, aber leichter aussehende, d.h. zu einer falschen Lösung ermutigende Aufgabe handelt» (Lienert & Raatz, 1998, S. 75, zit. nach Kelava & Moosbrugger, 2012, S. 80).

Item-Trennschärfe. Die Trennschärfe eines Items (*item discriminability*) gibt an, wie gross der Zusammenhang zwischen der Testpersonen-Differenzierung bei diesem Item (Itemwert) und derjenigen des gesamten Tests (Testwert) ist (Kelava & Moosbrugger, 2012, S. 84). Die Item-Trennschärfe (r_{it}) hängt von der Item-Schwierigkeit (M_i): Je extremer die Item-Schwierigkeit (sehr hohe bzw. sehr geringe Item-Mittelwerte), desto geringer ist die Item-Trennschärfe; Items mit mittlerer Schwierigkeit haben die höchste Trennschärfe (Döring & Bortz, 2016, S. 478f.).

Für Items mit dichotomen Antwortkategorien wird die klassische Item-Trennschärfe als Punkt-Moment-Korrelation zwischen den Punktwerten des Items und dem Gesamtwert des Tests berechnet. Die unkorrigierte Item-Skala-Korrelation ($r_{it-total}$, *item-total correlation*) ist die Korrelation des Itemwerts mit dem gesamten Testwert, die korrigierte Item-Skala-Korrelation ($r_{it-rest}$, *item-rest correlation*, *part-whole-corrected*) ist die Korrelation des Itemwerts mit dem Testwert aller anderen Items des Tests (Kelava & Moosbrugger, 2012, S. 84f.).⁴⁰ Für Items mit polytomen ungeordneten Antwortkategorien (z.B. Multiple-Choice) wird die sog. Kategorien-Trennschärfe als punktbiseriale Korrelation (*point-biserial correlation*, r_{ptbis}) zwischen jedem Kategorienwert und dem Testwert berechnet (OECD, 2005b, S. 123). Der Wertebereich der Trennschärfe liegt zwischen -1 und $+1$. Je

⁴⁰ Diese Korrektur für die klassische Trennschärfe ist besonders relevant bei kurzen Tests, wo ein Item den Testwert stark beeinflussen kann (Le, 2012).

nach Ausprägung der Trennschärfe können drei Fälle unterschieden werden (Kelava & Moosbrugger, 2012, S. 86):

- *Positive Trennschärfe* ($r_{it} \approx +1$): Das Item und der Testwert differenzieren ähnlich bezüglich des Personenmerkmals. Das Item wird von Testpersonen mit hohem Testwert richtig gelöst und von Personen mit geringem Testwert falsch bzw. nicht gelöst.
- *Geringe positive Trennschärfe* ($r_{it} \approx 0$): Das Item hat mit dem Testwert wenig gemein und ist ungeeignet, um zwischen Testpersonen mit hohem Testwert und jenen mit geringem Testwert zu unterscheiden.
- *Negative Trennschärfe* ($r_{it} \approx -1$): Das Item und der Testwert differenzieren bezüglich des Personenmerkmals umgekehrt als beabsichtigt. Das Item wird von Testpersonen mit geringem Testwert richtig gelöst und von Personen mit hohem Testwert falsch oder nicht gelöst. Dies kann durch Mängel in der Instruktion oder im Itemstamm verursacht sein.

Items mit negativen Trennschärfen oder Trennschärfen nahe bei 0 sollten modifiziert oder aus dem Itempool eliminiert werden (Kelava & Moosbrugger, 2012, S. 88). Als strenger Toleranzwert für die Item-Trennschärfe wird in der Literatur $r_{it} \geq .30$ angegeben (Döring & Bortz, 2016, S. 478).⁴¹ Item-Trennschärfen von $r_{it} > .50$ gelten als „sehr gut/hoch“, $.30 \leq r_{it} \leq .50$ als „gut/mittelmässig“ und $r_{it} < .30$ als „ungenügend/gering“ (Döring & Bortz, 2016, S. 478). Bei polytom ungeordneten Items (z.B. Multiple-Choice) sollte die Kategorien-Trennschärfe für die *richtige* Antwortkategorie positiv sein ($r_{ptbis} > 0$) und für alle *falschen* Antwortkategorien (Distraktoren) negativ ($r_{ptbis} < 0$) – oder zumindest nicht höher als jene der korrekten Kategorie (Lienert & Raatz, 1998, S. 101ff.; OECD, 2005b, S. 123). Bei polytom geordneten Items sollte die punktbiseriale Korrelation entlang der Antwortkategorien aufsteigend angeordnet sein (OECD, 2005b, S. 123; 2014b, S. 149).

Dimensionalität. Die Dimensionalität eines Tests gibt an, ob die verschiedenen Items ein Konstrukt (eindimensionaler Test) oder mehrere Teilkonstrukte erfassen (mehrdimensionaler Test mit zwei oder mehreren Subtests). Bei der Dimensionalitätsprüfung wird empirisch untersucht, ob die Anzahl der Faktoren (*test dimensionality*) und ob die Zuordnung der Items zu den Faktoren (*item dimensionality*) den inhaltlichen Spezifikationen des Konstrukts entspricht. In der KTT wird die faktorielle Struktur bzw. Dimensionalität des Messmodells mittels explorativer (EFA) oder konfirmatorischer Faktorenanalyse (CFA) untersucht (vgl. Moosbrugger & Schermelleh-Engel, 2012). Wird für den Test eine eindimensionale faktorielle Struktur angenommen, müssen alle manifesten

⁴¹ In der PISA-Studie 2003 wurde für die Item-Trennschärfe ein $r_{it} \geq .25$ als tolerierbar erachtet (OECD, 2005b, S. 123, S. 127) und für die Distraktoren-Trennschärfe ein $r_{ptbis} < .05$ (OECD, 2005b, S. 127). In PISA-Studie 2012 wurde für die Item-Trennschärfe und die Kategorien-Trennschärfe der richtigen Kategorie ein $r_{ptbis} > .20$ und für die Kategorien-Trennschärfe der falschen Kategorie ein $r_{ptbis} < .0$ für tolerierbar erachtet (OECD, 2014b, S. 149).

Indikatoren (Items) auf *einen* latenten Faktor laden, d.h. die Items des Tests werden durch *ein* Konstrukt repräsentiert. Wird hingegen für den Test eine mehrdimensionale faktorielle Struktur unterstellt, z.B. weil er aus mehreren Subtests besteht, so müssen die manifesten Indikatoren (Items) auf *mehrere* latente Faktoren laden und die inhaltlich zusammengehörigen Indikatoren auf denselben Faktor laden, d.h. die Items eines Subtests repräsentieren jeweils ein Konstrukt (AERA et al., 2014; Cronbach, 1988).

5.2.2 Probabilistische Testtheorie bzw. Item-Response-Theorie (IRT)

Die *probabilistische Testtheorie* (PTT, *probabilistic test theory*, *modern test theory*, *latent trait theory*) wird international auch als Item-Response-Theorie (IRT) bezeichnet.

5.2.2.a Annahmen der IRT

Die IRT postuliert einen *probabilistischen* Zusammenhang zwischen der Ausprägung des Personenmerkmals (latente Variable) und den Item-Antworten (manifeste Variablen) (vgl. Abbildung 5-3). Sie basiert auf der *Annahme*, dass die *Wahrscheinlichkeit* einer bestimmten Item-Antwort abhängig ist von der latenten Merkmalsausprägung der Person (kurz: Personenmerkmal, Personenfähigkeit) und von den Merkmalen des Items (Döring & Bortz, 2016, S. 461). Die geschätzten Personenmerkmale werden *Personenparameter* genannt, die geschätzten Item-Merkmale *Itemparameter*. Je nach Testmodell werden unterschiedliche Art und Anzahl von Itemparametern geschätzt: Item-Schwierigkeit, Item-Trennschärfe, Ratewahrscheinlichkeit.

Die Annahmen der IRT werden in sog. *Testmodellen* spezifiziert und evaluiert. Probabilistische Testmodelle können nur auf Datenstrukturen psychometrischer Tests (vgl. Kapitel 5.1.1) angewendet werden, d.h. auf eine Personen x Item-Datenmatrix. Eine *notwendige* Bedingung für die Gültigkeit eines Testmodells hinsichtlich einer Datenmatrix ist, dass mehrere Items vorliegen, die untereinander korrelieren. Eine *hinreichende* Bedingung ist die Item-Homogenität⁴², d.h. dass die Zusammenhänge zwischen den beobachteten Item-Antworten (manifeste Variablen) alleinig durch das konstruierte Personenmerkmal (latente Variable) erklärt werden, insbesondere von keinem anderen Merkmal systematisch beeinflusst sind (Moosbrugger, 2012a, S. 228f.; Rost, 2004, S. 29ff.).

⁴² Im Kontext des Rasch-Modells wird Item-Homogenität auch als Rasch-Homogenität bezeichnet (Eid & Schmidt, 2014, S. 146).

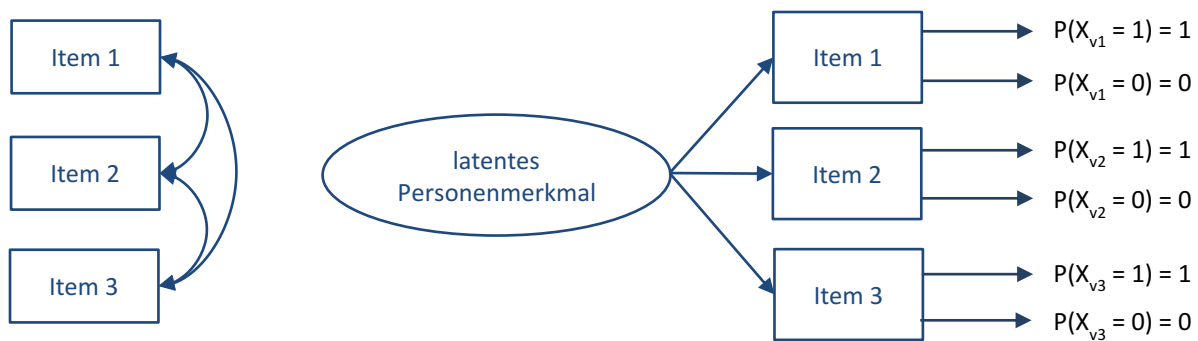


Abbildung 5-3. Annahmen der IRT

Hinweis: Der Zusammenhang zwischen den inhaltlich ähnlichen Items (manifesten Variablen) wird einzig durch das konstruierte Personenmerkmal (latente Variable) erklärt/beeinflusst.

Quelle: Eigene Darstellung i.A.a. Embretson (1999, S. 6) und (Rost, 2004, S. 30).

Die *Modellgültigkeit* ist zentral für eine wissenschaftlich fundierte Interpretation der Testergebnisse, z.B. Aussagen über Unterschiede in den Merkmalsausprägungen der Testpersonen (Rost, 2004, S. 28f.). Um die Gültigkeit von IRT-Modellen hinsichtlich der Item-Homogenität bzw. der lokalen stochastischen Unabhängigkeit zu prüfen, wird *eine bestimmte* Ausprägung des Personenmerkmals (Wert der latenten Variable) *konstant* gehalten und die Korrelationen der Item-Antworten (manifeste Variablen) bei allen Personen mit dieser Ausprägung, d.h. auf dieser „lokalen Stufe“, untersucht. Wenn die Korrelationen zwischen den Antwortvariablen verschwinden, darf von Item-Homogenität ausgegangen werden, d.h. dass das Personenmerkmal einzig ursächlich für die Item-Antworten ist. Wenn hingegen die Korrelationen deutlich von Null abweichen, darf nicht von Item-Homogenität ausgegangen werden; es wäre vielmehr ein Hinweis darauf, dass die Items nicht nur das intendierte Merkmal messen, sondern systematisch auch noch ein anderes nicht-intendiertes Merkmal. Die Item-Homogenität bzw. lokale stochastische Unabhängigkeit wird statistisch mittels des Multiplikationstheorems für stochastisch unabhängige Ereignisse ermittelt (Moosbrugger, 2012a, S. 229f.) (vgl. Kapitel 5.2.3.b).

5.2.2.b Klassifizierung von IRT-Modellen

Probabilistische Testmodelle bzw. IRT-Modelle lassen sich hinsichtlich verschiedener Kriterien klassifizieren (Döring & Bortz, 2016, S. 485ff.; Eid & Schmidt, 2014, S. 34ff.) (vgl. Abbildung 5-4): Skalenniveau der latenten Variable (*latent trait*, *latent class*), Antwortformat der manifesten Variable (*dichotom*, *polytom*), Anzahl der Modellparameter bzw. Itemparameter (Item-Schwierigkeit bzw. Kategorienschwierigkeit, Item-Trennschärfe, Ratewahrscheinlichkeit). Bei der Auswahl des geeigneten IRT-Modells gilt es zwei Aspekte zu berücksichtigen: Das Modell muss das *Testverhalten* korrekt *abbilden* und es muss gut *auf die Daten passen* (Rost, 2004, S. 28f.).

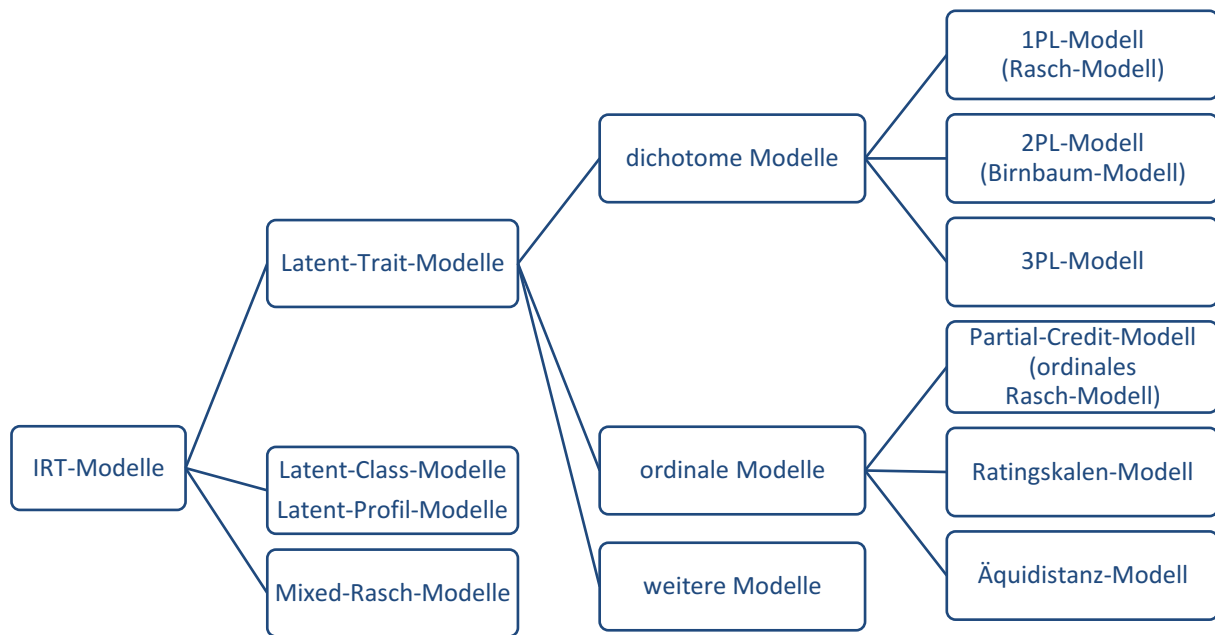


Abbildung 5-4. Klassifikation der IRT-Modelle

Hinweis: Klassifikation nach Skalenniveau der latenten Variablen (2. Ebene), nach Antwortformat der manifesten Variablen (3. Ebene) und nach Anzahl Modellparameter (4. Ebene, oben).

Quelle: Eigene Darstellung i.A.a. Döring & Bortz (2016, S. 485ff.) und Eid & Schmidt (2014, S. 34ff.).

Latent-Trait-Modelle (*latent trait models*) messen ein kontinuierliches kardinalskaliertes latentes Merkmal (z.B. Ausprägung von Intelligenz). *Latent-Class-Modelle* (*latent class models*) messen ein kategoriales nominalskaliertes latentes Merkmal (z.B. Einteilung der Personen in drei Gruppen). *Mixed-Rasch-Modelle* sind eine Kombination aus Latent-Trait- und Latent-Class-Modellen; sie messen ein kardinalskaliertes Merkmal für die durch ein nominalskaliertes Merkmal gebildeten Gruppen.

Dichotome Latent-Trait-Modelle basieren auf Items mit dichotomen Antwortvariablen (im Persönlichkeitstest: nein/ja; im Leistungstest: falsch/richtig). Die Lösungswahrscheinlichkeit des Items wird – in Abhängigkeit der Modellparameter (Personenfähigkeit, Item-Schwierigkeit und u.U. weiteren Itemparametern) – über die logistische itemcharakteristische Funktion (*item characteristic function*) berechnet und grafisch als *Item Characteristic Curve* (ICC) dargestellt. *Ordinale Latent-Trait-Modelle* (auch: Partial-Credit-Modell) basieren auf Items mit polytomen Antwortvariablen mit geordneten Antwortkategorien (im Persönlichkeitstest: Ablehnung/neutral/Zustimmung, nie/selten/oft); im Leistungstest: falsch/teilweise richtig/richtig).⁴³ Die Lösungswahrscheinlichkeit für

⁴³ Es müssen zwei Arten von polytomen Antwortkategorien unterschieden werden. Polytome *ungeordnete* Antwortkategorien resultieren aus gebundenen Antwortformaten mit mehreren Antwortoptionen (auch: Mehrfachwahlantworten, *multiple choice*), die in keiner Rangfolge zueinander stehen (Eid & Schmidt, 2014, S. 97ff.; Jonkisz et al., 2012, S. 39ff.; Rost, 2004, S. 59ff.). Für eine allfällige Distraktoren-Analyse werden zunächst alle Antwortoptionen kodiert (z.B. 1 = Antwort A, 2 = Antwort B, 3 = Antwort C, 4 = Antwort D) und später i.d.R. zu dichotomen Antwortvariablen

jede *Antwortkategorie* des Items wird über die logistische kategoriencharakteristische Funktion (*category characteristic function*) berechnet und grafisch als *Category Characteristic Curve* (CCC) dargestellt. Die Schnittpunkte der CC-Kurven kennzeichnen die Schwellenparameter. Die IC-Funktion eines ordinalen Items ergibt sich aus den CC-Funktionen aller Kategorien.

Das 1PL-Modell (einparametriges logistisches Modell, auch: dichotomes Rasch-Modell) berücksichtigt für die Lösungswahrscheinlichkeit des Items die Personenfähigkeit (Personenparameter, *theta*, θ) und *einen* weiteren Parameter: die Item-Schwierigkeit (Item-Schwierigkeitsparameter, *sigma*, σ). Das 2PL-Modell (zweiparametriges logistisches Modell, auch: Birnbaum-Modell) berücksichtigt für die Lösungswahrscheinlichkeit des Items die Personenfähigkeit sowie *zwei* weitere Parameter: die Item-Schwierigkeit und die Item-Trennschärfe (Item-Trennschärfeparameter, *beta*, β). Ein Item mit einer hohen Trennschärfe differenziert deutlicher zwischen fähigen und weniger fähigen Personen; es hat eine steilere ICC. Das 3PL-Modell (dreiparametriges logistisches Modell) berücksichtigt für die Lösungswahrscheinlichkeit des Items die Personenfähigkeit sowie *drei* weitere Parameter: die Item-Schwierigkeit, die Item-Trennschärfe und die Ratewahrscheinlichkeit (Rateparameter, *gamma*, γ). Ein Item, das durch Raten lösbar ist, hat auch für Personen mit sehr geringer Fähigkeit eine minimale Lösungswahrscheinlichkeit grösser als 0 % ($p > 0.0$).

Nachfolgend werden das dichotome Rasch-Modell und das ordinale Rasch-Modell (Partial-Credit-Rasch-Modell) dargestellt. Ersteres bildet die Grundlage für alle anderen IRT-Modelle. Letzteres erscheint für den WBK-T2 aufgrund seiner Antwortformaten und -kodierungen als geeignet (vgl. Kapitel 8.3.3 und 8.4).

5.2.3 Das Rasch-Modell (*dichotomes 1PL-Latent-Trait-Modell*)

Das Rasch-Modell⁴⁴ (Rasch, 1960) ist ein eindimensionales Modell für *dichotome* Antwortvariablen bzw. Antwortkategorien, bei dem neben dem Personenparameter nur *ein* Itemparameter (1PL-Modell) geschätzt und das latente Merkmal als eine kardinalskalierte Variable (*latent trait*) erfasst wird (Döring & Bortz, 2016, S. 484f.).

Nachfolgend werden die Gleichung, die Annahmen/Eigenschaften des Rasch-Modells, die Methoden für Parameterschätzung und Modellgültigkeitsprüfung (sog. Modelltests) zusammengefasst. Für eine ausführliche formale Darstellung des Rasch-Modells sei auf die einschlägige Literatur verwiesen (z.B. Rost, 2004, S. 115ff.; Strobl, 2010).

rekodiert (z.B. 2 = richtig, ELSE = falsch). Polytope *geordnete* Antwortkategorien resultieren aus freien Antwortformaten, deren Antworten mehrstufig kodiert sind (z.B. 0 = falsch, 1 = teilweise richtig, 2 = vollständig richtig).

⁴⁴ Georg Rasch (1901–1980) war dänischer Statistiker, von 1962 bis 1972 Professor für Statistik an der Universität Kopenhagen.

5.2.3.a Modellgleichung

In der Person x Item-Datenmatrix eines psychologischen Tests sind die Personen (v) in Spalten und die Items (i) in Zeilen angeordnet; in den Zellen der Datenmatrix stehen die kodierten Antworten jeder Person zu jedem Item (x_{vi}), z.B. 0 oder 1. Die Wahrscheinlichkeit, dass Person v bei der Beantwortung von Item i die Antwort x_{vi} gibt, ist $P(X_{vi} = x_{vi})$. P bezeichnet die Lösungswahrscheinlichkeit (englisch: *probability*) einer Person für ein Item, x die kodierte Antwort. Die Wahrscheinlichkeit für das Lösen $P(X_{vi} = 1)$ und die Gegenwahrscheinlichkeit für das Nichtlösen $P(X_{vi} = 0)$ addieren sich zu 1 (Rost, 2004, S. 28; Strobl, 2010, S. 5ff., S. 12f.):

$$P(x_{vi} = 1|\theta_v, \sigma_i) + P(x_{vi} = 0|\theta_v, \sigma_i) = 1,$$

wobei θ (griechisch: *theta*) die Personenfähigkeit und σ (griechisch: *sigma*) die Item-Schwierigkeit bezeichnet. Das Rasch-Modell bzw. die Modellgleichung des Rasch-Modells spezifiziert die Lösungswahrscheinlichkeit p einer Person v für jedes Item i in Abhängigkeit der Personenfähigkeit θ_v und der Item-Schwierigkeit σ_i über eine logistische Funktion⁴⁵ (Moosbrugger, 2012a, S. 236ff.; Rost, 2004, S. 115ff.; Strobl, 2010, S. 7ff., S. 12ff.):

$$P(X_{vi} = x_{vi}|\theta_v, \sigma_i) = \frac{e^{x_{vi} \cdot (\theta_v - \sigma_i)}}{1 + e^{\theta_v - \sigma_i}}, \quad v = 1, \dots, n \text{ und } i = 1, \dots, m$$

Der Item-Schwierigkeitsparameter σ_i ist definiert als jene Ausprägung des Personenparameters θ_v , bei dem die Lösungswahrscheinlichkeit für Item i (und ebenso die Gegenwahrscheinlichkeit) genau 0.5 beträgt: $P(X_{vi} = 1) = P(X_{vi} = 0) = 0.5$. Mit dieser Definition können der Personenparameter θ_v und der Itemparameter σ_i auf derselben Skala (*joint scale*) abgetragen und gemeinsam interpretiert werden (Moosbrugger, 2012a, S. 237).

⁴⁵ Eine logistische Funktion wird allgemein durch folgende Gleichung beschrieben: $y = \frac{e^x}{1+e^x}$, wobei $e = 2.718$. Durch die Logit-Transformation des Wettquotienten $\log \frac{P(X_{vi}=1|\theta)}{P(X_{vi}=0|\theta)} = \log \frac{P(X_{vi}=1|\theta)}{1-P(X_{vi}=1|\theta)} = \theta_v - \sigma_i$ wird der Wertebereich der Lösungswahrscheinlichkeit $[0,1]$ symmetrisch auf den Wertebereich der Logit-Variablen $]-\infty, +\infty[$ projiziert. Die Parameter haben somit das Skalenniveau der kontinuierlichen Logit-Variablen. Die Logit-Variable wird durch Translation in die Differenz der Parameter zerlegt. Die Lösungswahrscheinlichkeit $p = 0.5$ entspricht der Gleichwertigkeit der Parameter $\theta_v = \sigma_i$, d.h. dem Nullpunkt der Logit-Skala (Eid & Schmidt, 2014, S. 152f.; Rost, 2004, S. 166ff.).

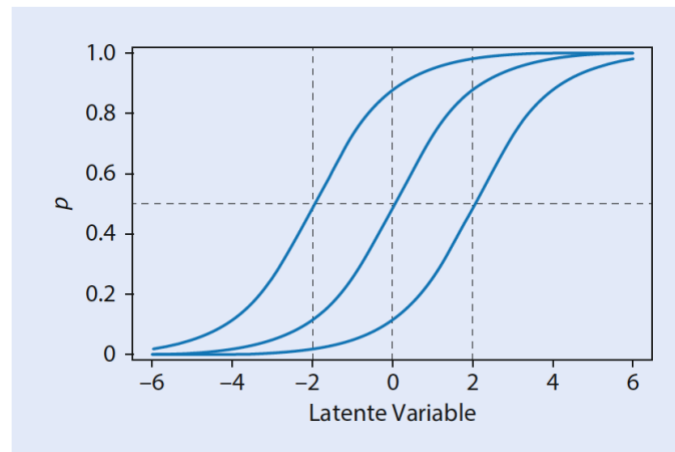


Abbildung 5-5. Item Characteristic Curves für drei dichotome Items

Hinweis: Auf der X-Achse sind Personenfähigkeiten (θ) und Item-Schwierigkeiten (σ) skaliert, auf der Y-Achse sind die Lösungswahrscheinlichkeit (p) abgetragen. Die IC-Funktionen der drei Items haben gleiche Steigungen, aber ungleiche Schwierigkeiten. Bei gleicher Lösungswahrscheinlichkeit (z.B. $p = 0.5$) resultiert für jedes Item eine andere Schwierigkeit bzw. es erfordert eine andere Personenfähigkeit: $\sigma_1 = \theta_v = -2$, $\sigma_2 = \theta_v = 0$, $\sigma_3 = \theta_v = +2$. Bei gleicher Personenfähigkeit (z.B. $\theta_v = 0$) resultiert für jedes Items eine andere Lösungswahrscheinlichkeit: $p(x_{v1}) = 0.1$, $p(x_{v2}) = 0.5$, $p(x_{v3}) = 0.9$.

Quelle: Döring & Bortz (2016, S. 490).

Die logistische Funktion wird als itemcharakteristische Funktion (*item characteristic function*, IC-Funktion) bezeichnet. Sie verläuft monoton, hat ihren Wendepunkt an der Stelle σ_i und einen Wertebereich zwischen 0 und 1. Sie wird grafisch als S-förmige itemcharakteristische Kurve (*item characteristic curve*, IC-Kurve) dargestellt (vgl. Abbildung 5-5). Die X-Achse ist eine kontinuierliche Logit-Skala, auf der die Personenparameter (θ) und die Itemparameter (σ) skaliert werden. Auf der Y-Achse wird die Lösungswahrscheinlichkeit mit dem Wertebereich $[0,1]$ abgetragen. Ein negativer Logit-Wert indiziert eine geringe Personenfähigkeit ($\theta < 0$) bzw. ein einfaches Item ($\sigma < 0$), ein positiver Logit-Wert eine hohe Personenfähigkeit ($\theta > 0$) bzw. ein schwieriges Item ($\sigma > 0$) (Döring & Bortz, 2016, S. 484f.; Rost, 2004, S. 120; Strobl, 2010, S. 10ff.).

Die Lösungswahrscheinlichkeit ist eine *bedingte* Wahrscheinlichkeit; sie hängt von der Personenfähigkeit und der Item-Schwierigkeit ab ($|\theta_v, \sigma_i$). Genauer gesagt, die Lösungswahrscheinlichkeit hängt von der Differenz dieser beiden Parameter ab ($\theta_v - \sigma_i$). Je nach Ausprägung der Parameter können drei Fälle unterschieden werden (Moosbrugger, 2012a, S. 238; Strobl, 2010, S. 9):

- $\theta_v > \sigma_i$: Die Person ist fähiger als das Item schwierig ist; es resultiert eine hohe Lösungswahrscheinlichkeit ($p > 0.5$), sie geht asymptotisch gegen 100 % ($p \approx 1.0$).
- $\theta_v = \sigma_i$: Die Personenfähigkeit entspricht der Item-Schwierigkeit; die Lösungswahrscheinlichkeit ist exakt 50 % ($p = 0.5$).
- $\theta_v < \sigma_i$: Die Person ist weniger fähig als das Item schwieriger als ist, es resultiert eine geringe Lösungswahrscheinlichkeit ($p < 0.5$), sie geht asymptotisch gegen 0 % ($p \approx 0.0$).

Das Rasch-Modell ist ein sog. 1PL-Modell und spezifiziert nur einen Itemparameter, den Item-Schwierigkeitsparameter (σ_i). Daher haben die Items eines Rasch-skalierten Tests zwar unterschiedliche Schwierigkeiten (parallel verschobene IC-Kurven), aber gleiche Trennschärfen (gleiche Steigung der IC-Kurven) (vgl. Abbildung 5-5).⁴⁶ Die *Trennschärfe* bezieht sich auf die Steigung im *mittleren* Bereich der IC-Kurve. Je steiler die IC-Kurve und je grösser die Trennschärfe, desto deutlicher ist der Unterschied in der Lösungswahrscheinlichkeit bei zwei Personen mit nahe beieinander liegenden Fähigkeiten. Mit anderen Worten: Ein Item differenziert zwischen den Personen am besten, wo die Steigung seiner IC-Kurve am grössten bzw. seine Trennschärfe am grössten ist (Strobl, 2010, S. 11f.). Umgekehrt bedeutet dies, dass die Personenfähigkeit am genauesten geschätzt werden kann mit einem Item, das an der Stelle eine hohe Trennschärfe hat, also $\theta_v = \sigma_i$ (Strobl, 2010, S. 35f.).

5.2.3.b Modellannahmen/Modelleigenschaften

Das Rasch-Modell beruht auf *theoretischen Annahmen*, die häufig auch als dessen Eigenschaften betrachtet werden. Bei der Anwendung der Modellgleichung auf die Daten eines konkreten Tests müssen alle Annahmen des Rasch-Modells empirisch geprüft werden, bevor man die theoretisch vorteilhaften Eigenschaften des Rasch-Modells für den Test beanspruchen darf (Strobl, 2010, S. 14). Nachfolgend werden die zentralen Annahmen des Rasch-Modells zusammengefasst. Von diesen fünf Annahmen gilt die „spezifische Objektivität“ explizit nur für Rasch-Modelle bzw. für 1PL-Modelle (Strobl, 2010, S. 14).

Suffiziente Statistik (auch: erschöpfende Statistik) (Strobl, 2010, S. 14ff.). Eine Statistik ist ein einzelner Kennwert, der aus Daten berechnet wird. Eine *suffiziente* Statistik enthält die gesamte relevante Information, die die Daten über den unbekannten Parameter liefern. Im Rasch-Modell sind die Randsummen der Person x Item-Datenmatrix suffiziente Statistiken: für die Person v enthält die Zeilen-Randsumme $r_v = \sum_{i=1}^m x_{vi}$ alle Information über den Personenparameter θ_v ; für das Item i enthält die Spalten-Randsumme $s_i = \sum_{v=1}^n x_{vi}$ alle Information über den Itemparameter σ_i . Für die Schätzung der Personenparameter muss man also nicht wissen, *welche* Items eine Person gelöst hat (Antwortmuster), sondern lediglich *wie viele* Items diese Person gelöst hat (Antwortanzahl). Entsprechend braucht man für die Schätzung der Itemparameter nicht zu wissen, welche Person ein Item gelöst hat, sondern nur wie viele Personen das Item gelöst haben. Eine Person mit einer geringen Fähigkeit löst *wahrscheinlich* nur die einfachen Items des Tests und insgesamt weniger

⁴⁶ In einem 2PL-Modell (vgl. Abbildung 5-4) werden zwei Itemparameter spezifiziert; ein Schwierigkeitsparameter (σ_i) und ein Trennschärfeparameter (β_i) für ungleiche Steigungen der IC-Funktionen. In einem 3PL-Modell werden drei Itemparameter spezifiziert; ein Schwierigkeitsparameter (σ_i), ein Trennschärfeparameter (β_i) und ein Item-Rateparameter (γ_i), so dass die IC-Funktionen gegen $p > 0$ konvergieren (Strobl, 2010, S. 50ff.).

Items. Hingegen löst eine Person mit einer hohen Fähigkeit *wahrscheinlich* die einfachen und die schwierigen Items und insgesamt mehr Items.

Lokale stochastische Unabhängigkeit (Eid & Schmidt, 2014, S. 154ff.; Moosbrugger, 2012a, S. 229ff.; Strobl, 2010, S. 16ff.). Stochastische Unabhängigkeit von zwei Ereignissen liegt vor, wenn das *Ergebnis* des ersten Ereignisses die *Wahrscheinlichkeit* des Ergebnisses des zweiten Ereignisses *nicht verändert*. Sind die Ereignisse voneinander stochastisch unabhängig, so kann ihre gemeinsame Wahrscheinlichkeit als Produkt der einzelnen Wahrscheinlichkeiten berechnet werden (sog. Multiplikationstheorem für unabhängige Ereignisse). Im Rasch-Modell ergibt sich die gemeinsame Lösungswahrscheinlichkeit aller Personen $v = 1, \dots, n$ für alle Items $i = 1, \dots, m$ als:

$$\prod_{v=1}^n \prod_{i=1}^m P(X_{vi} = x_{vi} | \theta_v, \sigma_i).$$

Im Rasch-Modell muss *lokale* stochastische Unabhängigkeit für die *Items* gelten, d.h. die Lösungswahrscheinlichkeit *einer* Person – oder mehrerer Personen mit *gleicher* Personenfähigkeit – für ein Item darf *nicht* davon abhängen, ob *diese eine* Person ein anderes Item lösen kann. Es darf aber durchaus sein, dass eine Person mit höherer Fähigkeit alle Aufgaben mit höherer Wahrscheinlichkeit löst als eine Person mit geringerer Fähigkeit. Bei der Testentwicklung muss also darauf geachtet werden, dass die Items nicht aufeinander aufbauen, d.h. dass die Lösung des zweiten Items nicht aus der Lösung des ersten Items folgt.⁴⁷ Lokale stochastische Abhängigkeit muss im Rasch-Modell auch für die *Personen* gelten, d.h. die Lösungswahrscheinlichkeit *einer* Person für ein bestimmtes Item darf nicht davon abhängen, ob eine *andere* Person dieses Item lösen kann. Bei der Testdurchführung muss also darauf geachtet werden, dass die Testpersonen einander nicht einflüstern oder voneinander abschreiben. Lokale stochastische Unabhängigkeit wird im Rahmen der Parameterschätzung geprüft (vgl. Kapitel 5.2.3.c).

Spezifische Objektivität (Eid & Schmidt, 2014, S. 153f.; Rost, 2004, S. 121f.; Strobl, 2010, S. 20ff.).⁴⁸ Spezifische Objektivität bedeutet, dass der Vergleich zweier Messobjekte (Testpersonen) *nicht* vom Messinstrument (Testitems) abhängt, und *vice versa*, der Vergleich zweier

⁴⁷ In diesem Zusammenhang wird auch auf die Effekte der Item-Reihenfolge hingewiesen (vgl. Döring & Bortz, 2016, S. 490; Eid & Schmidt, 2014, S. 88ff.).

⁴⁸ Die spezifische Objektivität wird gelegentlich auch als *Stichprobenunabhängigkeit* bezeichnet, was jedoch irreführend, wenn nicht gar falsch ist. Die Parameter des Rasch-Modells sind nur dann stichprobenunabhängig, wenn das Rasch-Modell in der gezogenen Stichprobe *und* in der adressierten Grundgesamtheit gilt. Erstens ist es für die Gültigkeit des Rasch-Modells entscheidend, welche Personen und Items man auswählt/untersucht (z.B. ist ein modellgültiger Test für Gymnasiast*innen nicht automatisch auch modellgültig für Berufsschüler*innen). Zweitens ist auch bei Gültigkeit des Rasch-Modells die *Genauigkeit* der Parameterschätzung von der Verteilung der Personen- und Itemparameter in der Stichprobe abhängig (Rost, 2004, S. 121).

Messinstrumente *nicht* vom Messobjekt. Mit anderen Worten: Eine Aussage über die Fähigkeit zweier Personen ist unabhängig davon, anhand welcher Items sie verglichen werden; eine Aussage über die Schwierigkeit zweier Items ist unabhängig davon, anhand welcher Personen sie verglichen werden. Die Differenz der Lösungswahrscheinlichkeiten bzw. der Logit-Variablen von zwei Personen v (θ_1 und θ_2) für *alle* Items i ist:

$$\log \frac{P(X_v=1|\theta=\theta_2)}{P(X_v=0|\theta=\theta_2)} - \log \frac{P(X_v=1|\theta=\theta_1)}{P(X_v=0|\theta=\theta_1)} = \theta_2 - \theta_1.$$

Die Differenz der Lösungswahrscheinlichkeiten bzw. der Logit-Variablen von *allen* Personen v für zwei Items i (σ_1 und σ_2) ist:

$$\log \frac{P(X_2=1|\theta=\theta_v)}{P(X_2=0|\theta=\theta_v)} - \log \frac{P(X_1=1|\theta=\theta_v)}{P(X_1=0|\theta=\theta_v)} = \sigma_2 - \sigma_1.$$

Die spezifische Objektivität ist nur durch parallele IC-Kurven gewährleistet, d.h. sie gilt nur für 1PL-Modelle. Stellt sich ein Item für unterschiedliche Personengruppen (z.B. Geschlecht, Migrationshintergrund, Schulform/Ausbildung) als unterschiedlich schwierig heraus, spricht man von *differential item functioning* (DIF).⁴⁹ DIF kann durch sog. Modelltests geprüft werden (vgl. Kapitel 5.2.3.d).

Eindimensionalität (Moosbrugger, 2012a, S. 229f.; Strobl, 2010, S. 23). Eindimensionalität im weiteren Sinne bedeutet, dass die Items eines Tests inhaltlich ähnlich sind und dass sie ein und dasselbe latente Personenmerkmal messen. Falls sie zusätzlich zum Personenmerkmal noch ein anderes Merkmal messen, dann dürfen sie es nicht systematisch tun.⁵⁰ Eindimensionalität des Tests im engeren Sinne bedeutet, dass die Personen- und Itemparameter auf *einer* gemeinsamen latenten Dimension liegen.

Skalenniveau (Eid & Schmidt, 2014, S. 148ff.; Rost, 2004, S. 121; Strobl, 2010, S. 23ff.). Das Skalenniveau der Personen- und Itemparameter ist ausgehend von der Logit-Variablen metrisch. Je nach Darstellung der Modellgleichung wird in der Literatur für eine Intervallskala (willkürlicher Skalen-Nullpunkt und willkürliche Skaleneinheiten, d.h. Dehnung/Stauchung der Skala) oder eine Differenzenskala argumentiert (willkürlicher Skalen-Nullpunkt und fixe Skaleneinheiten, d.h. nur

⁴⁹ Beispiel: Wenn in einem Test zur Mathematik-Kompetenz ein Item (Textaufgabe) ein selten verwendetes Wort enthält, so ist dieses Item für Schüler*innen mit Deutsch als Fremdsprache schwieriger zu lösen als für Schüler*innen mit Deutsch als Erstsprache, auch wenn ihre Mathematikfähigkeit gleich ist.

⁵⁰ Beispiel: Ein Test zur Mathematik-Kompetenz soll nur Mathematik-Kompetenz messen, nicht auch noch Sprach-Kompetenz.

additive Transformation der Skala). Bei einer Differenzenskala bleiben die *Differenzen* der Parameter durch zulässige, d.h. additive Transformationen unverändert. Aussagen über die Differenz von Personenparametern und über die Differenz von Itemparametern sind also – unabhängig vom jeweiligen Parameterwert – *bedeutsam* (vgl. spezifische Objektivität).

Normierung (Eid & Schmidt, 2014, S. 148ff.; Moosbrugger, 2012a, S. 243, S. 254; Rost, 2004, S. 121). Normierung ist das Festlegen von *a priori* nicht eindeutigen Parametern. Im Rasch-Modell hängt die Lösungswahrscheinlichkeit einer Person v für ein Item i einzig von der Differenz der beiden Parameter ab: $\theta_v - \sigma_i$. Es gibt jedoch unendlich viele Kombinationen der Parameter, die zum selben Differenzwert führen. Die Personenparameter θ_v und Itemparameter σ_i sind also eindeutig bestimmt bis auf Translationen (additive Skalen-Transformation). Für eine eindeutige Bestimmung der Parameter muss eine Normierung vorgenommen werden. In der Praxis wird konventionell die Summennormierung der Itemparameter gewählt, d.h. die Summe aller Schwierigkeitsparameter über alle Items hinweg wird auf 0 fixiert: $\sum_{i=1}^m \sigma_i = 0$. Mit dieser Summennormierung der Itemparameter sind nun auch die Personenparameter bestimmt. Negative Personenparameter korrespondieren mit geringen Fähigkeitsausprägungen, positive Personenparameter mit hohen Fähigkeitsausprägungen. Die Parameter nehmen auf der gemeinsamen Logit-Skala i.d.R. Werte im Intervall zwischen -3 und $+3$ an.

5.2.3.c Parameterschätzung

Die Schätzung der unbekannten Item- und Personenparameter aus der Datenmatrix des psychologischen Tests erfolgt nach einem iterativen Algorithmus. Für die Parameterschätzung gibt es verschiedene Ansätze und Methoden; bei Rasch-Modellen werden die Parameter am häufigsten nach dem *Maximum-Likelihood-Ansatz* (ML) geschätzt (Rost, 2004, S. 309ff.; Strobl, 2010, S. 27f., S. 34f.).⁵¹ Der ML-Ansatz verwendet eine Likelihood-Funktion (englisch *likelihood*: Plausibilität), die definiert ist als das Produkt der Wahrscheinlichkeiten über alle Personen und alle Items (Rost, 2004, S. 123ff.; Strobl, 2010, S. 27f.).⁵²

$$L_{x_{vi}}(\theta_v, \sigma_i) = P(x_{vi} | \theta_v, \sigma_i) = \frac{e^{v_i \cdot (\theta_v - \sigma_i)}}{1 + e^{\theta_v - \sigma_i}} = \prod_{v=1}^n \prod_{i=1}^m \frac{e^{x_{vi} \cdot (\theta_v - \sigma_i)}}{1 + e^{\theta_v - \sigma_i}}, \quad v = 1, \dots, n \text{ und } i = 1, \dots, m$$

⁵¹ Neben den ML-Ansätzen für die Parameterschätzung gibt es noch andere, wie z.B. die paarweise Schätzmethode (*pair wise*) für die Itemparameter oder der Bayes-Ansatz für die Personenparameter.

⁵² Diese Darstellung der Likelihood-Funktion entspricht einer *joint maximum likelihood* (JML), bei der Item- und Personenparameter gleichzeitig geschätzt werden.

Die Likelihood-Funktion beschreibt also die Wahrscheinlichkeit der Daten in Abhängigkeit der Modellparameter und unter der Annahme der Modellgültigkeit: beim Maximalwert der Likelihood-Funktion finden sich die besten Parameterschätzungen; je höher der Wert der Likelihood-Funktion, desto besser passt das Modell auf die Daten. Die Likelihood kann Werte zwischen 0 und 1 annehmen. Die Multiplikation der Wahrscheinlichkeiten setzt stochastische Unabhängigkeit bezüglich Personen und Items voraus (Moosbrugger, 2012a, S. 239f.; Rost, 2004, S. 112f.) (vgl. Kapitel 5.2.3.b, lokale stochastische Unabhängigkeit).

Innerhalb des ML-Ansatzes können drei Methoden angewendet werden; sie unterscheiden sich dahingehend, ob die Item- und Personenparameter gleichzeitig oder nacheinander geschätzt werden und wie mit den Personenparametern umgegangen wird. Nachfolgend werden die Methoden zusammenfassend beschrieben. Für eine formale Darstellung drei Likelihood-Funktionen sei auf die einschlägige Literatur verwiesen (Eid & Schmidt, 2014, S. 156ff.; Rost, 2004, S. 123ff., S. 309ff.; Strobl, 2010, S. 27ff.).

- Bei der unbedingten ML-Schätzung werden die Itemparameter und die Personenparameter gemeinsam in der *joint maximum likelihood* (JML-Funktion) anhand der Datenmatrix geschätzt. Allerdings liefert die JML-Funktion für die *Itemparameter* keine konsistenten Schätzer⁵³, d.h. die Schätzgenauigkeit der Itemparameter ist verzerrt. Die JML-Methode liefert für die *Personenparameter* zwei Punktschätzer, die *maximum likelihood estimates* (MLE) und die *weighted likelihood estimates* (WLE) (vgl. Warm, 1989). Die WLE haben gegenüber den MLE geringere absolute Schätzwerte und eine geringe Varianz der Schätzwerte, d.h. die WLE sind die besten *Punktschätzer* für die Personenparameter.
- Bei der bedingten ML-Schätzung werden die *Itemparameter* in der *conditional maximum likelihood* (CML-Funktion) geschätzt. Die Personenparameter werden durch die Personen-Randsummen „ersetzt“ (vgl. Kapitel 5.2.3.b, suffiziente Statistik). Mit der CML-Funktion können die Itemparameter konsistent geschätzt werden. Allerdings lassen sich mit der CML-Funktion für Personen, die keine oder alle Items gelöst haben, keine Personenfähigkeit geschätzt werden. Umgekehrt können für Items, die von keiner oder von allen Personen gelöst wurden, keine Item-Schwierigkeit geschätzt werden.
- Bei der marginalen ML-Schätzung werden die *Itemparameter* in der *marginal maximum likelihood* (MML-Funktion) geschätzt. Die Personenparameter werden durch eine Personen-

⁵³ Ein konsistenter Schätzwert liegt vor, wenn sich die Statistik mit wachsendem Stichprobenumfang dem zu schätzenden Parameter annähert (Bortz & Schuster, 2010, S. 88f.). Bei einem konsistenten Schätzer wird durch eine Vergrößerung der Stichprobe der Mittelwert der Schätzfunktion stabiler, d.h. weniger von einzelnen Ausreißern abhängig, und die Varianz der Schätzfunktion geringer. Die JML-Funktion liefert nur dann konsistente Schätzer, wenn die Anzahl Personen und die Anzahl Items sehr gross sind, d.h. gegen unendlich konvergieren. Für einen psychologischen Test lässt sich die Anzahl Personen relativ „leicht“ vergrössern, jedoch ist die Anzahl Items typischerweise durch die Testlänge fixiert.

Randverteilung „ersetzt“, wobei dafür meist eine Normalverteilung angenommen wird. Mit der MML-Funktion muss die Verteilung der Personenparameter *a priori* festgelegt werden, so dass eine falsch gewählte Verteilung die Schätzung der Personenparameter verzerren kann. Die MML-Methode liefert keine Punktschätzer für die Personenparameter, jedoch die bayesianisch geschätzten Verteilungsparameter (Erwartungswert und Varianz) der latenten Variablen, sog. *expected a posteriori* Schätzer (EAP).

Die drei Methoden liefern für die Itemparameter dieselben Schätzwerte. Hingegen zeigen sie für die Personenparameter eine unterschiedliche *Varianz* der Schätzwerte (Rost, 2004, S. 309): Der WLE hat eine höhere Erwartungstreue (d.h. eine geringste systematische Verzerrung, v.a. in den Extrembereichen), der EAP hat einen geringeren bedingten Standardfehler (Eid & Schmidt, 2014, S. 180f.; Frey, 2012, S. 284f.).

Mit der unbedingten ML-Methode lassen sich – indirekt über die Testinformations-Funktion – die Standardfehler der geschätzten Personenparameter bestimmen. Die Standardfehler für die geschätzten Personenparameter widerspiegeln die Genauigkeit der Schätzungen. Die Testinformationsfunktion (*test information function*) gibt an, wie viel Information in den Daten zur Schätzung der Personenparameter enthalten ist. Ein Wert der Testinformationsfunktion kann bestimmt werden, indem die geschätzten Itemparameter in die Modellgleichung des Rasch-Modells eingesetzt und hierüber die bedingten Varianzen geschätzt werden. Ein Wert der Testinformationsfunktion ist die Summe der bedingten Item-Varianzen $Var(X_i|\theta = \theta_v)$ an der Stelle $\theta = \theta_v$ (Eid & Schmidt, 2014, S. 173ff.):

$$I(\theta) = \sum_{i=1}^m P(X_i = 1|\theta) \cdot P(X_i = 0|\theta) = \sum_{i=1}^m Var(X_i|\theta)$$

Die bedingte Item-Varianz kann auch als Item-Informationsfunktion ausgedrückt werden: $I_i(\theta) = Var(Y_i|\theta)$. Die Item-Informationsfunktion $I_i(\theta)$ ist die erste Ableitung der itemcharakteristischen Funktion (IC-Funktion) und sieht einer Normalverteilung ähnlich. Sie hat einen maximalen Wert von 0.25, und zwar an der Stelle $\theta = \sigma_i$, wo die IC-Funktion ihren Wendepunkt hat, also bei der Lösungswahrscheinlichkeit $p = 0.5$. An dieser Stelle liefert das Item am meisten Information, da die beide Antwortkategorien 0 und 1 gleich wahrscheinlich sind und die Trennschärfe des Items am höchsten ist (Eid & Schmidt, 2014, S. 174f.; Strobl, 2010, S. 36).

Der Standardfehler für einen geschätzten Personenparameters $\hat{\theta}_v$ lässt sich über den geschätzten Wert der Test-Informationsfunktion $\hat{I}(\theta = \theta_v)$ an der Stelle $\theta = \theta_v$ bestimmen (Eid & Schmidt, 2014, S. 175f.):

$$SE_{\hat{\theta}_i} = \frac{1}{\sqrt{\hat{I}(\theta = \hat{\theta}_i)}}$$

Der Standardfehler ist umso kleiner, je grösser der Wert der Test-Informationsfunktion bzw. der bedingten Varianz an der Stelle des geschätzten Personenparameters ist. Der Standardfehler für einen geschätzten Personenparameter nimmt einen theoretischen Minimalwert an, wenn alle Items denselben Item-Schwierigkeitsparameter aufweisen *und* wenn der geschätzte Personenparameter genau dem Itemparameter entspricht (Eid & Schmidt, 2014, S. 175f.). Der Standardfehler bzw. die Genauigkeit der geschätzten Personenparameter hängt von zwei Faktoren ab: Erstens von der Anzahl Items; je mehr Items ein Test umfasst, desto präziser lassen sich die Personenparameter schätzen. Zweitens von der Passung der Itemparameter und Personenparameter; je mehr Items des Tests eine Item-Schwierigkeit nahe der zu schätzenden Personenfähigkeit aufweisen, desto präziser lassen sich die Personenparameter schätzen. Insofern ist ein sehr einfaches Item für die präzise Schätzung einer sehr fähigen Person nicht geeignet, und umgekehrt, ein sehr schwieriges Item nicht für eine sehr unfähige Person. Diese Eigenschaften der Standardfehler können für die Testentwicklung und Testevaluation genutzt werden, insbesondere auch um Aspekte der Schätzgenauigkeit und Testökonomie zu verbinden (Eid & Schmidt, 2014, S. 175f.).

Je nach Testzweck ist also eine unterschiedliche Zusammenstellung der Items bzw. eine unterschiedliche Streuung der Item-Schwierigkeitsparameter entlang der Personenparameter anzustreben.⁵⁴ Wenn mit dem Test z.B. die ganze Bandbreite der Personenfähigkeit erfasst werden soll, müssen die Item-Schwierigkeiten über die gesamte Logit-Skala streuen (Eid & Schmidt, 2014, S. 176). Zudem muss der Test mehr Items mit mittlerer Schwierigkeit sowie weniger Items mit geringer und hoher Schwierigkeit umfassen. In diesem Falle liefert der Test am meisten Information im mittleren Logit-Bereich und weniger Information in den Randbereichen. Die Standardfehler der Personenparameter sind in den Randbereichen grösser, d.h. die Schätzung der Personenparameter also ungenauer (Strobl, 2010, S. 37, S. 75).

⁵⁴ Wenn der Test für hochbegabte Personen konzipiert ist, muss er v.a. aus schwierigen Items bestehen, um für die für hohe Personenfähigkeiten eine präzise Schätzung zu ermöglichen, und *vice versa*.

5.2.3.d Modellgeltungstests

Modellgeltungstests sind Methoden zur (statistischen) Überprüfung der Gültigkeit des Rasch-Modells, d.h. sie überprüfen, ob die empirischen Daten die Modellannahmen erfüllen. Die mit der Likelihood-Funktion geschätzten Parameter sagen nämlich noch nichts darüber aus, ob die Modellannahmen erfüllt sind. Je nach dem interessierenden Aspekt der Modellgültigkeit werden in der Literatur verschiedene Modelltests vorgeschlagen (vgl. Eid & Schmidt, 2014, S. 182ff.; Rost, 2004, S. 330ff.; Strobl, 2010, S. 39ff.):

- Gleichheit der Itemparameter in Subpopulationen (sog. DIF-Analyse): grafischer Modelltest, Andersen-Test (auch: bedingter Likelihood-Quotienten-Test), Wald-Test, Mixed-Rasch-Test;
- globale Modellgültigkeit und Wahrscheinlichkeitsverteilung der Antwortmuster: Pearson- χ^2 -Test, Likelihood-Ratio-Test;
- Gleichheit der Personenparameter in reduzierten Rasch-Modellen: Martin-Löf-Test.

Nachfolgend wird der Aspekt „Gleichheit der Itemparameter“ bzw. *differential item functioning* (DIF) und einige dafür konzipierte Modelltests zusammenfassend vorgestellt. Eine ausführliche formale Darstellung der verschiedenen Modelltests findet sich in der einschlägigen Literatur (vgl. Eid & Schmidt, 2014, S. 182ff.; Rost, 2004, S. 330ff.; Strobl, 2010, S. 40ff.).

Aus den Annahmen des Rasch-Modells folgt, dass die Itemparameter für alle Subpopulationen identisch sein müssen (sog. Item-Homogenität). Sind die Subpopulationen bekannt, kann die Stichprobe nach dem Personenmerkmal (Personen-Testwerte) oder nach einem anderen relevanten Kriterium (z.B. Geschlecht, Schulform) in zwei oder mehrere Substichproben unterteilt werden. Für jede Substichprobe werden getrennte Schätzungen der Itemparameter vorgenommen. DIF liegt vor, wenn sich die geschätzten Itemparameter zwischen den Personengruppen der Stichprobe bedeutsam und signifikant unterscheiden, und zwar bei Personen mit *gleichen* Personenparametern. Mit anderen Worten, bei DIF sind die Items des Tests für die Gruppen *systematisch* einfacher bzw. schwieriger.

«If the responses to an item have different frequencies for different subgroups, that is evidence of differential impact of the item on those subgroups [...] Instead DIF studies focus on whether respondents at the same locations give similar responses across the different subgroups» (Wilson, 2005, S. 165).

Wenn DIF vorliegt, ist die Annahme der Item-Homogenität verletzt und die Gültigkeit des Rasch-Modells nicht erwiesen (Eid & Schmidt, 2014, S. 183ff.; Moosbrugger, 2012a, S. 244; Strobl, 2010, S. 39).

Differential item functioning (DIF) kann grafisch oder numerisch, d.h. mittels statistischer Tests, überprüft werden. Bei den statistischen Tests wird als Nullhypothese Modellkonformität und als Alternativhypothese Nicht-Modellkonformität formuliert. Die Nullhypothese ist also hier die

„Wunschhypothese“ (Döring & Bortz, 2016, S. 491, S. 885): Die geschätzten Itemparameter sind für beide Gruppen gleich bzw. unterscheiden sich nicht systematisch. Eine *signifikante* Statistik bei den Modelltests bedeutet, dass das Rasch-Modell *nicht* gilt (Strobl, 2010, S. 43).

Bei der *grafischen Methode* werden die Itemparameter von zwei Gruppen (z.B. Personenparameter $\theta < 0$ Logits und Personenparameter $\theta > 0$ Logits, Mädchen und Jungen) in getrennten Modellen geschätzt und die geschätzten Itemparameter als Koordinatenpunkte in einem Diagramm abgetragen (Eid & Schmidt, 2014, S. 183ff.; Strobl, 2010, S. 39ff.). Jeder Punkt im Koordinatensystem repräsentiert ein Item, das aufgrund seiner Position als „DIF zugunsten der Fokusgruppe“ (DIF *against reference group*) oder als „DIF zugunsten der Referenzgruppe“ (DIF *against focal group*) interpretiert werden kann. Items auf der 45°-Diagonale haben keine Abweichung in den Itemparametern ($|\Delta \sigma| = 0$ Logits): Die geschätzten Itemparameter sind für beide Gruppen identisch, die Items sind für die Gruppen gleich schwierig – es liegt in diesem Falle kein grafisches DIF vor. Items oberhalb und unterhalb der Diagonalen haben Abweichungen in den Itemparametern ($|\Delta \sigma| > 0$ Logits). Eine geringe Abweichung von der 45°-Diagonalen ($|\Delta \sigma| < 0.5$ Logits) ist aufgrund von zufälligen Schwankungen im Schätzwert tolerierbar. Eine starke Abweichung von der 45°-Diagonalen, $|\Delta \sigma| > 0.5$ Logits) weist darauf hin, dass sich die geschätzten Itemparameter in den Gruppen systematisch unterscheiden – es liegt grafisches DIF vor (Draba, 1977; Strobl, 2010, S. 40).

Bei der *numerischen Methode* werden die Itemparameter der Gruppen in einem gemeinsamen Modell geschätzt und die Differenzen der Itemparameter verglichen. Numerische Modelltests können auf Skalenebene, d.h. Prüfung des DIF für den gesamten Test, oder auf Itemebene, d.h. Prüfung des DIF für einzelne Items, durchgeführt werden (Strobl, 2010, S. 41ff.). Als Modelltest auf Skalenebene bieten sich der Likelihood-Quotienten-Test (Andersen, 1973), der globale Wald-Test oder der X^2 -Differenz-Test an. Als Modelltest auf Itemebene kann der Item-spezifische Wald-Test, der DIF-Rasch-Test oder der Mantel-Haenszel-Test (Paek & Wilson, 2011; Sun, 2012; Zwick, Thayer, & Lewis, 1999) verwendet werden.⁵⁵

Beim X^2 -Differenz-Test wird ein restriktives Modell mit gleichen Itemparametern und ein weniger restriktives Modell mit unterschiedlichen Itemparametern spezifiziert. Die Differenz der relativen Modellgüte-Indizes (Δ Deviance) und die Differenz der Freiheitsgrade (Δ df) wird an der X^2 -Statistik auf statistische Signifikanz geprüft. Die statistische Nullhypothese (H_0) postuliert gleiche

⁵⁵ Der X^2 -Differenz-Test, der DIF-Rasch-Test und der Mantel-Haenszel-Test sind in ACER ConQuest Version 4 verfügbar (Adams & Wu, 2010b; Adams, Wu, & Wilson, 2015; Paek & Wilson, 2011; Sun, 2012; Wu, Adams, Wilson, & Haldane, 2007, S. 89ff.). Der Likelihood-Quotienten-Test und die Wald-Tests können im Statistikprogramm R als Pakete installiert werden (Mair & Hatzinger, 2007, 2009; R-Team, 2010; Strobl, 2010).

Itemparameter für die Gruppen bzw. keine signifikante Modellgüte-Differenz ($p > .05$). H_0 muss abgelehnt werden, wenn die Modellgüte-Differenz signifikant ist (vgl. Eid & Schmidt, 2014, S. 185ff.).

Beim DIF-Rasch-Test werden die Differenzen der gruppenspezifischen Itemparameter geschätzt und an der t-Statistik auf statistische Signifikanz geprüft. Auch hier lautet die statistische Nullhypothese (H_0), dass der jeweilige Itemparameter für die Gruppen gleich ist bzw. keine signifikante Itemparameter-Differenz (DIF) ($|t| < 1.96$ bzw. $p > .05$) vorliegt. H_0 muss abgelehnt werden, wenn die Itemparameter-Differenz signifikant ist (vgl. Eid & Schmidt, 2014, S. 187ff.). Wird für den DIF-Rasch-Test eine Fokusgruppe und Referenzgruppe definiert, so bedeutet ein $DIF > 0$ DIF zugunsten der Fokusgruppe (*against reference group*), ein $DIF < 0$ DIF zulasten der Fokusgruppe (*against focal group*). Der Educational Testing Service (ETS) verwendet für die Beurteilung der Effektstärke des DIF folgende Kategorien und Toleranzwerte (Paek & Wilson, 2011; Zwick et al., 1999):

- Kategorie A "negligible effect": nicht signifikante Itemparameter-Differenz ($p > .05$ bzw. $|z| \leq 1.96$).
- Kategorie A "negligible effect": signifikante Itemparameter-Differenz ($p < .05$ bzw. $|z| > 1.96$) von $|DIF| \leq 0.426$ Logits; A+ für Fokusgruppe, A– für Referenzgruppe.
- Kategorie B "medium effect": signifikante Itemparameter-Differenz von 0.426 Logits $< |DIF| < 0.638$ Logits; B+ für Fokusgruppe, B– für Referenzgruppe.
- Kategorie C "large effect": signifikante Itemparameter-Differenz von $|DIF| \geq 0.638$ Logits; C+ für Fokusgruppe, C– für Referenzgruppe.

5.2.3.e Itemfit-Masse

Itemfit-Masse geben Hinweise auf die Qualität eines Items, d.h. wie gut ein Item das latente Merkmal abbildet und wie gut es zum verwendeten Testmodell passt. Itemfit-Masse können – ergänzend zur DIF-Analyse – beigezogen werden, um nicht-modellkonforme Items zu identifizieren und ggf. zu selektieren (Eid & Schmidt, 2014, S. 198f.). Die inhaltliche Analyse der modellkonformen und nicht-modellkonformen Items liefert zudem Hinweise für die Konstruktion neuer Items und/oder die Modifikation bestehender Items. Jedoch müssen auch diese neukonstruierten und modifizierten Items einer erneuten Modellgültigkeitsprüfung unterzogen werden (Döring & Bortz, 2016, S. 491).

In der Literatur werden residuenbasierte und likelihood-basierte Itemfit-Masse unterschieden (Eid & Schmidt, 2014, S. 198ff.; Rost, 2004, S. 371ff.). *Likelihood-basierte* Masse beziehen sich auf die Wahrscheinlichkeit der beobachteten Item-Antworten. *Residuenbasierte* Itemfit-Masse beziehen sich auf die Abweichung von empirisch beobachteten und theoretisch erwarteten Item-Antworten. Je geringer die Abweichung zwischen beobachteten und erwarteten Lösungshäufigkeiten ist, desto besser ist der Item-Infit. Oder anders ausgedrückt: Je mehr Personen eine erwartete Antwort auf das Item geben, desto besser ist der Item-Infit. Je nach Itemfit-Mass werden die erwarteten Item-

Antworten unterschiedlich berechnet, die Differenzen unterschiedlich gewichtet und eine andere Prüfstatistik abgeleitet. Ein häufig verwendetes Residualmass sind die standardisierten quadrierten Abweichungen, die über alle Personen v summiert werden: $Infit_i = \sum_{v=1}^n z_{vi}^2$.

Für das Rasch-Modell können die ungewichteten (*unweighted mean squares*, MNSQ) und die gewichteten Abweichungsquadrate (*weighted mean squares*, wMNSQ) zwischen beobachteten und erwarteten Lösungshäufigkeiten als residuenbasierte Item-Infit-Statistik verwendet werden.⁵⁶ Der wMNSQ wird bevorzugt verwendet, da er weniger sensitiv für Ausreisser ist (Winther, 2010, S. 143ff.). Anhand der ICC lässt sich der Item-Infit auch grafisch ablesen: die theoretische ICC repräsentiert die erwarteten Lösungshäufigkeiten, die empirische ICC die beobachteten Lösungshäufigkeiten (Wu et al., 2007, S. 28f.). Es können drei Fälle unterschieden werden (OECD, 2014b, S. 148; Winther, 2010, S. 144f.; Wu et al., 2007, S. 40):

- wMNSQ = 1 (*perfect fit*): Der wMNSQ ist gleich seinem Erwartungswert von 1. Das Item ist modellkonform. Die beiden ICC sind im Logit-Bereich des *Item-Schwierigkeitsparameters* gleich steil.
- wMNSQ < 1 (*overfit*): Der wMNSQ ist *kleiner* als sein Erwartungswert. Die empirische ICC ist im Logit-Bereich des *Item-Schwierigkeitsparameters* *steiler* als die theoretische. Die Item-Trennschärfe ist *höher* als vom Modell erwartet.
- wMNSQ > 1 (*underfit*): Der wMNSQ ist *grösser* als sein Erwartungswert. Die empirische ICC ist im Logit-Bereich des *Item-Schwierigkeitsparameters* *flacher* als die theoretische, d.h. die Item-Trennschärfe ist *geringer* als vom Modell erwartet.

Für den Item-Infit werden in der Literatur Toleranzintervalle vorgeschlagen und verwendet: Das Intervall [0.80, 1.20] gilt als konservativ (OECD, 2014b, S. 151)⁵⁷, das Intervall [0.75, 1.33] als weniger konservativ (Adams & Khoo, 1996, zit. nach Wilson, 2005, S. 129). Die statistische Nullhypothese ist wMNSQ = 1 und wird an einer t-Statistik auf Signifikanz geprüft ($|t| < 1.96$ bzw. $p < .05$), wobei ein nicht signifikanter Itemfit ideal, aber nicht zwingend ist (Wu et al., 2007, S. 25). Ist der wMNSQ statistisch *signifikant* von 1 verschieden, wird die Nullhypothese abgelehnt – das Item ist *nicht* modellkonform.

⁵⁶ Im Programm ACER ConQuest werden für den Item-Infit die *unweighted mean squares* (MNSQ) und die *weighted mean squares* (wMNSQ) jeweils mit t-Statistik und 95 %-Konfidenzintervall ausgegeben (Wu et al., 2007, S. 25).

⁵⁷ In PISA-Studie 2012 wurde ein wMNSQ zwischen 0.80 und 1.20 als tolerierbar erachtet (OECD, 2014b, S. 151).

5.2.4 Das Partial-Credit-Rasch-Modell (ordinales 1PL-Latent-Trait-Modell)

Das ordinale Rasch-Modell bzw. das auf Leistungstests passende Partial-Credit-Rasch-Modell (Masters, 1982) ist ein eindimensionales Modell für *polytome* Antwortvariablen mit *geordneten* Antwortkategorien, bei dem neben dem Personenparameter nur *ein* Itemparameter (1PL-Modell) geschätzt und das latente Merkmal als eine kardinalskalierte Variable (*latent trait*) erfasst wird (Döring & Bortz, 2016, S. 486ff.). Im Partial-Credit-Rasch-Modell können Items mit unterschiedlicher Anzahl Antwortkategorien k_i verwendet werden, z.B. $k_i = 2$ Kategorien (d.h. dichotom 0/1) und $k_i = 3$ Kategorien (d.h. polytom geordnet 0/1/2). Die Antwortkategorien werden für jedes Item aufsteigend kodiert: Der ersten Kategorie wird der Wert $c = 0$ zugeordnet, der zweiten der Wert $c = 1$, der k -ten Kategorie der Wert $c = k-1$ (Eid & Schmidt, 2014, S. 230). Insofern stellt das Partial-Credit-Rasch-Modell eine Erweiterung des Rasch-Modells dar, oder umgekehrt, das Rasch-Modell ist ein Spezialfall des Partial-Credit-Rasch-Modells. Nachfolgend wird die Gleichung des Partial-Credit-Rasch-Modells zusammenfassend dargestellt. Für eine ausführliche formale Darstellung des Rasch-Modells sei auf die einschlägige Literatur verwiesen (Rost, 2004, S. 203ff.).

Das Partial-Credit-Rasch-Modell bzw. die Modellgleichung des Partial-Credit-Rasch-Modells spezifiziert die Lösungswahrscheinlichkeit p einer Person v für *jede Kategorienantwort* c eines Items i über eine logistische Funktion. Die Lösungswahrscheinlichkeit der Kategorienantwort c ist abhängig von der Personenfähigkeit (θ_v) und der Kategorien-Schwierigkeit (σ_{ic}) (Rost, 2004, S. 203ff.; Strobl, 2010, S. 54ff.):

$$P(X_{vi} = c | \theta_v, \sigma_i) = \frac{e^{c \cdot \theta_v - \sigma_{ic}}}{\sum_{s=0}^{k-1} e^{s \cdot \theta_v - \sigma_{is}}}, \quad c = 0, \dots, k-1, \text{ wobei } \sigma_{ic} = \sum_{s=0}^c \tau_{is} \text{ und } \sigma_{i0} = 0$$

Der Personenparameter θ_v wird mit dem Kategorienwert c gewichtet und mit allen darunter liegenden Kategorien-Schwierigkeiten σ_{ic} verglichen. Die Kategorien-Schwierigkeit σ_{ic} ergibt sich aus der Summe der Schwellenparameter τ_{is} , die kleiner oder gleich c sind: die Summe aller Schwellen, die eine Person überschritten hat, wenn ihre Antwort mit dem Kategorienwert c kodiert wurde. Der Kategorienparameter σ_{ic} ist also ein *kumulierter* Parameter. Der Schwellenparameter τ_{is} bezeichnet die Schwierigkeit, die Schwelle von der unteren zur oberen Kategorie zu überschreiten: $p(x_{vi}|c) = p(x_{vi}|c-1)$; er ist ein *dekumulierter* Parameter. Er ist definiert also jene Ausprägung des Personenparameters θ_v , bei dem die *Schwellenwahrscheinlichkeit* (nicht die Kategorienwahrscheinlichkeit) genau 50 % beträgt. Der Mittelwert aller Schwellenparameter τ_{is} wird als *Lokalisationsparameter* (auch: Lageparameter) des Items i bezeichnet (Eid & Schmidt, 2014, S. 234ff.; Rost, 2004, 205ff.) und kann mit dem Schwierigkeitsparameter eines Items mit dichotomen Antwortkategorien verglichen werden.

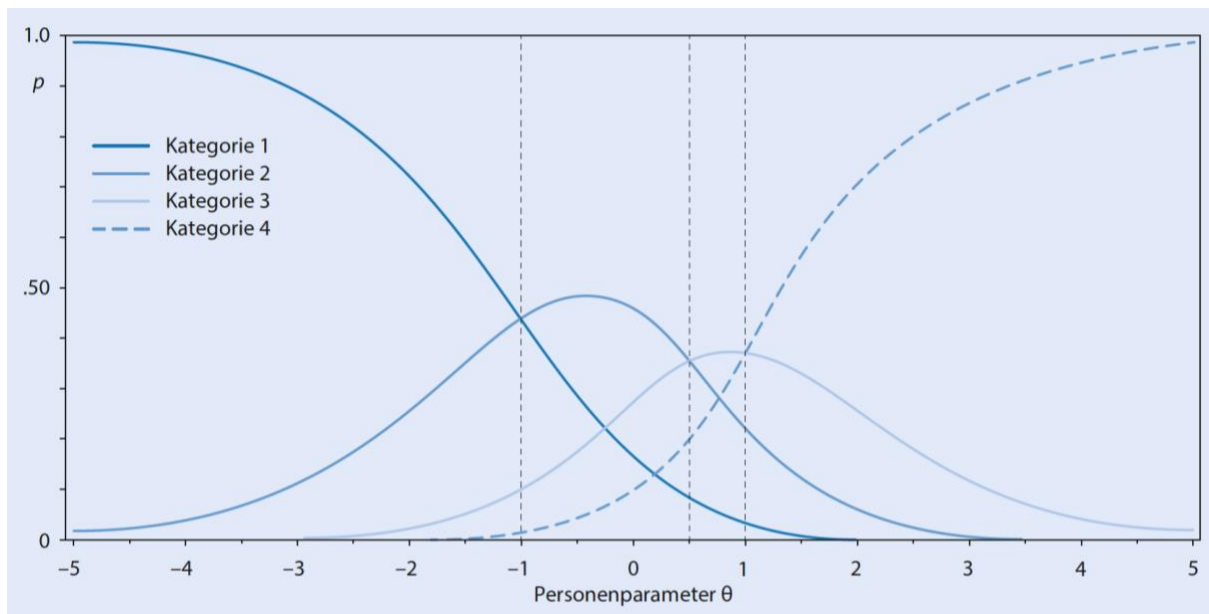


Abbildung 5-6. Category Characteristic Curves für vier Kategorien eines ordinalen Items

Hinweis: Die X-Achse ist eine Logit-Skala, auf der die Personenfähigkeiten (θ) und die Item-Schwierigkeiten (σ) abgetragen werden. Auf der Y-Achse wird die Lösungswahrscheinlichkeit (p) abgetragen. Die drei Schwellenparameter (τ) der vier Kategorienantworten (c) sind der Schwierigkeit nach aufsteigend angeordnet.

Quelle: Döring & Bortz (2016, S. 484).

Die logistische Funktion wird als kategoriencharakteristische Funktion (*category characteristic function*, CC-Funktion) bezeichnet und grafisch als kategoriencharakteristische Kurve (*category characteristic curve*, CC-Kurve) dargestellt (vgl. Abbildung 5-6). Die CC-Kurven des Items entsprechen der Lösungswahrscheinlichkeit für jede einzelne Antwortkategorie (Döring & Bortz, 2016, S. 486f.; Rost, 2004, S. 203ff.; Strobl, 2010, S. 54ff.): Bei der ersten Kategorie ist die Lösungswahrscheinlichkeit für Personen mit geringer Fähigkeit (z.B. $\theta = -4$) sehr hoch ($p \approx 1.0$), für jene mit hoher Fähigkeit (z.B. $\theta = +4$) hingegen sehr gering ($p \approx 0.0$). Bei den mittleren Kategorien ist die Lösungswahrscheinlichkeit zuerst gering, dann am höchsten (z.B. $\theta = 0.5$ und $\theta = 1$), zuletzt wiederum gering. Bei der letzten Kategorie ist die Lösungswahrscheinlichkeit für Personen mit geringer Fähigkeit (z.B. $\theta = -2$) sehr gering ($p \approx 0.0$), für jene mit hoher Fähigkeit (z.B. $\theta = +4$) sehr hoch ($p \approx 1.0$). Die Kategorienwahrscheinlichkeiten müssen sich zu 1 addieren: $\sum_{s=0}^{k-1} p_x = 1$.

Die Schnittpunkte der CC-Kurven eines Items i ergeben die Schwellenparameter (τ_{is}) (vgl. Abbildung 5-6, gestrichelte Linien). Bei k Antwortkategorien ergeben sich $k-1$ Schwellenparameter τ_{js} . Die Interpretation des Schwellenparameters im Partial-Credit-Rasch-Modell entspricht der Interpretation des Schwierigkeitsparameters im Rasch-Modell. Je nach Ausprägung der Parameter können drei Fälle unterschieden werden (Eid & Schmidt, 2014, S. 233f.):

- $\theta_v > \tau_{ic}$: Der Personenparameter ist grösser als der Schwellenparameter; die Lösungswahrscheinlichkeit für die *obere* Kategorie ist *grösser* als jene für die untere.
- $\theta_v = \tau_{ic}$: Der Personenparameter entspricht dem Schwellenparameter; die Lösungswahrscheinlichkeit für die untere und obere Kategorie ist *gleich* ($p = 0.5$).
- $\theta_v < \tau_{ic}$: Der Personenparameter ist kleiner als der Schwellenparameter; die Lösungswahrscheinlichkeit für die *untere* Kategorie ist *grösser* als jene für die obere.

Anhand der Anordnung der Schwellenparameter kann die ordinale Struktur der Antwortkategorien interpretiert und evaluiert werden. Das Partial-Credit-Modell verlangt geordnete Schwellenparameter, d.h. $\tau_{i1} < \tau_{i2} < \dots < \tau_{ik-1}$, aber die Abstände zwischen den Schwellenparametern dürfen unterschiedlich gross sein (Döring & Bortz, 2016, S. 486; Eid & Schmidt, 2014, S. 237ff.; Rost, 2004, S. 203ff.; Strobl, 2010, S. 54f.).⁵⁸ Sind die Schwellenparameter ungeordnet, so gibt dies wertvolle Hinweise für die Item-Konstruktion und -Revision, insbesondere zur Item-Formulierung und Item-Kodierung (vgl. Kapitel 6.2.2).

Die zentralen Annahmen des Rasch-Modells (vgl. Kapitel 5.2.3.b) gelten auch für das Partial-Credit-Rasch-Modell, wobei es sich hierbei um *Item- und Kategorien-Homogenität* handelt. Die *Normierung* der Personenparameter erfolgt im Partial-Credit-Rasch-Modell i.d.R. ebenfalls über die Summennormierung, d.h. die Summe aller Schwellenparameter über alle Items hinweg wird auf 0 fixiert: $\sum_{v=1}^n \sum_{s=0}^{k-1} \tau_{ic} = 0$. Über die Summennormierung können die Werte der Personenparameter und der Itemparameter eindeutig bestimmt werden (Rost, 2004, S. 211).

Die Methoden der Parameterschätzung und der Modellgültigkeitsprüfung des dichotomen Rasch-Modells (vgl. Kapitel 5.2.3.c und 5.2.3.d) sind auch für das Partial-Credit-Rasch-Modell anwendbar (Eid & Schmidt, 2014, S. 241ff.).

⁵⁸ Das ordinale Äquidistanz-Modell (vgl. Abbildung 5-4) fordert geordnete Schwellenparameter und zusätzlich gleich grosse Abstände zwischen den Schwellenparametern. Wenn diese beiden Anforderungen für alle ordinalen Items des Tests erfüllt sind, dann kann das Skalenniveau der Items als intervallskaliert betrachtet werden (Döring & Bortz, 2016, S. 488; Rost, 2004, S. 215ff.; Strobl, 2010, S. 57).

6 Standards für pädagogisches-psychologisches Testen

Standards für pädagogisches und psychologisches Testen, kurz Teststandards, sind vereinheitlichte Leitlinien und Handlungsanweisungen, die sich auf verschiedene Phasen bzw. Bereiche des pädagogischen-psychologischen Testens beziehen und deren Optimierung anstreben (Moosbrugger & Höfling, 2012, S. 204). In den Standards widerspiegeln sich Zielsetzungen für die *Testentwicklung* (Konstruktion und Evaluation des Tests), die *Testadaptation* (Übersetzung und Anpassung des Tests), die *Testanwendung* (Durchführung, Auswertung und Interpretation des Tests) und die *Qualitätsbeurteilung* (Überprüfung der Einhaltung der Standards). Standards für die Testentwicklung sind kongruent mit den Gütekriterien für einen psychologischen Test (Moosbrugger & Höfling, 2012, S. 206f.) (vgl. Kapitel 5.1.2). Standards für die Testanwendung sind eine Voraussetzung für objektive, reliable und valide Testwerte (Moosbrugger & Höfling, 2012, S. 211f.).

Teststandards werden von verschiedenen wissenschaftlichen Organisationen erarbeitet und herausgegeben. In der empirischen Bildungsforschung sind insbesondere die *Standards for Educational and Psychological Testing*, kurz *Standards*, einschlägig, die erstmals 1954 erschienen sind und mittlerweile in der sechsten Auflage vorliegen (AERA et al., 2014). Eine Zusammenfassung dieser Standards findet sich in vielen Lehrbüchern zur Testtheorie und Testkonstruktion (vgl. z.B. Bühner, 2011; Eid & Schmidt, 2014; Moosbrugger & Kelava, 2012b; Rost, 2004).

Dieses Kapitel erläutert die für diese Dissertation wichtigen Standards für psychologisches Testen, nämlich jene der evidenzbasierten Validierung (Kapitel 6.1), der Testentwicklung (Kapitel 6.2) und der Testanwendung (Kapitel 6.3). Für Erläuterungen zu weiteren Standards sei auf die einschlägige Literatur zu den Standards verwiesen (vgl. AERA et al., 2014; Jonkisz et al., 2012; Moosbrugger & Höfling, 2012).

6.1 Standards für evidenzbasierte Validierung

Dieses Kapitel erläutert zuerst die Schritte des Validierungsprozesses, der sowohl qualitative als auch quantitative Ansätze beinhaltet (Kapitel 6.1.1). Danach werden die fünf Evidenzquellen der Validität unterschieden (Kapitel 6.1.2).

6.1.1 Validierungsprozess

Der Prozess der Validierung umfasst das Sammeln von Evidenz für die spezifische Interpretation der Testwerte. Es wird also die Interpretation der Testwerte und nicht der Test per se evaluiert (AERA et al., 2014, S. 11).

«The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores for proposed uses that are evaluated, not the test itself» (AERA et al., 2014, S. 11).

Der Validierungsprozess beginnt mit einer Aussage zur empfohlenen Interpretation der Testwerte, begleitet von einer Begründung bezüglich der Bedeutung dieser Interpretation für den Testeinsatz, sog. *argument-based approach to validation* (Kane, 1992, 2013).

«Validation can be viewed as a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use» (AERA et al., 2014, S. 11).

Im Verlaufe des Validierungsprozesses sollen verschiedene Evidenzquellen zusammengetragen werden, um eine wissenschaftliche Grundlage für die beabsichtigte Interpretation der Testwerte zu schaffen (vgl. Kapitel 6.1.2). Wenn die Testwerte auf verschiedene Art interpretiert werden sollen (z.B. normorientiert und/oder kriteriumsorientiert, diagnostisch und/oder prognostisch), dann muss auch jede beabsichtigte Interpretation evaluiert werden (AERA et al., 2014, S. 11, S. 14). Findet man für eine Interpretation der Testwerte starke oder keine Evidenz, muss trotzdem für alle anderen Interpretationen Evidenz gesammelt werden (AERA et al., 2014, S. 13). Mit anderen Worten: Für jede spezifische Annahme zum Konstrukt ist Evidenz erforderlich, und Evidenz für eine spezifische Annahme erlaubt nicht das Verkürzen oder Weglassen der Evidenz für eine andere Annahme (AERA et al., 2014, S. 14).

Empirische Evidenz kann sowohl vom vorliegenden Test und seinem Kontext gewonnen werden als auch von ähnlichen Tests und ihren Kontexten (AERA et al., 2014, S. 13). Um die massgeblichen Arten von Evidenz zu klären, ist es hilfreich, eine Reihe von Annahmen zum Konstrukt zu formulieren, welche die vorgeschlagene Interpretation der Testwerte bzw. des Testeinsatzes stützen (AERA et al., 2014, S. 12). Um eine nicht-vorgeschlagene Interpretation der Testwerte zu finden, können konkurrierende Hypothesen formuliert werden, die sich auf sog. *construct deficiency* (auch: *construct underrepresentation*) oder auf sog. *construct contamination* (auch: *construct irrelevance*) beziehen (AERA et al., 2014, S. 12). Bei einer *construct deficiency* deckt der Test nicht alle relevanten Aspekte⁵⁹ des beschriebenen Konstrukts ab; die Testwerte werden also hinsichtlich des Konstrukts zu eng interpretiert. Umgekehrt umfasst der Test bei der *construct contamination* für das Konstrukt irrelevante Aspekte⁶⁰; die Testwerte werden also durch Konstrukt-externe und Konstrukt-irrelevante Prozesse beeinflusst und hinsichtlich des Konstrukts zu weit interpretiert.

Zum Validierungsprozess gehört – neben der relevanten und adäquaten Konstrukt-Repräsentation im Rahmen der Testentwicklung und -evaluation – auch eine sorgfältige Testadministration (z.B. bzgl. Testbedingungen). Denn beides kann die beabsichtigte Interpretation der Testwerte limitieren bzw. qualifizieren. Insofern kann der Validierungsprozess zu einer Revision

⁵⁹ Bspw. Inhaltselemente, Denkprozesse, Antwortformate.

⁶⁰ Bspw. zu hohes/geringes Sprachniveau der Testmaterialien, emotionale Reaktion auf Testinhalte, Vertrautheit mit Testinhalten, Schreibkompetenz bei freien Antwortformaten.

des Konstrukts und/oder des Tests führen und die Revision ihrerseits eine erneute Validierung der Interpretation der Testwerte nach sich ziehen (AERA et al., 2014, S. 13).

6.1.2 Evidenzquellen für Validität

In den *Standards* werden fünf Evidenzquellen für Validität beschrieben, die verschiedene Aspekte der Validierung beleuchten (AERA et al., 2014, S. 13ff.): Testinhalt, Denkprozesse, Interne Struktur, Beziehung zu externen Merkmalen, Konsequenzen der Testung. Für den Validierungsprozess sind nicht alle Evidenzquellen erforderlich, jedoch braucht es in der Regel mehrere Evidenzquellen, um die vorgeschlagene Interpretation der Testwerte und den beabsichtigten Testeinsatz zu stützen.

Bei Evidenz zum Testinhalt (*evidence based on test content*) geht es um die Beziehung zwischen dem Inhalt des Tests und der Inhaltsdomäne des Konstrukts (AERA et al., 2014, S. 14f.). Der Testinhalt basiert i.d.R. auf detaillierten Inhaltsspezifikationen des Tests, dazu gehören Inhaltsbereiche, Itemtypen, Kodierung (*scoring*) und Sprache (*wording*). Evidenz des Testinhalts lässt sich auf zwei Arten gewinnen: Zum einen kann die Relevanz (Bedeutsamkeit) und die Adäquanz (Angemessenheit) mit Bezug auf die Inhaltsspezifikationen des Tests theoretisch analysiert werden: Wie relevant ist der Testinhalt für die Inhaltsdomäne des Konstrukts und für die vorgeschlagenen Interpretation der Testwerte? Wie adäquat repräsentiert der Testinhalt die Inhaltsdomäne des Konstrukts? Zum anderen können Experten die Relevanz und Adäquanz der Testitems hinsichtlich des Konstrukts empirisch einschätzen: Inwiefern beinhalten die Items konstrukt-irrelevante Schwierigkeit/Einfachheit? Evidenz zum Testinhalt gibt u.a. Hinweise zu *construct underrepresentation* oder *construct irrelevance*, die verschiedene Zielsubgruppen unfair behandelt, d.h. unbeabsichtigt begünstigt oder benachteiligt. Evidenz zum Testinhalt kann auch für die Interpretation der Unterschiede in den Testwerten bei verschiedenen Zielsubgruppen verwendet werden.

Evidenz zu Denkprozessen (*evidence based on response processes*) befasst sich mit der Beziehung zwischen den tatsächlichen Antworten der Testpersonen und den für das Konstrukt theoretisch angenommenen Denkprozessen (AERA et al., 2014, S. 15f.). Dabei werden die individuellen Antworten der Testpersonen mittels Audios oder Videos aufgezeichnet und inhaltsanalytisch ausgewertet. Diese Analysen können sowohl die Definition des Konstrukts anreichern als auch die Revision einzelner Iteminhalte und -formate anregen.

Bei Evidenz zur internen Struktur (*evidence based on internal structure*) geht es um die Beziehung zwischen einzelnen Testitems und Testkomponenten sowie um die Übereinstimmung dieser Beziehung mit dem Konstrukt (AERA et al., 2014, S. 16). Wird für das Konstrukt Eindimensionalität postuliert, braucht es Evidenz dafür, dass alle Items des Tests ähnlich sind und dasselbe Konstrukt abbilden (Test-Eindimensionalität). Wird hingegen ein mehrdimensionales

Konstrukt angenommen, braucht es Evidenz dafür, dass die Items innerhalb einer Testkomponente ähnlich sind und die Items zwischen den Testkomponenten verschieden sind (Test-Mehrdimensionalität). Des Weiteren geht es bei der Evidenz zur internen Struktur um die Beziehung zwischen einzelnen Items und Stichprobengruppen, d.h. ob die Items für bestimmte Gruppen gleich funktionieren (Item-Homogenität), sog. *differential item functioning* (DIF) (vgl. Kapitel 5.2.3.d).

Evidenz zur Beziehung zu externen Merkmalen (*evidence based on relations to other variables*) behandelt die Beziehung von Test-externen Variablen zum Konstrukt (AERA et al., 2014, S. 16ff.). Dabei können zwei Arten von Beziehungen analysiert werden: erstens die Beziehung des Konstrukts mit einem anderen Konstrukt, zweitens die Beziehung des Konstrukts mit einem Aussenkriterium (sog. *test-criterion relationship*). Wird das zu messende Konstrukt mit einem gleichen, ähnlichen oder eng verwandten Konstrukt verglichen (sog. *convergent evidence*), ist eine hohe Korrelation der jeweiligen Testwerte zu erwarten.⁶¹ Wird hingegen das Konstrukt mit einem ungleichen, unähnlichen oder entfernt verwandten Konstrukt verglichen (sog. *divergent evidence*), ist eine weniger hohe Korrelation der Testwerte zu erwarten.⁶² Die Beziehung des zu messenden Konstrukts zu einer externen Kriteriumsvariablen (z.B. Gruppenzugehörigkeit) ist eine forschungsleitende Fragestellung und eine prüfbare Hypothese (Inwiefern ist der Test prädiktiv für die Kriteriumsvariable?). Sie kann entweder zeitgleich (sog. *concurrent evidence*), d.h. die Variablen werden zum gleichen Zeitpunkt erhoben, oder zeitverschoben (sog. *predictive evidence*), d.h. die Variablen werden zu unterschiedlichen Zeitpunkten erhoben, untersucht werden.

Evidenz zu Konsequenzen der Testung (*evidence for consequences of testing*) kann in drei Kategorien eingeteilt werden (AERA et al., 2014, S. 19ff.). 1) Konsequenzen zum Testeinsatz ergeben sich direkt aus der Interpretation der Testwerte und dem beabsichtigten Testeinsatz; 2) Konsequenzen zum Testeinsatz ergeben sich nicht direkt aus der Interpretation der Testwerte oder dem beabsichtigten Testeinsatz; 3) Konsequenzen zum Testeinsatz sind von den Testentwickler*innen nicht beabsichtigt.

In dieser Dissertation wird für die Validierung Evidenz aus drei Quellen zusammentragen: zum Testinhalt (vgl. Kapitel 9), zur internen Struktur (vgl. Kapitel 11) und zu externen Merkmalen (vgl. Kapitel 12). Weitere Evidenz, insbesondere zu Denkprozessen, könnte vor einem erneuten Testeinsatz ebenfalls gesammelt werden.

⁶¹ Bspw. Lesekompetenz gemessen mit einem Multiple-Choice Test und einem Essay-Test.

⁶² Bspw. Lesekompetenz und Rechenkompetenz.

6.2 Standards für Testentwicklung

Testentwicklung ist ein iterativer Prozess, bei dem ein Instrument zur Messung eines individuellen Persönlichkeitsmerkmals (z.B. Wissen, Fertigkeiten, Interesse, Einstellung) konstruiert, evaluiert und revidiert wird. Die Phasen der Testentwicklung sind mit den Testgütekriterien Validität, Reliabilität und Objektivität verwoben (vgl. Kapitel 5.1.2). Insbesondere muss die Testentwicklung die Validität stützen, d.h. die Interpretation der Testergebnisse für ihre beabsichtigte Verwendung (vgl. Kapitel 6.1). Für die Testentwicklung braucht es einen konzeptionellen Rahmen, d.h. eine elaborierte Beschreibung des Konstrukts, das mit dem Test gemessen werden soll: Umfang des Konstrukts, Aspekte des Konstrukts, Unterscheidung des Konstrukts von anderen Konstrukten, Beziehung zu anderen Variablen. Dieser konzeptionelle Rahmen verweist denn auch auf die Arten von Evidenz, die im Rahmen des Validierungsprozesses gesammelt werden sollen, um die vorgeschlagene Interpretation der Testergebnisse zu evaluieren (AERA et al., 2014, S. 11f.).

Der Prozess der Testentwicklung lässt sich in vier Phasen unterteilen (AERA et al., 2014, S. 75ff.): (1) Testspezifikationen entwickeln und evaluieren; (2) Testitems konstruieren, pilotieren und revidieren; (3) Testformen bzw. Testhefte konfektionieren und evaluieren; (4) Instruktionen/Materialien für Testadministration entwickeln. Nachfolgend werden die ersten drei Phasen erläutert.

6.2.1 Testspezifikationen

Die Testspezifikationen umfassen detaillierte Beschreibungen zu Testzweck, Testinhalt, Testformat, Testlänge, Kodierungs-/Bewertungsverfahren, psychometrische Eigenschaften von Items und Test sowie Testadministration (AERA et al., 2014, S. 76ff.; Jonkisz et al., 2012, S. 32ff.).

- Beim *Testzweck* werden Aussagen gemacht zum Konstrukt (insbesondere zur Inhaltsdomäne), zur beabsichtigten Testanwendung (insbesondere normorientierte oder kriteriumsorientierte Interpretation der Testergebnisse), zu den vorgesehenen Testpersonen (Zielgruppe) und zu den vorgesehenen Testanwender*innen.
- Die *Inhaltsspezifikationen* beschreiben den Umfang und die Aspekte des zu messenden Konstrukts bzw. der Inhaltsdomäne. Sie wird entweder theoretisch hergeleitet oder empirisch analysiert (Domänenanalyse).
- Die *Formatspezifikationen* beschreiben das Itemformat (Frage, Aufgabe, Assessment, Simulation, Portfolio), das Antwortformat (*selected response*, *constructed response*) (vgl. Kapitel 6.2.2.b) und das Kodierungsformat (dichotom, polytom) (vgl. Kapitel 6.2.2.c). Sie hängen vom Testzweck, vom Testinhalt und vom Testverfahren ab, nichtdestotrotz sollten sie die Validität der beabsichtigten Interpretation der Testergebnisse stützen. So sollten bspw. Formate vermieden werden, mit denen die Testpersonen oder bestimmte Gruppen von

Testpersonen nicht vertraut sind. Die Formatspezifikationen sollten zudem Überlegungen zu alternativen Formaten beinhalten, mit denen Konstrukt-irrelevante Varianz in den Testergebnissen reduziert oder beseitigt werden kann.

- Mit der *Testlänge* wird die Anzahl Items in jedem Testteil (Inhaltsbereich) bzw. im gesamten Test definiert, die jeder Testperson vorgelegt wird. Dafür müssen Kriterien der Testökonomie, der verfügbaren/zumutbaren Testzeit, der Genauigkeit der Testergebnisse sowie der Verteilung von Itemtypen (Kognitionsprozess, Antwortformat) berücksichtigt werden.
- Die *Kodierungs-/Bewertungsspezifikationen* (*scoring specifications*) beschreiben die Bewertung der einzelnen Items und die Verrechnung der Itemwerte zu einem Testwert. Dabei muss für jedes Item ein Kodierungs- bzw. Bewertungsschema erarbeitet werden: Bei Selected-Response-Items wird festgelegt, für welche Antwortoption welcher *Wert* vergeben wird; bei Constructed-Response-Items werden für die Bewertung Kriterien und Stufen bestimmt. Ebenfalls muss für die Kodierungsspezifikationen festgelegt werden, über welche Qualifikation die Kodierer*innen verfügen müssen, wie die Kodierer*innen geschult und kontrolliert werden, wie Nichtübereinstimmungen in der Kodierung erkannt und beseitigt werden und wie Übereinstimmungen überprüft wird.
- Die *psychometrischen Spezifikationen* bezeichnen die gewünschte statistische Qualität der einzelnen Items (z.B. Item-Schwierigkeit, Item-Trennschärfe, Inter-Item-Korrelation) und des gesamten Tests (z.B. Test-Schwierigkeit, Test-Genauigkeit, Item-Verteilung über Inhaltsbereiche und Kognitionsstufen). Bei einer Modellschätzung basierend auf der IRT, wird die Passung des psychometrischen Testmodells auf die Daten evaluiert, d.h. inwiefern die Annahmen des Testmodells erfüllt sind (z.B. Eindimensionalität des Tests, Homogenität der Items) (vgl. Kapitel 5.2.3).
- Die *Testadministrationsspezifikationen* legen das Testverfahren (*Paper-and-Pencil*, Computer), Zeitbegrenzungen, Instruktionen und Materialien für Testleitende und Testpersonen, Aufbewahrung von Testmaterialien etc. fest.

6.2.2 Item-Konstruktion

Items sind Bestandteile eines Tests (z.B. Fragen, Aufgaben, Bilder), mit denen ein Verhalten, d.h. eine Antwort hervorgerufen werden soll. Ein Item besteht aus Itemstamm und Antwortformat. Für jedes Item braucht es ein Kodierungsschema, damit die Antworten objektiv und reliabel ausgewertet sowie valide interpretiert werden können.

Bei der Item-Konstruktion wird i.d.R. ein Itempool zusammengestellt, der die benötigte Anzahl Items für den Test bzw. die Testform übersteigt. Dieses Vorgehen erlaubt es, eine genügend grosse Anzahl Items theoretisch (*review*) und empirisch (*pretesting*) zu erproben (AERA et al., 2014, S. 81f.). Durch die psychometrischen Analysen zur Gültigkeit des probabilistischen Testmodells (vgl. Kapitel

5.2.3.d) erweisen sich erfahrungsgemäss bei neu entwickelten Tests viele Items als nicht modellkonform und bedürfen einer Revision. Für die finale Testform reichen i.d.R. 20 Items aus (Döring & Bortz, 2016, S. 491).

Beim *Review* werden die Items bezüglich Klarheit, Korrektheit und Konstrukt-irrelevanter Aspekte von Experten beurteilt und ggf. revidiert (Evidenzquelle Testinhalt, vgl. Kapitel 6.1.2). Beim *Cognitive Pretesting* werden mit ausgewählten Probanden kognitive Interviews (*think aloud*) geführt, um nicht beabsichtigte Antwortbarrieren aufzudecken und Evidenz für Denkprozesse zu sammeln (Evidenzquelle Denkprozesse, vgl. Kapitel 6.1.2). Beim *Standard Pretesting* werden die Items einer Gruppe von repräsentativen Probanden vorgelegt, um die psychometrischen Eigenschaften der Items (z.B. Item-Schwierigkeit, Item-Trennschärfe, DIF) zu analysieren und Items ggf. zu revidieren oder zu eliminieren. Ebenfalls werden beim *Pretesting* die Kodierungsschemata der Items erprobt.

Bei Leistungstests gilt es zudem zu beurteilen, ob genügend Items verschiedener Schwierigkeitsstufen vorhanden sind. Dies ist nötig, damit der Test nicht nur im mittleren Leistungsbereich, sondern auch im unteren und oberen zwischen den Testpersonen differenzieren kann (Jonkisz et al., 2012, S. 66f.; Kelava & Moosbrugger, 2012, S. 76ff.).

6.2.2.a Itemstamm

Ein Item besteht aus zwei Komponenten, dem Itemstamm und dem Antwortformat (Jonkisz et al., 2012, S. 39; Rost, 2004, S. 55). Der *Itemstamm* ist bei Leistungstests eine Frage oder Aufgabe, zu der die Testperson die Lösung angeben soll; bei Persönlichkeitstests eine Frage oder Aussage, zu der die Testperson Stellung beziehen soll. Bei der Formulierung des Itemstamms sind die sprachliche Klarheit und sprachliche Eindeutigkeit besonders wichtig. Die Testpersonen sollten die Items beim erstmaligen Lesen verstehen; somit können Fehlinterpretationen und Motivationseinbussen vermieden werden. Alle Testpersonen sollten den Iteminhalt in gleicher Weise verstehen, um eine Vergleichbarkeit der Testwerte zu ermöglichen. Es gilt folgende Aspekte zu beachten (Jonkisz et al., 2012, S. 64ff.):

- Items positiv formulieren: Verneinungen (Negationen), insbesondere doppelte Verneinungen vermeiden.
- Items mit klaren Satzkonstruktionen formulieren: umständliche und komplizierte Satzkonstruktionen vermeiden.
- Wortabkürzungen in Items vermeiden, wie beispielsweise „z.B.“, „u.a.“, „u.U.“, „i.d.R.“.
- Items sprachlich der Zielgruppe anpassen: Fachbegriffe in Items vermeiden, insbesondere wenn sie nur einem kleinen Teil der Zielgruppe geläufig sind.
- Universalausdrücke in Items vermeiden, wie beispielsweise „immer“, „nie“, „alle“.
- Schwierige Begriffe definieren, bevor die eigentliche Frage/Aufgabe gestellt wird.

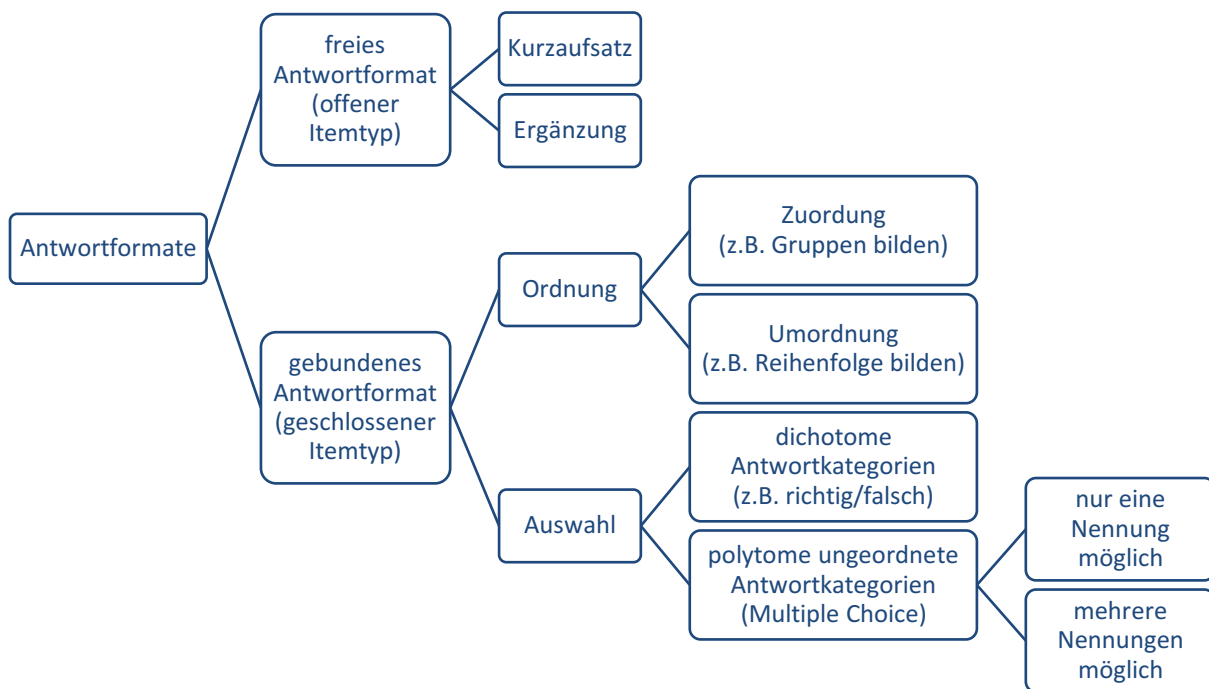


Abbildung 6-1. Antwortformate für Leistungstest

Quelle: Eigene Darstellung i.A.a. Jonkisz et al. (2012, S. 39) und Eid & Schmidt (2014, S. 100).

6.2.2.b Antwortformate

Das *Antwortformat* kann bei Leistungstests frei (auch: offener Itemtyp) oder gebunden (auch: geschlossener Itemtyp) sein (Eid & Schmidt, 2014, S. 97ff.; Jonkisz et al., 2012, S. 39ff.; Rost, 2004, S. 59ff.) (vgl. Abbildung 6-1). Bei *freien Antwortformaten* wird die Antwortformulierung der Testperson überlassen (sog. *constructed response*). Dabei wird zwischen zwei Formaten unterschieden: Bei *Kurzaufsatz-Items* besteht die Antwort aus einigen Worten, einige Sätzen oder einem kurzen Essay; bei *Ergänzungs-Items*, wird die Antwort um Zahlenreihen, Wortgruppen oder Lückentexte vervollständigt.

Bei *gebundenen Antwortformaten* wird die Antwortformulierung der Testpersonen vorgegeben. Bei gebundenen *Ordnungs-Items* müssen Beziehungen zwischen Elementen geordnet werden, entweder als Zuordnung (z.B. sinnvolle Gruppen bilden) oder als Umordnung (z.B. in sinnvolle Reihenfolge bringen). Bei gebundenen *Auswahl-Items* (sog. *selected response*) muss eine Antwortoption ausgewählt werden: Bei dichotomen Antwortkategorien stehen zwei Antwortoptionen zur Auswahl (auch: Alternativantworten, z.B. ja/nein, richtig/falsch); bei polytomen ungeordneten Antwortkategorien werden mehrere Antwortoptionen vorgegeben (auch: Mehrfachwahlantworten, *multiple choice*, z.B. A/B/C/D), wobei eine Antwort (*single answer question*) oder mehrere Antworten (*multiple answer question*) richtig sein können.

Ein Vorteil der freien Antwortformate ist, dass mit ihnen höhere kognitive Prozesse abgerufen werden können (Reproduktionsleistung) und Ratewahrscheinlichkeit sehr gering ist. Ein erster

Nachteil ist, dass sie bei der Testdurchführung eine längere Bearbeitungszeit beanspruchen und dass sie Testpersonen mit Schreibschwierigkeiten benachteiligen. Ein zweiter Nachteil ist, dass die Wörter bzw. Sätze für die Testauswertung in Ziffern (*codes*) überführt werden müssen (Antwortkodierung) und dass dadurch die Auswertungsobjektivität eingeschränkt ist (Döring & Bortz, 2016, S. 244, S. 454f.; Jonkisz et al., 2012, S. 40ff.) (vgl. Kapitel 5.1.2.a).

Ein Vorteil der gebundenen Antwortformate ist die Testökonomie, d.h. ein geringer Zeit- und Kostenaufwand bei der Testdurchführung und bei der Testauswertung durch maschinelle oder elektronische Auswertung. Ein zweiter Vorteil ist die eindeutige Antwortkodierung und somit eine sehr hohe Auswertungsobjektivität (Döring & Bortz, 2016, S. 455; Jonkisz et al., 2012, S. 43ff.) (vgl. Kapitel 5.1.2.a). Der grösste Nachteil von gebundenen Mehrfachauswahl-Items ist die hohe Ratewahrscheinlichkeit: je weniger Distraktoren, desto höher die Ratewahrscheinlichkeit (Jonkisz et al., 2012, S. 49f.).⁶³ Der zweite Nachteil liegt in der Testfairness: Manche Testpersonen kreuzen eine Antwortoption an, auch wenn sie sich über deren Richtigkeit unsicher sind, erzielen also durch Raten einen höheren Testwert; andere Testpersonen kreuzen eine Antwortoption nur an, wenn sie von deren Richtigkeit überzeugt sind (Döring & Bortz, 2014, S. 455). Ein dritter Nachteil ist, dass das erfolgreiche Bearbeiten des Items lediglich eine Rekognitionsleistung (Wiedererkennung) voraussetzt (Jonkisz et al., 2012, S. 44). Einen vierten Nachteil der Mehrfachauswahl-Items stellt die zeitaufwändige Konstruktion der Antwortoptionen dar (Döring & Bortz, 2016, S. 455; Jonkisz et al., 2012, S. 45f.):

- Die Antwortoptionen müssen disjunkt sein, d.h. sich gegenseitig ausschliessen und keine Schnittmenge bilden.
- Nur eine der Antwortoptionen soll richtig sein, nicht mehrere.
- Die falschen Antwortoptionen (auch: Distraktoren) müssen derart attraktiv und plausibel sein, dass sie von der richtigen Antwortoption ablenken und von der Testperson mit ungefähr gleich hoher Wahrscheinlichkeit für richtig eingeschätzt werden.
- Es muss zwischen Qualität und Anzahl der Antwortoptionen der Antwortoptionen abgewogen werden: einen zusätzlichen, gleichwertigen Distraktor zu formulieren ist schwierig; je mehr Distraktoren vorliegen, desto geringer die Ratewahrscheinlichkeit). Es sollten i.d.R. drei Antwortoptionen ausreichend (Rodriguez, 2005).

⁶³ Bei zwei Antwortkategorien beträgt die Ratewahrscheinlichkeit 50 %, bei vier ungeordneten Antwortkategorien 25 %.

Tabelle 6-1. Exemplarisches Kodierungsschema für Items in einem Leistungstest

Item A			Item B		
Code	Auswertung	Antwortbeispiel	Code	Auswertung	Antwortbeispiel
1	richtig	...	2	vollständig richtig	...
0	falsch	...	1	teilweise richtig	...
			0	falsch	...
888	nicht kodierbar	Juxantwort	888	nicht kodierbar	Juxantwort
999	keine Antwort	---	999	keine Antwort	---

Hinweis: Aus der Kodierung von Item A folgt eine dichotome Antwortvariable, aus jener von Item B eine polytome geordnete Antwortvariable.

Quelle: Eigene Darstellung i.A.a. Eid & Schmidt (2014, S. 124f.).

6.2.2.c Item-Kodierung

Kodieren ist das numerische Bewerten von gebundenen und freien Item-Antworten aus einem Leistungstest, d.h. das Überführen von Text in Zahl (z.B. 0/1, 0/1/2) (Eid & Schmidt, 2014, S. 124f.). Durch die Item-Kodierung entstehen sog. Antwortvariablen. Für die Item-Kodierung braucht es ein detailliertes Kodierungs- bzw. Bewertungsschema, anhand dessen die Antworten der Testpersonen inhaltlich beurteilt (z.B. falsch/richtig, falsch/teilweise richtig/vollständig richtig) und numerisch bewertet werden. Das Kodierungsschema sollte mit Bewertungshinweisen, Antwortbeschreibungen und Beispielantworten versehen sein (vgl. Tabelle 6-1). Es ist z.B. durchaus denkbar, dass es unterschiedliche richtige Antworten gibt, die mit demselben Code versehen werden, da Testpersonen selten „die Musterlösung“ produzieren.

Bei einem Leistungstest sind auch zusätzliche Kodierungen denkbar, z.B. betreffend Sprachgebrauch (z.B. Satzbau, Wortarten, Fachtermini) oder Argumentationsstruktur⁶⁴ (z.B. bejahende/verneinende Behauptung, stützende/entkräftende Begründung, alltägliche oder wissenschaftliche Begründung), die zu kategorialen (nominalskaliert: Arten) und kardinalen (intervallskaliert: Anzahl) Antwortvariablen führen.

Die Codes der Item-Kodierung werden in eine Person x Item-Datenmatrix überführt und für quantitative Analysen verwendet (Eid & Schmidt, 2014, S. 124f.). Die Antwortvariablen eines Leistungstests sind i.d.R. dichotom oder polytom geordnet (Eid & Schmidt, 2014, S. 130, S. 224). Insofern bestimmen das Antwortformat *und* die Item-Kodierung das Skalenniveau der Daten und jenes wiederum das passende probabilistische Testmodell (vgl. Kapitel 5.2.2).

⁶⁴ Eine explorative Studie zur Argumentationsstruktur der Schüler*innen-Antworten im WBK-T2 ist derzeit noch in Arbeit (Ackermann & Kavadarli, 2019, in progress).

6.2.3 Testheft

Beim Zusammenstellen der Items zu einem Testheft gilt es folgende Aspekte zu beachten (Jonkisz et al., 2012, S. 68f.):

- Reihenfolge der Items: Bei Leistungstests werden die Items üblicherweise in aufsteigender Schwierigkeit angeordnet, um eine Überforderung und Demotivation der Testpersonen zu vermeiden: einfache, von den meisten Probanden lösbare Items am Anfang des Tests (sog. „Eisbrecher“); schwierige, von den wenigsten Probanden lösbare Items am Ende des Tests. Es kann aber auch sinnvoll sein, die schwierigen Items oder Testteile eher am Anfang des Tests zu platzieren, solange die Konzentration noch hoch ist. Zwischen den Items darf es keine inhaltlichen Abhängigkeiten geben, insbesondere dürfen sich aufeinanderfolgende Items nicht wechselseitig erschweren oder erleichtern (sog. Positionseffekt, Aktualisierungseffekt).
- Instruktion (Testanweisung): Die Instruktion am Testanfang soll die Probanden zur Mitarbeit ermuntern. Sie enthält klare Handlungsanweisungen, so z.B. eine Erläuterung zum Antwortmodus der Items, ein Beispielitem und eine Beispielantwort.
- Layout (Gestaltung): Das Layout des Testmaterials soll einfach und übersichtlich sein, optisch ansprechend und der Zielgruppe angepasst sein, aber v.a. die Bearbeitung des Tests erleichtern.
- Anonymität: Bei wissenschaftlichen Studien ist es wichtig, die Testpersonen explizit auf Anonymität der erhobenen Daten bei der Testauswertung hinzuweisen und eine Kontaktperson der Forschungsinstitution anzugeben. Dies ist einerseits für den Datenschutz unerlässlich, andererseits können die Probanden offener auf die Items antworten.

6.3 Standards für Testanwendung

Der Prozess der Testanwendung lässt sich in zwei Hauptphasen unterteilen (Moosbrugger & Höfling, 2012, S. 213ff.), nämlich (1) Testdurchführung und (2) Testauswertung. Zu den Phasen der Testauswertung gehören: (1a) Vorbereitungsphase und (1b) Testphase; zu den Phasen der Testauswertung zählen: (2a) Ergebnisse, (2b) Interpretation, (2c) Sicherung. Eine systematische Zusammenstellung der betreffenden Standards und Kommentare findet sich in der einschlägigen Literatur (AERA et al., 2014, S. 111ff., S. 123ff.; Moosbrugger & Höfling, 2012, S. 211ff.).

7 Entwicklungsgeschichte des WBK-Tests

In diesem Kapitel wird der Leitfrage (5) nachgegangen (vgl. Kapitel 1.2):

- (5) *Welche fachdidaktischen und psychometrischen Kriterien sollten bei der Testevaluation und -revision des WBK-T1 berücksichtigt werden, um die Qualität des WBK-T2 zu optimieren?*

Der Test zur wirtschaftsbürgerlichen Kompetenz (WBK-Test) wurde im Rahmen des Forschungsprojekts CoBALIT entwickelt und hinsichtlich einiger Aspekte validiert (Eberle et al., 2016; Schumann et al., 2017). Später wurde der WBK-Test im Rahmen des Forschungsprojekts WBKgym und insbesondere für die vorliegende Dissertation evaluiert, revidiert und hinsichtlich weiterer Aspekte validiert (vgl. Kapitel 9, 11 und 12). Eine ausführliche Darstellung der Testentwicklung und Testrevision findet sich in der „Dokumentation der Entwicklung und Revision des WBK-Tests“ (Ackermann, 2018a) und ist auf Anfrage bei der Autorin erhältlich. In dieser Dissertation wird die originale Version des WBK-Tests als WBK-T1 bezeichnet, die – von der Autorin – revidierte Version als WBK-T2.

Dieses Kapitel erzählt die Entwicklungsgeschichte des WBK-Tests von seiner Konstruktion und Adaption (Kapitel 7.1) bis zu seiner Revision (Kapitel 7.2).

7.1 CoBALIT und WBK-T1

In diesem Kapitel wird der Hintergrund des Forschungsprojekts CoBALIT geschildert (Kapitel 7.1.1), die Entwicklung und Adaption des WBK-T1 (Kapitel 7.1.2) sowie die psychometrische Qualität des WBK-T1 (Kapitel 7.1.3).

7.1.1 Hintergrund des Projekts CoBALIT

Das Verbundprojekt CoBALIT (*Competencies in the Field of Business and Administration, Learning, Instruction, and Transition*) war eines von mehreren Projekten, das im Rahmen der Forschungsinitiative ASCOT (*Technology-based Assessment of Skills and Competencies in Vocational Education and Training*) des deutschen Bundesministeriums für Bildung und Forschung (BMBF) zwischen 2011 und 2015 gefördert wurde (K. Beck et al., 2016; BMBF, 2012). Die Ziele des Verbundprojekts CoBALIT waren zum einen die Entwicklung von technologiebasierten Instrumenten zur Erfassung kaufmännischer Kompetenzen, zum anderen die Messung dieser Kompetenzen und der Ausbildungsbedingungen bei Lernenden in bestimmten kaufmännischen Berufen bzw. Branchen (S. Weber et al., 2016; Winther, Seeber, Festner, Sangmeister, & Liedtke, 2016).

Die im Verbundprojekt CoBALIT integrierte Studie „Modellierung und Messung wirtschaftsbürgerlicher Kompetenz im kaufmännischen Bereich“ wurde vom Schweizer Staatssekretariat für Bildung, Forschung und Innovation (SBFI) zwischen 2012 und 2016 gefördert

(Eberle et al., 2016; Schumann et al., 2017). Mit dieser Studie wurde erstmals die wirtschaftsbürgerliche Kompetenz von kaufmännischen Lernenden technologiebasiert erfasst und für einen Branchen- und Ländervergleich analysiert (Schumann et al., 2017). Die Piloterhebung umfasste 522 kaufmännische Lernende, davon 250 aus der Schweiz und 272 aus Deutschland, an denen das neu entwickelte Testinstrument (vgl. Kapitel 7.1.2) erprobt wurde. Für die Haupterhebung in der Schweiz wurden insgesamt 420 Auszubildende der kaufmännischen Branchen „Maschinen, Elektro- und Metallindustrie“ (SWISSMEM) (N = 295) und „Internationale Speditionslogistik“ (SPEDLOGSWISS) (N = 125) getestet und befragt, in Deutschland waren es insgesamt 839 Auszubildende in den vergleichbaren Berufen „Industriekaufmann/-frau“ (N = 512) und „Kaufmann/Kauffrau für Spedition und Logistikdienstleistungen“ (N = 327).

7.1.2 Entwicklung und Adaption des WBK-T1

Zur Messung des Konstrukts wirtschaftsbürgerliche Kompetenz (WBK) wurde im Rahmen des Projekts CoBALIT ein neuer Leistungstest entwickelt. In Anlehnung an die Definition der WBK (Eberle, 2015; Eberle et al., 2016) (vgl. Kapitel 4.1) wurde der neue Test als wirtschaftsbürgerlicher Kompetenztest (WBK-Test) bezeichnet. Eine zusammenfassende Beschreibung zur Entwicklung und Spezifikation des WBK-T1, insbesondere zur Domänenanalyse und Item-Konstruktion findet sich bei Eberle et al. (2016, S. 101ff.).

Der WBK-Test unterscheidet sich durch zwei wesentliche Merkmale von anderen, damals bereits bestehenden und erprobten Leistungstests für *economic literacy* bzw. ökonomische Kompetenz (vgl. Kapitel 3.4), woraus sich auch die Notwendigkeit dieses neuen Tests ableiten lässt (vgl. Kapitel 5.1.2.d, Nützlichkeit). Erstens, der WBK-Test orientiert sich an einer lebenssituationalen Konzeption ökonomischer Bildung und ist nach einer Themenlogik statt einer Fachlogik strukturiert (Eberle et al., 2016). Im Gegensatz dazu folgen z.B. der Wirtschaftskundliche Bildungstest (WBT) (K. Beck, 1993) und der OEKOMA-Test (Schumann & Eberle, 2014a; Schumann et al., 2010; Schumann et al., 2011) einer fachwissenschaftlichen Konzeption ökonomischer Bildung. Entsprechend wurde bei diesen beiden Tests die Inhaltsauswahl entlang klassischer ökonomischer Kategorien getroffen. Zweitens, der WBK-Test umfasst etwa gleich viele gebundene wie freie Antwortformate. Dies ermöglicht es, mit den Testitems höhere kognitive Prozesse abzurufen (Reproduktionsleistung statt nur Rekognitionsleistung) (vgl. Jonkisz et al., 2012, S. 40ff.) (vgl. Kapitel 6.2.2.b). Im Vergleich dazu besteht der WBT aus nur gebundenen Antwortformaten, der OEKOMA-Test aus ca. 90 % gebundenen Antwortformaten.

Der WBK-Test hat mit dem OEKOMA-Test drei Merkmale gemein (Eberle et al., 2016; Schumann & Eberle, 2014a). Erstens, beide Tests operationalisieren ein Konstrukt, das als Referenzfigur den «mündigen Wirtschafts- und Gesellschaftsbürger» (Schumann & Eberle, 2014a, S. 107) bemüht bzw. «die aktuelle und künftige allgemeine Teilhabe junger Leute am Geschehen in

Wirtschaft und Gesellschaft» (Eberle et al., 2016, S. 94) im Blickfeld hat. Diese (jungen) Menschen sind in der Lage, wirtschaftsbezogene Problemstellungen zu verstehen, zu analysieren und begründete Schlüsse für (potenzielle) Lösungen daraus zu ziehen. Zweitens, beide Tests adressieren eine Zielgruppe am Ende der Sekundarstufe II (Eberle et al., 2016, S. 95, S. 105; Schumann & Eberle, 2014a, S. 106, S. 110). Drittens, beide Tests verwenden modifizierte Medientexte als Itemstämme zur inhaltlichen Strukturierung und Kontextualisierung (Eberle et al., 2016, S. 102; Schumann & Eberle, 2014a, S. 112).

Da mit der Schweizer Studie des Verbundprojekts CoBALIT (Eberle et al., 2016; Schumann et al., 2017) ein Vergleich der wirtschaftsbürgerlichen Kompetenz von kaufmännischen Lernenden in der Schweiz und in Deutschland angestrebt wurde, mussten zwei parallele Versionen des WBK-Tests entwickelt werden (Version CH und eine Version DE), die den spezifischen Kontext der beiden Länder berücksichtigten. Zuerst entwickelte das Forschungsteam zusammen mit Schweizer Experten die Version CH des WBK-T1 für den Schweizer Kontext, die quasi den Prototypen darstellte. Danach adaptierte das Team zusammen mit deutschen Experten die Version DE des WBK-T1 für den deutschen Kontext. Die Adaption bezog sich hauptsächlich auf sprachlich-kulturelle, politisch-institutionelle und technisch-faktische Aspekte (vgl. Tabelle 7-1). Wo immer möglich, wurden für beide Versionen Begriffe und Sätze „neutralisiert“, so dass sie für beide Länder zutreffend und in beiden Ländern verständlich waren. Nichtsdestotrotz mussten aus inhaltlichen Gründen für die Version DE des WBK-T1 zwei Problemsituationen (LAN, STEU) weggelassen werden, die für die Version CH entwickelt worden waren (vgl. Tabelle 7-2).

7.1.3 Psychometrische Qualität des WBK-T1

Bei der CoBALIT-Piloterhebung wurden acht Problemsituationen für die Schweiz mit 74 Items und sechs Problemsituationen für Deutschland mit 57 Items in einem rotierten Testdesign erprobt (vgl. Frey, Hartig, & Rupp, 2009). Aufgrund der klassischen und probabilistischen Itemanalysen wurden danach ca. 20 Items modifiziert und 11 Items ausgeschlossen. Daraus resultierte der WBK-T1 mit 63 Items in der Version CH und 47 in der Version DE (vgl. Tabelle 7-2).

Bei der CoBALIT-Haupterhebung in der Schweiz wurden aus dem WBK-T1 nur zwei Problemsituationen (ENE, STA) mit insgesamt 14 Items eingesetzt. Aufgrund der verfügbaren Erhebungsdauer war es zum einen nicht möglich, mehr als zwei Problemsituationen pro Testperson darzubieten, zum anderen war es aufgrund der kleinen Stichprobe nicht möglich ein *booklet design* einzusetzen. In Deutschland wurden wiederum sechs Problemsituationen (AHV, ENE, STA, MAN, JUG, EUR) mit insgesamt 48 Items des WBK-T1 in einem rotierten *booklet design* eingesetzt, d.h. je zwei Problemsituationen pro Testperson (vgl. Frey et al., 2009). In allen Testheften wurden insgesamt 26 Items aus anderen Instrumenten eingesetzt: 12 Items aus dem WBT und 14 aus dem OEKOMA-Test.

Tabelle 7-1. Beispiele für Adaption des WBK-T1

Kategorie	Version CH (spezifisch)	Version DE (spezifisch)	beide Versionen („neutralisiert“)
sprachlich-kulturell	Lohn, Salär, Gehalt	Gehalt	Gehalt
	Bruttoinlandprodukt	Bruttoinlandsprodukt	---
	Massnahme	Maßnahme	---
politisch-institutionell	Alters- und Hinterlassenenversicherung (AHV)	Gesetzliche Rentenversicherung	---
	Gesamtarbeitsvertrag	Tarifvertrag	---
	Franken, CHF	Euro, EUR	---
	Nationalrat, Ständerat, Bundesversammlung	Bundesrat	Parlament
	Bundesrat	Bundesregierung, Bundeskabinett	Regierung
Technisch-faktisch	Energieverbrauch und Energiestrategie Schweiz	Energieverbrauch und Energiestrategie Deutschland	---
	BIP und Schuldenquote Schweiz	BIP und Schuldenquote Deutschland	---

Mit den Daten der CoBALIT-Haupterhebung wurden in einem ersten Schritt klassische und probabilistische Itemanalysen durchgeführt (Eberle et al., 2016, S. 107ff.). Die Item-Infits (wMNSQ) waren für die Version CH mehrheitlich befriedigend, für die Version DE weniger befriedigend. Bei der Version CH mit 40 Items war die Reliabilität gut ($EAP/PV = .76$), bei der Version DE mit 74 Items war sie entsprechend höher ($EAP/PV = .87$).⁶⁵ Allerdings wurden bei der CoBALIT-Erhebung auch „fremde“ Items aus anderen etablierten Tests, in die Testhefte aufgenommen (12 Items aus dem WBT, 14 Items aus dem OEKOMA-Test) und bei den Analysen integriert, so dass die berichteten Befunde nicht eindeutig und einzig als Qualität des WBK-T1 interpretiert werden können.

In einem zweiten Schritt wurde eine Dimensionalitätsanalyse durchgeführt, hierbei ging es um die Prüfung der domänenverbundenen und domänenspezifischen Kompetenz des Modells der Wirtschaftskompetenz (Eberle et al., 2016, S. 109f.) (vgl. Kapitel 3.4.3). Dafür wurden alle eingesetzten Items vorgängig von Experten den beiden Dimensionen zugeordnet (Eberle et al., 2016, S. 103). Das zweidimensionale Modell zeigte gemäss den Modellfit-Indizes (Devianz, BIC, AIC) eine bessere Anpassung auf die Daten, sowohl für beide Länder getrennt als auch gesamthaft (CH: Δ Deviance = 89, Δ df = 2, $p < .001$; DE: Δ Deviance = 376, Δ df = 2, $p < .001$; Gesamt: Δ Deviance = 417, Δ df = 2, $p < .001$). Die latenten Korrelationen zwischen den beiden Dimensionen deuteten ebenfalls

⁶⁵ Die Reliabilität eines Tests hängt positiv mit der Testlänge, d.h. der Anzahl Items, zusammen (Moosbrugger, 2012b, S. 112).

auf eine zweidimensionale Struktur hin ($r_{CH} = .63$, $r_{DE} = .60$, $r_{gesamt} = .61$). Nichtsdestotrotz wurde für die weiteren inferenzstatistischen Analysen zu Ländern und Branchen/Berufen das eindimensionale Modell verwendet (Schumann et al., 2017).

7.2 WBKgym und WBK-T2

Dieses Kapitel beschreibt den Hintergrund des Forschungsprojekts WBKgym (Kapitel 7.2.1) und die Revision für den WBK-T2 (Kapitel 7.2.2).

7.2.1 Hintergrund des Projekts WBKgym

Das Forschungsprojekt „Wirtschaftsbürgerliche Kompetenz von Deutschschweizer Gymnasiast*innen“ (WBKgym) wurde von der Autorin als Dissertationsvorhaben initiiert, geplant und durchgeführt. Das Ziel des Projekts WBKgym war u.a. die Evaluation und Revision des WBK-Tests und die Validierung der Testergebnisse anhand einer Stichprobe von Gymnasiast*innen (vgl. Kapitel 1.2). Eine Testevaluation und -revision drängte sich aufgrund der Ergebnisse der Itemanalyse und der Erkenntnisse aus der Antwortkodierung auf. Aus der Testrevision des originalen WBK-Tests (WBK-T1) resultierte der revidierte WBK-Test (WBK-T2). Das Vorgehen und die Befunde der Testrevision bilden einen wesentlichen Bestandteil der vorliegenden Dissertation.

7.2.2 Revision des WBK-T1 und Entwicklung des WBK-T2

Der Prozess der Testrevision verlief in mehreren Phasen, die nachfolgend zusammenfassend dargestellt werden. Das Ergebnis der Testrevision wird bei den Testspezifikationen vorgestellt (vgl. Kapitel 8). Eine übersichtliche Darstellung der Testspezifikationen findet sich im Anhang (vgl. Anhang Tabelle A- 4).

Die erste Phase widmete sich der Evaluation der Testinhalte, d.h. der sozioökonomischen Problemsituationen. Die bestehenden Problemsituationen des WBK-T1 wurden bezüglich ihrer Passung mit definierten Kriterien, v.a. der politischen Relevanz geprüft (vgl. Kapitel 8.2.1). Die verschobenen Erhebungszeitpunkte der Projekte CoBALIT und WBKgym legte eine solche Überprüfung nahe (AERA et al., 2014, S. 83f.): Die Problemsituationen für das Projekt CoBALIT waren im Jahr 2012/2013 entwickelt und im Jahr 2013/2014 eingesetzt worden, die Erhebung des Projekts WBKgym war für das Jahr 2017 geplant. Im Ergebnis blieben im WBK-T2 vier Problemsituationen (AHV, ENE, STA, MAN) erhalten (vgl. Tabelle 7-2), nicht zuletzt auch aus Gründen der Forschungsökonomie.

In der zweiten Phase wurden alle Items des WBK-T1 mittels fachdidaktischer Kriterien und psychometrischer Indikatoren systematisch evaluiert. Die fachdidaktischen Kriterien, insbesondere die inhaltliche Relevanz und Adäquanz (AERA et al., 2014, S. 14), stützten sich zum einen auf die Notizen der Autorin im Kodierungsmanual des Projekts CoBALIT, zum anderen auf eine erneute

fachdidaktische Inhaltsanalyse der sozioökonomischen Problemsituationen (vgl. Kapitel 8.2.2). Die psychometrischen Indikatoren wurden aus dem Projekt CoBALIT beigezogen: die klassische Item-Schwierigkeit (prozentuale Lösungshäufigkeit) und die klassische Item-Trennschärfe (Item-Test-Korrelation, r_{it}) sowie die probabilistische Item-Schwierigkeit (Itemparameter, δ) und der Item-Infit (wMNSQ). Die Evaluation des WBK-T1 ergab schliesslich für die Items drei Revisionskategorien: a) ausschliessen, b) unverändert übernehmen, c) modifiziert übernehmen (vgl. Tabelle 7-2, exemplarisch Anhang Tabelle A- 3). Die Kategorie „ausschliessen“ galt für Items, die sowohl die inhaltliche Relevanz als auch die psychometrische Qualität verfehlten. Mit „modifiziert übernehmen“ wurden Items kategorisiert, wenn sie inhaltlich relevant, aber fachdidaktisch und psychometrisch unbefriedigend waren sowie modifizierbar schienen.

In der dritten Phase der Testrevision wurden die zu modifizierenden Items sprachlich (Formulierung, z.B. Fremdwörter und Fachbegriffe), inhaltlich (z.B. Antwortkategorien) und formal (Antwortformat, Antwortstrukturierung, Kodierung) überarbeitet, wobei die Itemtypen des erweiterten Konstruktionsschemas als Referenzrahmen dienten (vgl. Tabelle 8-4).

Bei der vierten Phase wurden externe Experten für eine Validierung des Testinhalts hinzugezogen. Anhand eines Kriterienrasters gaben sie ihre Einschätzung zu allen unveränderten, modifizierten und neukonstruierten Items (vgl. Kapitel 9.1). Schliesslich fanden kognitive Interviews mit Schüler*innen statt, um weitere sprachliche, formale und inhaltliche Unstimmigkeiten aufzudecken (vgl. Kapitel 9.2).

Aus dieser umfassenden Testrevision entstand der WBK-T2 mit vier Problemsituationen und mit insgesamt 32 Items, davon wurden ca. 40 % (13 Items) unverändert aus dem WBK-T1 übernommen, ca. 60 % modifiziert (12 Items) und neukonstruiert (7 Items) (vgl. Tabelle 7-2). Zum Vergleich: Der WBK-T1 umfasst in der Version CH acht sozioökonomische Problemsituationen mit 63 Items, in der Version DE sechs Problemsituationen mit 46 Items. Für den WBK-T2 wurden also vier Problemsituationen (LAN, STEU, JUG, EUR) bzw. 31 Items ausgeschlossen.

Des Weiteren wurden für den WBK-T2 einige der Problemsituationen geringfügig umbenannt (vgl. Tabelle 7-3). Der Titel sollte sich an der spezifischen Problemsituation orientieren statt am allgemeinen Politikfeld und er sollte für die Version CH und Version DE identisch sein. So wurde bspw. „Energiepolitik“ im WBK-T1 zu „Energieversorgung“ im WBK-T2, „AHV“ (Version CH) bzw. „Rentenversicherung“ (Version DE) im WBK-T1 zu „Altersvorsorge“ im WBK-T2.

Tabelle 7-2. Testinhalte und Testlänge des WBK-T1 und WBK-T2

Problemsituation	Version	Pretest		WBK-T1 total	WBK-T2				
		neuk.	ausg.		total	ausg.	unv.	mod.	neuk.
Altersvorsorge (AHV)	CH/DE	12	0	12	9	4	5	3	1
Energieversorgung (ENE)	CH/DE	11	3	8	7	2	2	4	1
Staatsverschuldung (STA)	CH/DE	8	2	6	9	0	4	2	3
Managervergütungen (MAN)	CH/DE	8	2	6	7	1	2	3	2
<i>Subtotal</i>		39	7	32	32	7	13	12	7
Agrarhandel (LAN)	CH	8	0	8	---				
Steuerwettbewerb/ Finanzausgleich (STEU)	CH	9	0	9	---				
Jugendverschuldung (JUG)	CH/DE	10	3	7	---				
Eurokrise (EUR)	CH/DE	8	1	7	---				
<i>Subtotal</i>		35	4	31	0				
<i>Total</i>		74		63	32				

Hinweis: Pretest eingesetzt bei Pilotierung CoBALIT, WBK-T1 eingesetzt bei Erhebung CoBALIT, WBK-T2 eingesetzt bei Erhebung WBKgy.

ausg. = ausgeschlossen, unv. = unverändert, mod. = modifiziert, neuk. = neukonstruiert.

Tabelle 7-3. Benennung der Problemsituationen im WBK-T1 und WBK-T2

WBK-T1	WBK-T2
AHV bzw. Rentenversicherung (AHV)	Altersvorsorge (AHV)
Energiepolitik (ENE)	Energieversorgung (ENE)
Staatsverschuldung (STA)	Staatsverschuldung (STA)
Managerlöhne bzw. Managergehälter (MAN)	Managervergütungen (MAN)
Landwirtschaftspolitik (LAN)	[Agrarhandel (LAN)]
Steuerwettbewerb (STEU)	[Steuerwettbewerb/Finanzausgleich (STEU)]
Jugendverschuldung (JUG)	---
Eurokrise (EUR)	---

Die konkreten inhaltlichen und formellen Veränderungen der Items und Begründungen für diese Veränderungen werden exemplarisch anhand der Problemsituation Altersvorsorge (AHV) veranschaulicht (vgl. Anhang Tabelle A- 3). Der Ausschluss von Items ist entweder auf die inhaltliche Relevanz (Repräsentativität, Adäquanz) oder auf die psychometrische Qualität (Item-Kennwerte) zurückzuführen. Bei der Modifikation von Items war die inhaltliche Präzisierung (Frageformulierung, Antwortkategorien) und die formelle Gestaltung (Antwortformat, Antwortstrukturierung/Umfang, Antwortkodierung) ausschlaggebend.

8 Spezifikationen des WBK-T2

In diesem Kapitel wird der Leitfrage (6) nachgegangen (vgl. Kapitel 1.2):

(6) *Welche Spezifikationen müssen bei der Testentwicklung des WBK-T2 vorgenommen werden, um die sozioökonomische Facette der WBK repräsentativ und adäquat abzubilden?*

Das Rahmenmodell der wirtschaftsbürgerlichen Kompetenz (WBK-Modell) (vgl. Kapitel 4.2) bildet die theoretische Grundlage und den konzeptionellen Rahmen für die Beschreibung und Operationalisierung des Konstrukts sozioökonomische Facette der WBK (WBK-soek) sowie für die Entwicklung des WBK-T2. Inwiefern die Testwerte des WBK-T2 hinsichtlich des Konstrukts valide interpretiert werden können, wird in dieser Dissertation durch verschiedene evidenzbasierte Validierungsaspekte untersucht (vgl. Kapitel 9, 11 und 12). Eine übersichtliche Darstellung der Spezifikationen des WBK-T2 findet sich in der „Dokumentation des WBK-T2“ (Ackermann, 2018c) und ist auf Anfrage bei der Autorin erhältlich.

Dieses Kapitel beschreibt die Spezifikationen des WBK-T2 nach den *Standards for Educational and Psychological Testing*, kurz *Standards* (AERA et al., 2014). Zuerst wird der Testzweck beschrieben (Kapitel 8.1), danach werden die Spezifikationen zu Inhalt (Kapitel 8.2), Format (Kapitel 8.3), Kodierung (Kapitel 8.4) und Testadministration (Kapitel 8.5) dargestellt.

8.1 Testzweck

In diesem Kapitel wird das zu messende Konstrukt eingegrenzt (Kapitel 8.1.1), die beabsichtigte Testanwendung (Kapitel 8.1.2) und die vorgesehenen Testpersonen (Kapitel 8.1.3) beschrieben sowie Annahmen zum Konstrukt und zur Interpretation der Testwerte formuliert (Kapitel 8.1.4).

8.1.1 Eingrenzung des Konstrukts

Die Entwicklung und Revision des WBK-T2 erfolgte entlang des WBK-Modells (vgl. Kapitel 4.2) mit forschungspragmatischen Eingrenzungen auf der Strukturebene und Prozessebene. Auf der *Strukturebene* werden sozioökonomische Problemsituationen im gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereich (s3) operationalisiert (vgl. Kapitel 4.3.2). Dies wird im Folgenden als *sozioökonomische Facette der WBK* (WBK-soek) bezeichnet. Diese Eingrenzung der Strukturebene auf nur eine inhaltliche Domäne des WBK-Modells fusst auf konstruktionslogischen (gleiche bzw. ähnliche Anforderungssituationen in der Domäne) und testtheoretischen (Skalierung eines eindimensionalen Testmodells) Gründen. Die Anforderungssituationen zwischen den Lebensbereichen sind heterogen und lassen sich kaum

repräsentativ und adäquat in einem einzigen psychologischen Test integrieren und als inhaltsvalide interpretieren.

Aus der Vielzahl sozioökonomischer Problemsituationen im gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereich (vgl. Kapitel 4.3.2) wurden für die Testentwicklung bzw. Testrevison schliesslich diejenigen ausgewählt, die basierend auf fünf Kriterien das Konstrukt in seiner Breite und Tiefe am besten abzubilden vermögen (vgl. Kapitel 8.2.1). Die Fokussierung auf vier Problemsituationen (vgl. Tabelle 7-2) innerhalb des gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereichs hat forschungsökonomische und testökonomische Gründe (personelle und finanzielle Ressourcen, zumutbare Testlänge).

Auf der Prozessebene werden von der Informationsverarbeitung alle drei Stufen (k1) „wiedererkennen/wiedergeben“, (k2) „verstehen/anwenden“ und (k3) „vergleichen/beurteilen/entscheiden“ operationalisiert (vgl. Kapitel 4.4.1). Vom Problemlösen werden die Phasen (p2) „Problem multiperspektiv analysieren“ und (p3) „alternative Lösungen beurteilen“ operationalisiert (vgl. Kapitel 4.4.2). Die vorgelagerte Phase (p1) wird aus konstruktionslogischen und testökonomischen Gründen (Komplexitätsreduktion, Testdauer) im Einleitungstext und/oder im Itemstamm durch Informationen teilweise vorgegeben und vorstrukturiert, d.h. das „Problem“ benannt, einige bekannte Lösungsansätze skizziert wie auch Ziel- und Interessenskonflikte angedeutet. Die nachgelagerte Phase (p4) wird aus domänenlogischen Gründen ausgeklammert; diese Phase passiert in Realität zeitlich versetzt und wird durch die Behörden vollzogen. Es geht also bei der sozioökonomischen Facette der WBK (WBK-soek) um die abstrakte und fiktive Bearbeitung relevanter, komplexer und kontroverser sozioökonomischer Problemsituationen.

8.1.2 Testanwendung

Mit dem WBK-T2 wird ein *diagnostischer* Zweck verfolgt: Die Testwerte sollen als Kompetenzausprägung, also als Ausprägung in der sozioökonomischen Facette der WBK (WBK-soek) der Testpersonen interpretiert werden. Inwiefern die Testwerte auch einen prognostischen Zweck haben und als prädiktiv für andere Konstrukte und Merkmale betrachtet werden können, wurde bislang weder theoretisch fundiert noch empirisch untersucht.

Die Interpretation der Testwerte soll *kriteriumsorientiert* erfolgen, wobei als Kriterium der adjustierte Testwert bzw. die Lösungsquote dient. Eine Lösungsquote von 50 % wird als *mimimal standard* interpretiert (vgl. Döring & Bortz, 2016, S. 433; Eberle et al., 2008, S. 383); unter 50 % gilt als geringe WBK-soek (*low competence level*), zwischen 50 % und 70 % als mittlere WBK-soek (*moderate competence level*) und über 70 % als hohe WBK-soek (*high competence level*).

Eine kriteriumsorientierte Interpretation der Personenfähigkeiten anhand eines Kompetenzniveaumodells wird erstmals in dieser Dissertation versucht (vgl. Kapitel 11.6). Eine

normorientierte Interpretation der Testwerte gemäss des Nebengütekriteriums Normierung (vgl. Kapitel 5.1.2.d) ist bislang nicht möglich, da der WBK-T2 noch nicht an einer genügend grossen und repräsentativen Stichprobe (Eichstichprobe) getestet wurde und (noch) keine Normwerte für die Population vorliegen (vgl. Döring & Bortz, 2016, S. 432f.).

Der WBK-T2 richtet sich an Forschende der empirischen Bildungsforschung, insbesondere der Wirtschaftspädagogik/-didaktik. Er kann für vielfältige Forschungsdesigns verwendet werden: bei Querschnittstudien um den Kompetenzstand der Testpersonen zu vergleichen (Schumann et al., 2017; Siegfried & Ackermann, under review), bei Interventions- bzw. Längsschnittstudien um die Kompetenzentwicklung der Testpersonen zu verfolgen (Siegfried & Ackermann, in progress).

8.1.3 Testpersonen: Zielgruppe und Zielsubgruppen

Der WBK-T2 ist für Jugendliche am Ende der Sekundarstufe II konzipiert. Die Altersgruppe der ca. 18-jährigen erscheint aus entwicklungspsychologischer und bildungsbiografischer Perspektive sinnvoll. Dies einerseits, weil die sozioökonomischen Problemsituationen des WBK-T2 (junge) Menschen in ihrer Rolle als Staatsbürger*innen betreffen und andererseits, weil für die Bearbeitung dieser Problemsituationen ein ökonomisches und politisches Orientierungswissen erforderlich ist.

Als Zielgruppe des WBK-T2 sind Schüler*innen der folgenden Schulformen/Bildungsgänge vorgesehen: kaufmännische Berufsfachschulen, Berufsmaturitätsschulen im Schwerpunkt WuR und gymnasiale Maturitätsschulen im Schwerpunkt WuR. Von diesen Schüler*innen wird angenommen, dass sie über curricular erworbenes domänenspezifisches Wissen (ökonomische und politische Konzepte) und über domänenspezifisches Können (Problemlösemethoden) verfügen. Als Kontrastgruppe fungieren Schüler*innen der folgenden Schulformen/Ausbildungsgänge: nicht-kaufmännische Berufsfachschulen (z.B. technisch, gewerblich, gestalterisch, gesundheitlich, sozial), Berufsmaturitätsschulen mit anderem Schwerpunktfach und gymnasiale Maturitätsschulen mit anderem Schwerpunktfach. Diese Schüler*innen verfügen vermutlich über geringeres curricular erworbenes domänenspezifisches Wissen und Können (vgl. Eberle, 2006b, S. 21). Anwendet wurde der WBK-T1 bislang erst bei kaufmännischen Lernenden in zwei ausgewählten Branchen (Holtsch & Eberle, 2018; Siegfried & Ackermann, in progress) und der WBK-T2 bei Gymnasiast*innen in einem ausgewählten Kanton bzw. Bundesland (Eberle et al., 2016; Schumann et al., 2017).

Der WBK-T2 scheint *prima vista* für die Sekundarstufe I nicht geeignet. Schüler*innen auf dieser Stufe fehlt es erstens am ökonomischen und politischen Orientierungswissen, das für die Bearbeitung der sozioökonomischen Problemsituationen notwendig ist. Zweitens mangelt es an der persönlichen Reife, sich aktiv mit wirtschafts- und sozialpolitischen Problemstellungen auseinandersetzen zu wollen und zu können. Für die Sekundarstufe I scheint es hingegen sinnvoller und zielführender zu sein, die persönlich-finanzielle Facette der WBK (vgl. Kapitel 4.3.1) mittels anderer bereits bestehender Leistungstests zu untersuchen (z.B. OECD, 2012, 2014a, 2017b;

Rudeloff, 2019). Dies auch in Anbetracht des in der Schweiz neu eingeführten Lehrplans 21 mit dem Kompetenzbereich „Wirtschaft, Arbeit, Haushalt“ (WAH), in welchem sich u.a. Themenaspekte zum Umgang mit Geld (WAH.2) und zur Gestaltung des Konsums (WAH.3) finden (D-EDK, 2014).

Inwiefern sich der WBK-T2 für Erwachsene eignet, ist derzeit noch ungeklärt. Diese Frage müsste in einer weiteren Validierungsstudie mit Testpersonen, z.B. aus der Erwachsenen-Maturitätsschule und der Lehrer*innen-Bildung für Sekundarstufe I und II, untersucht werden.

8.1.4 Annahmen zum Konstrukt und zur Interpretation der Testwerte

Für den WBK-T2 und für die Interpretation seiner Testwerte hinsichtlich des Konstrukts WBK-soek werden nachfolgend eine Reihe von Annahmen formuliert, die es im Rahmen der evidenzbasierten Validierung zu untersuchen und überprüfen gilt (vgl. Kapitel 9, 11 und 12). Diese Annahmen werden aus den *Standards* zur Validierung (AERA et al., 2014) abgeleitet und beziehen sich auf den Testinhalt, die innere Struktur des Tests und die Beziehung der Testwerte zu anderen Merkmalen (vgl. Kapitel 6.1).

8.1.4.a Annahmen zum Testinhalt

Evidenz zum Inhalt des Tests kann auf zwei Arten gesammelt werden (AERA et al., 2014, S. 14f.) (vgl. Kapitel 6.1.2): zum einen durch eine detaillierte Beschreibung des Konstrukts und der Inhaltsspezifikationen des Tests; zum anderen durch die kriteriengeleitete Einschätzung des Testinhalts durch Experten.

Für den WBK-T2 wird ausgehend von der umfangreichen Domänenanalyse (vgl. Kapitel 4.3.2.c) eine detaillierte Inhaltsspezifikation vorgenommen (vgl. Kapitel 8.2). Die Beurteilung des Testinhalts des WBK-T2 erfolgt durch Experten-Interviews zu den beiden Kriterien Repräsentativität und Adäquanz. Für den Testinhalt des WBK-T2 werden folgende Annahmen i.S.v. Kriterien formuliert und weiter unten im Rahmen der Validierung untersucht (vgl. Kapitel 9):

- **Annahme I zur Repräsentativität:**

Die Items des WBK-T2 sind repräsentativ für die inhaltsbezogene und kognitionsbezogene Domäne der WBK-soek.

- **Annahme II zur Adäquanz:**

Die Items des WBK-T2 sind adäquat für die Zielgruppe, weniger adäquat für die Kontrastgruppe.

8.1.4.b Annahmen zur internen Struktur

Um Evidenz zur internen Struktur des Tests zu sammeln, soll untersucht werden, inwiefern die empirische Beziehungen der Testitems zum Test bzw. zu den Testkomponenten mit dem theoretischen Konstrukt konform sind (AERA et al., 2014, S. 16) (vgl. Kapitel 6.1.2). Wird für den Test

eine eindimensionale faktorielle Struktur angenommen, müssen alle manifesten Indikatoren (Items) auf *einen* latenten Faktor laden, d.h. die Items des Tests werden durch *ein* Konstrukt repräsentiert (AERA et al., 2014; Cronbach, 1988) (zur Dimensionalität vgl. Kapitel 5.2.1.b; zur Eindimensionalität vgl. Kapitel 5.2.3.b). Wird hingegen für den Test eine mehrdimensionale faktorielle Struktur unterstellt, z.B. weil er aus mehreren Subtests besteht, so müssen die manifesten Indikatoren (Items) auf *mehrere* latente Faktoren und die inhaltlich zusammengehörigen Indikatoren auf *denselben* Faktor laden. Das bedeutet die Items eines Subtests repräsentieren jeweils ein Konstrukt (AERA et al., 2014; Cronbach, 1988).

Für das Konstrukt WBK-soek bzw. für den WBK-T2 wird aufgrund der theoretischen Modellierung (vgl. Kapitel 4.2), der testtheoretischen Überlegungen (vgl. Kapitel 5.2) und der Testkonstruktion (vgl. Kapitel 8.2 und 8.3) eine eindimensionale Struktur postuliert. Insbesondere sollen die Inhaltsbereiche (Problemsituationen), die Kognitionsstufen und die Antwortformate des WBK-T2 nicht als Facetten des Konstrukts WBK-soek betrachtet werden. In den vorangegangenen Studien mit dem WBK-T1 wurde eine zweidimensionale Struktur angenommen und gefunden (Eberle et al., 2016; Holtsch & Eberle, 2018). Diesbezüglich sind zwei Einwände vorzubringen: Zum einen wurden für Dimensionsanalyse auch „fremde“ Items aus dem WBT und aus dem OEKOMA-Test einbezogen (vgl. Kapitel 3.4 und 7.1.3), zum anderen wurden die Items in Anlehnung an das Modell der Wirtschaftskompetenz einer domänenverbundenen und domänenspezifischen Kompetenz zugeordnet (vgl. Eberle, 2015).

Wenn der Test basierend auf der IRT mit einem Rasch-Modell skaliert wird, müssen die Annahmen des Testmodells bezüglich der Item-Homogenität erfüllt sein (vgl. Kapitel 5.2.3.b). Für den WBK-T2 wird aufgrund seiner Kombination aus dichotomen und ordinalen Antwortkategorien ein Partial-Credit-Rasch-Modell spezifiziert und die Parameter des Modells mittels MML geschätzt. Die Schätzgenauigkeit der einzelnen Parameter kann mittels der Testinformation bzw. der Standardfehler beurteilt werden, die Messgenauigkeit der globalen Parameter und Testwerte mit den Test-Reliabilitäts-Koeffizienten.

Für die interne Struktur des WBK-T2 werden folgende Annahmen i.S.v. *Anforderungen* formuliert und weiter unten im Rahmen der Validierung untersucht (vgl. Kapitel 11):

- *Annahme III zur faktoriellen Struktur bzw. Dimensionalität des Konstrukts (vgl. Kapitel 11.1):
Der WBK-T2 zeigt eine eindimensionale faktorielle Struktur (Test-Eindimensionalität).*
- *Annahme IV zur Schätzung und Skalierung der Parameter (vgl. Kapitel 11.2):
Die Personenparameter der Stichprobe WBKgym und die Itemparameter des WBK-T2 lassen sich konsistent schätzen und auf einer kontinuierlichen Logit-Skala skalieren.*

- *Annahme V zur Messgenauigkeit des Tests (vgl. Kapitel 11.3):*
Die Personenparameter und Personen-Testwerte lassen sich mit dem WBK-T2 genügend genau schätzen (Test-Reliabilität).
- *Annahme VI zur Gültigkeit des Testmodells (vgl. Kapitel 11.4):*
Die geschätzten Itemparameter des WBK-T2 erfüllen die Annahmen des Partial-Credit-Rasch-Modells (Item-Homogenität).

Um Evidenz zur psychometrischen Qualität eines Tests und seiner Items zu sammeln, sollen die in den Testspezifikationen vorgegebenen psychometrischen Eigenschaften mittels klassischer und probabilistischer Itemanalysen ermittelt und überprüft werden. Zu den Kennwerten der klassischen Itemanalyse gehören u.a. die Testwerte-Verteilung, die Item-Schwierigkeiten und die Item-Trennschärfe (AERA et al., 2014, S. 37ff., S. 79) (vgl. Kapitel 5.1.2.b und 6.2.1). Für die psychometrische Qualität des WBK-T2 werden folgende Annahmen i.S.v. *Anforderungen* formuliert und weiter unten im Rahmen der Validierung untersucht (vgl. Kapitel 11):

- *Annahme VII zur Verteilung der Testwerte (vgl. Kapitel 11.5):*
Die Testwerte des WBK-T2 zeigen eine Normalverteilung.
- *Annahme VIII zur Qualität der Items (vgl. Kapitel 11.6):*
Die Items des WBK-T2 decken den mittleren Schwierigkeitsbereich ab (klassische Item-Schwierigkeit). Die Items des WBK-T2 unterscheiden zwischen Testpersonen mit hohen und geringen Testwerten (klassische Item-Trennschärfe).

Um Evidenz zur Graduierung eines Tests zu sammeln, soll die Logit-Skala in Abschnitte unterteilt und jeder Skalenabschnitt anhand der in ihm lokalisierten Items beschrieben werden (Hartig, 2007). Auf diese Weise kann ein Kompetenzniveaumodell etabliert werden (vgl. Kapitel 3.3). Für die Graduierung des WBK-T2 wird folgende Annahme formuliert und weiter unten im Rahmen der Validierung untersucht (vgl. Kapitel 11):

- *Annahme IX zur Graduierung des Tests (vgl. Kapitel 11.6):*
Für den WBK-T2 lassen sich anhand von schwierigkeitsbestimmenden Merkmalen und den probabilistischen Item-Schwierigkeiten Kompetenzschwellen quantitativ festlegen und Kompetenzniveaus qualitativ beschreiben.

8.1.4.c Annahmen zur Beziehung mit externen Merkmalen

Evidenz zur Beziehung der Testwerte mit externen Merkmalen kann durch zwei verschiedene Ansätze gesammelt werden (AERA et al., 2014, S. 16ff.) (vgl. Kapitel 6.1.2): Zum einen durch die Beziehung zwischen Testwerten und einem manifesten Aussenkriterium (sog. kriteriale Validität),

zum anderen durch die Beziehung zwischen Testwerten und anderen Konstrukten (sog. konvergente und divergente Validität).

Für die Beziehung der WBK-soek bzw. WBK-T2 zu externen Merkmalen werden Annahmen i.S.v. statistisch prüfbar *Hypothesen* formuliert und weiter unten im Rahmen der Validierung untersucht (Kapitel 12). Für die *kriteriale Validität* der WBK-soek werden verschiedene Annahmen zu Zielgruppe (WuR) und Kontrastgruppe (Nicht-WuR) sowie zu verschiedenen Zielsubgruppen (biologisches Geschlecht, Erstsprache, Staatsbürgerschaft, Schulform/Bildungsgang, Politiksystem/Land) getroffen. Für die *divergenten Validität* werden Annahmen zur Beziehung der WBK-soek zu anderen Leistungsmerkmalen und zu Interessensmerkmalen formuliert.

Das curriculare Vorwissen, das sich formell aus der Stundentafel des Faches WuR an gymnasialen Maturitätsschulen ableiten lässt, liefert Hinweise für Unterschiede in der WBK-soek (EDK, 1994; St.Gallen, 2006). Schüler*innen mit „Wirtschaft & Recht“ im Schwerpunktfach (Zielgruppe WuR) haben eine höhere Stundentafel und breitere/tiefere Lerninhalte als Schüler*innen mit einem anderen Schwerpunktfach bzw. mit „Wirtschaft & Recht“ im Einführungsfach (Kontrastgruppe Gruppe Nicht-WuR). Die Zielgruppe WuR bringt daher mehr curriculares Vorwissen mit und ist im WBK-T2 vermutlich begünstigt gegenüber der Kontrastgruppe Nicht-WuR. Jedoch bezieht sich das curriculare Vorwissen der Gymnasiast*innen hauptsächlich auf *fachbezogenes* Konzeptwissen (deklaratives Wissen), nicht unbedingt auf *sachbezogenes* Handlungswissen (Problemlösen i.S.v. Lösungsvorschläge für sozioökonomische Problemsituationen vergleichen und beurteilen). Frühere empirische Studien zu ökonomischen Kompetenzen weisen auf einen grossen positiven Effekt des curricularen Vorwissens in Form eines wirtschaftlichen Schwerpunktfaches oder eines kaufmännischen Ausbildungsprofils hin (vgl. K. Beck, 1993; Schumann & Eberle, 2014a; Schumann, Eberle, & Oepke, 2013).⁶⁶

- *Annahme X zum curricularen Vorwissen bzw. Schwerpunktfach (vgl. Kapitel 12.1):*
*Gymnasiast*innen im Schwerpunktfach WuR (WuR) verfügen über eine höhere WBK-soek als Gymnasiast*innen in anderen Schwerpunktfächern (Nicht-WuR).*

Bei der Modellierung der WBK bzw. der WBK-soek und bei der Spezifizierung des WBK-T2 wurden keine Unterschiede zwischen Mädchen und Jungen intendiert. Bei den Spezifikationen des WBK-T2 wurde sorgfältig darauf geachtet, dass die Problemsituationen alle „stereotypen“ Geschlechter ansprechen (z.B. durch sozial-, umwelt- und finanzpolitische Themen), dass die Einleitungstexte und Itemstämme *gender-neutral* formuliert sind, und dass die Darstellungsformen

⁶⁶ Für Schweizer Schüler*innen an gymnasialen Maturitätsschulen: grosser Effekt des Schwerpunktfachs WuR auf ökonomische Kompetenz (Cohens $|d| = 0.89$). Für Schweizer Schüler*innen an Berufsmaturitätsschulen: grosser Effekt des kaufmännischen Profils auf ökonomische Kompetenz (Cohens $|d| = 0.70$) (Schumann & Eberle, 2014a; Schumann et al., 2013).

und Antwortformate der Items derart variieren, damit kein biologisches Geschlecht begünstigt/benachteiligt ist (vgl. Kapitel 11.4.3.b und 12.2.1). Entsprechend sollte sich die WBK-soek zwischen den Geschlechtern nicht unterscheiden und der WBK-T2 keinen *gender gap* zeigen. Viele empirische Studien liefern jedoch Hinweise für einen geringen Geschlechtereffekt zugunsten der Jungen bei ökonomischen Kompetenzen von Jugendlichen (vgl. K. Beck, 1993; Brückner, Förster, Zlatkin-Troitschanskaia, & Walstad, 2015; Förster & Happ, under review; Rudeloff, Brahm, & Pumptow, under review; Schumann & Eberle, 2014a; Schumann et al., 2013; Walstad, Rebeck, & Butters, 2013a).⁶⁷ Einige wenige Studien können diesen Effekt jedoch nur für ihre Teilstichproben, wie z.B. einzelne Länder und einzelne kaufmännische Branchen, bestätigen (Schumann et al., 2017).⁶⁸

- *Annahme XI zum biologisches Geschlecht (vgl. Kapitel 12.2.1):*

Mädchen und Jungen zeigen keine substanziellen Unterschiede in der WBK-soek.

In der Beschreibung der inhalts- und kognitionsbezogenen Domäne der WBK-soek (vgl. Kapitel 4.3.2 und 4.4) finden sich Hinweise dafür, dass sprachliche Fähigkeiten und politisches Interesse für die WBK-soek förderlich sind. Deutsch als Erstsprache begünstigt das sprachliche Verständnis (i.S.v. Textinterpretation, Textproduktion) von sozioökonomischen Problemstellungen, das sowohl im WBK-T2 mit seinen textbasierten Items als auch im Alltag für die Lektüre von Medienberichten zentral ist. In einer früheren Studie wurde ein geringer positiver Effekt der Familiensprache auf ökonomische Kompetenzen nachgewiesen (vgl. Jüttler & Schumann, 2016).⁶⁹ Die Schweizer Staatsbürgerschaft erzeugt persönliche Betroffenheit und Interessiertheit für sozioökonomische Problemstellungen, da sich Schweizer Staatsbürger*innen regelmässig damit befassen „müssen“.

- *Annahme XII zur Erstsprache (vgl. Kapitel 12.2.2):*

*Gymnasiast*innen mit Deutsch als Erstsprache haben eine höhere WBK-soek als jene mit Zweitsprache Deutsch.*

⁶⁷ Für Schweizer Schüler*innen an gymnasialen Maturitätsschulen und Berufsmaturitätsschulen: geringer Effekt des biologischen Geschlechts auf ökonomische Kompetenz (Cohens $|d| = 0.44$) (Schumann & Eberle, 2014a; Schumann et al., 2013).

⁶⁸ Für Schweizer kaufmännische Lernende der Branchen SWISSMEM und SPEDLOGSWISS: kein Effekt des biologischen Geschlechts auf wirtschaftsbürgerliche Kompetenz (Cohens $|d| = 0.01$, $p > .05$). Für deutsche kaufmännische Lernende der Berufe Industriekaufleute und Speditionskaufleute: geringer Effekt des biologischen Geschlechts auf wirtschaftsbürgerliche Kompetenz (Cohens $|d| = 0.19$, $p < .05$) (Schumann et al., 2017).

Für kaufmännische Lernende der Branche SWISSMEM bzw. des Berufs Industriekauffrau/-mann: kein bzw. geringer Effekt des biologischen Geschlechts auf wirtschaftsbürgerliche Kompetenz (Cohens $|d| = 0.13$, $p > .05$ bzw. Cohens $|d| = 0.28$, $p < .01$). Für kaufmännische Lernende der Branche SPEDLOGSWISS bzw. des Berufs Speditionskauffrau/-mann: kein bzw. geringer Effekt des biologischen Geschlechts auf wirtschaftsbürgerliche Kompetenz (Cohens $|d| = 0.11$, $p > .05$ bzw. Cohens $|d| = 0.38$, $p < .05$) (Schumann et al., 2017).

⁶⁹ Für Schweizer Schüler*innen an gymnasialen Maturitätsschulen und Berufsmaturitätsschulen: geringer Effekt der Familiensprache Schweizerdeutsch/Deutsch auf ökonomische Kompetenz (Cohens $|d| = 0.19$, $p < .05$) (Jüttler & Schumann, 2016).

▪ *Annahme XIII zur Staatsbürgerschaft (vgl. Kapitel 12.2.2):*

*Gymnasiast*innen mit Schweizer Staatsbürgerschaft haben eine höhere WBK-soek als jene mit anderer Staatsbürgerschaft.*

Für den Übertritt von der Sekundarstufe I (8. oder 9. Klasse) ans Gymnasium müssen z.B. im Kanton St.Gallen schriftliche Aufnahmeprüfungen von einer bis drei Stunden in den Fächern Deutsch, Französisch und Mathematik absolviert werden. Kandidat*innen mit einem ungenügenden Prüfungsergebnis werden zusätzlich für mündliche Prüfungen in den Fächern Deutsch und Französisch aufgeboten. Die definitive Aufnahme ans Gymnasium erfolgt nach einer Probezeit von einem Semester (Aufnahmereglement der Mittelschule vom 25. Juni 2011, Kanton St.Gallen). Für den Beginn einer (kaufmännischen) Berufsausbildung ist hingegen weder eine Aufnahmeprüfung noch ein Eignungstest erforderlich. Allerdings stellen die Lehrbetriebe branchen- oder betriebsspezifische Anforderungen an die Lernenden, die Organisationen der Arbeitswelt (OdA) regeln die Bildungsinhalte und das Qualifikationsverfahren für den Abschluss Kauffrau/Kaufmann EFZ (BBG; SKKAB, 2011a).

Aufgrund dieser Selektionsmechanismen des Schweizer Bildungssystems ist davon auszugehen, dass Gymnasiast*innen über höhere allgemeine kognitive Fähigkeiten verfügen als kaufmännische Berufslernende und sie – auch bei einer geringeren Stundentafel im Fach WuR bzw. WuG – mit sozioökonomischen Problemsituationen souveräner umzugehen wissen. Hinweise dafür ergeben sich auch aus einer früheren Studie, wo für Gymnasiast*innen höhere kognitive Fähigkeiten berichtet wurden als für Berufsmaturand*innen (Schumann & Eberle, 2014b) und ein positiver, wenn auch geringer Zusammenhang zwischen allgemeinen kognitiven Fähigkeiten und ökonomischen Kompetenzen festgestellt wurde (Schumann et al., 2013).⁷⁰ Diese Befunde bestätigen die Annahme, dass «das Gymnasium [...] überwiegend von jenen Jugendlichen eines Jahrgangs besucht [wird], die über die höchsten allgemeinen kognitiven Fähigkeiten verfügen, auch wenn sich die Häufigkeitsverteilungen der parallelen Schulniveaustufen überlappen» (Eberle & Brüggenbrock, 2013, S. 102). Die allgemeinen kognitiven Fähigkeiten gehören zu den stabilen Bedingungsfaktoren von Schulleistungen, die sich durch veränderbare Bedingungsfaktoren (z.B. Leistungsmotivation, Lerngelegenheiten, Unterrichtsqualität) nur teilweise kompensieren lassen (ebd., S. 102).

▪ *Annahme XIV zur Schulform/Bildungsgang (vgl. Kapitel 12.2.3):*

*Schüler*innen an gymnasialen Maturitätsschulen verfügen über eine höhere WBK-soek als jene an kaufmännischer Berufsfachschulen.*

⁷⁰ Bei Schweizer Gymnasiast*innen: $M = 519$ ($SD = 90$), bei Schweizer Berufsmaturand*innen: $M = 461$ ($SD = 98$); standardisierte Testwerte von $M = 500$ und $SD = 100$ (Schumann & Eberle, 2014b, S. 224f.). Für Schweizer Schüler*innen an gymnasialen Maturitätsschulen und Berufsmaturitätsschulen: geringe positive Korrelation zwischen ökonomischer Kompetenz und kognitiver Grundfähigkeit ($r = .24$, $p < .01$) (Schumann & Eberle, 2014a, S. 116f.).

Aufgrund des politischen Systems der Schweiz, das Staatsbürger*innen mit direktdemokratischen Rechten ausstattet, werden Schweizer Schüler*innen intensiver in politische Debatten und Entscheidungen zu sozioökonomischen Problemstellungen eingebunden als deutsche Schüler*innen. In einer früheren Studie wurde bei kaufmännischen Lernenden ein geringer Effekt des Landes auf die wirtschaftsbürgerliche Kompetenz festgestellt, wenn für die formale Vorbildung (Matura bzw. Abitur) kontrolliert wird (Schumann et al., 2017).⁷¹

▪ *Annahme XV zum Politiksystem/Land (vgl. Kapitel 12.2.4):*

*Schweizer Schüler*innen auf der Sekundarstufe II, insbesondere an gymnasialen Maturitätsschulen und an kaufmännischen Berufsfachschulen, haben eine höhere WBK-soek als deutsche Schüler*innen vergleichbarer Schulformen.*

Ausgehend vom WBK-Modell (vgl. Kapitel 4.2) kann ein struktureller Zusammenhang zwischen der WBK-soek sowie anderen kognitiven Leistungsmerkmalen unterstellt werden: Die sozioökonomischen Problemsituationen im WBK-T2 erfordern domänenspezifisches Fakten- und Konzeptwissen (v.a. ökonomische und politische Konzepte) sowie sprachliche Fähigkeiten (i.S.v. Textinterpretation und Textproduktion) (vgl. Kapitel 8.3). Allerdings sind die Schulleistungen im Fach WuR nur bedingt prädiktiv für die WBK-soek, da der WBK-T2 nicht curricular konzipiert ist (vgl. Kapitel 8.2): Das Curriculum des Faches WuR ist inhaltlich viel breiter als die WBK-soek, d.h. es vermittelt Wissen und Können in den Fachbereichen VWL, BWL (inkl. FRW) und Recht. Auch die Schulleistungen im Fach Deutsch sind nur bedingt prädiktiv für die WBK-soek: Das Curriculum des Faches Deutsch ist inhaltlich viel breiter (Orthografie, Grammatik, Textinterpretation, Textproduktion, Literatur etc.), als es die WBK-soek erfordert, d.h. Informationen recherchieren, selektieren, interpretieren. In früheren empirischen Studien aus der Schweiz wird ein geringer positiver Zusammenhang der ökonomischen Kompetenz bzw. wirtschaftsbürgerlichen Kompetenz mit Deutschfähigkeiten und Mathematikfähigkeiten berichtet (vgl. Eberle, Schumann, Ackermann, Jüttler, & Kaufmann, 2017a; Eberle, Schumann, Ackermann, Jüttler, & Kaufmann, 2017b; Schumann & Eberle, 2014a, 2014b).⁷² Für das Projekt WBKgym mussten aus forschungsökonomischen Gründen

⁷¹ Für kaufmännische Lernende der Branche SWISSMEM bzw. des Berufs Industriekaufrau/-mann: geringer Effekt des Landes auf wirtschaftsbürgerliche Kompetenz (ohne Kontrolle Vorbildung: Cohens $|d| = 0.13$, $p > .05$; mit Kontrolle Vorbildung: Cohens $|d| = 0.28$, $p < .01$). Für kaufmännische Lernende der Branche SPEDLOGSWISS bzw. des Berufs Speditionskauffrau/-mann: geringer Effekt des Landes auf wirtschaftsbürgerliche Kompetenz (ohne Kontrolle Vorbildung: Cohens $|d| = 0.11$, $p > .05$; mit Kontrolle Vorbildung: Cohens $|d| = 0.38$, $p < .05$) (Schumann et al., 2017).

⁷² Für Schweizer Schüler*innen an gymnasialen Maturitätsschulen und Berufsmaturitätsschulen: geringe positive Korrelation zwischen ökonomischer Kompetenz und Deutschfähigkeit ($r = .24$, $p < .01$), geringe Korrelation zwischen ökonomische Kompetenz und Mathematikfähigkeit ($r = .16$, $p < .01$) (Schumann & Eberle, 2014a). Für Schweizer kaufmännische Lernende: geringe positive Korrelation zwischen wirtschaftsbürgerlicher Kompetenz und Deutschfähigkeit (SWISSMEM: $r = .26$, $p < .01$; SPEDLOGWISS: $r = .31$, $p < .01$), moderate positive Korrelation zwischen wirtschaftsbürgerlicher Kompetenz und Mathematikfähigkeit (SWISSMEM: $r = .31$, $p < .01$; SPEDLOGWISS: $r = .38$, $p < .01$) (Eberle et al., 2017a, 2017b). Für Schweizer kaufmännische Lernende: hohe positive Prädiktion der

die Schulnoten der Fächer WuR und Deutsch als manifeste Proxy für das fachliche Orientierungswissen zu „Wirtschaft“ und die Lesefähigkeit in der Erstsprache verwendet werden.

- *Annahme XVI zu Schulnoten (vgl. Kapitel 12.3):*

Die Testwerte des WBK-T2 haben einen positiven aber geringen Zusammenhang mit den Schulleistungen in den Fächern WuG und Deutsch.

Nach der pädagogischen Interessenstheorie ist Interesse als Relation zwischen Person und Gegenstand definiert (Prenzel, Krapp, & Schiefele, 1986, S. 165ff.). Der Gegenstand ist ein Umweltausschnitt, den die Person von anderen Ausschnitten abgegrenzt, strukturiert und anhand bestimmter Referenzobjekte rekonstruiert. Die Relation der Person zum Gegenstand kann aus zwei Perspektiven betrachtet werden: 1) die aktuelle Auseinandersetzung als zeit-/situationsspezifische Beziehung, oder 2) die andauernde Auseinandersetzung als zeit-/situationsübergreifender Bezug. Das Interesse bzw. die Person-Gegenstand-Relation wird durch drei strukturelle Merkmale bestimmt: den kognitiven Bereich (hohe gegenstandsbezogene Komplexität, differenzierte und variierende Auffassung des Gegenstands, differenziertes und integriertes Wissen über den Gegenstand), den emotionalen Bereich (anregendes und angenehmes gegenstandsbezogenes Erleben) und den Wertebereich (selbstintentionale Auseinandersetzung mit dem Gegenstand, identitätsstiftende Funktion des Gegenstands).

Im theoretischen Modell zur „Wirkungsweise von Interesse“ sind Kognition, Emotion Werte/Steuerung als unabhängige Variablen beschrieben, Persistenz (Aufrechterhaltung des Gegenstandsbezugs) und Selektivität (inhaltliche Ausrichtung des Gegenstandsbezugs) als abhängige (Prenzel et al., 1986, S. 169ff.). Als persistenzfördernd werden im Modell u.a. folgende Faktoren angenommen: bestehende, aber auflösbar erscheinende kognitive Diskrepanzen; Kompetenzzuwachs (bzgl. Wissensbeständen und Handlungsrepertoire); Spannung bzw. versunkenes Agieren (*flow*); Kompetenzgefühl (Freude über das Gelingen). Im theoretischen Modell zu „Interesse, Lernen und Leistung“ wird Interesse (individuelles oder situationales Interesse) als unabhängige Variable (Bedingung) und die bewertete Leistung (z.B. Testwerte, Schulnoten) als abhängige Variable (Effekt) modelliert (Krapp, 1992, S. 751f.). In einer Metaanalyse zu Fachinteresse und Schulleistungen wurde ein durchschnittlich geringer bis moderater Zusammenhang berichtet (Schiefele, Krapp, & Schreyer, 1993).⁷³ In einer früheren Studie wurde zwischen dem Fachinteresse und der ökonomischen Kompetenz ein positiver, aber geringer Zusammenhang festgestellt (Schumann &

Eingangsvoraussetzungen (Mathematikfähigkeit, Deutschfähigkeit und kognitive Fähigkeit) auf wirtschaftsbürgerliche Kompetenz (domänenverbunden: $\beta = .71$, $p < .001$; domänenspezifisch: $\beta = .47$, $p < .001$) (Sticca et al., 2018).

⁷³ Durchschnittliche Korrelation von $r = .30$.

Eberle, 2014a, 2014b).⁷⁴ Das Interesse an sozioökonomischen Problemsituationen ist ein domänenspezifisches Interesse, d.h. ein Interesse an der Sache, nicht unbedingt am Fach (WuR). Daher wird ein Zusammenhang zwischen der Leistung im WBK-T2 und dem Interesse an den sozioökonomischen Problemsituationen postuliert.

- *Annahme XVII zum Sachinteresse (vgl. Kapitel 12.3):*

Die Testwerte des WBK-T2 haben einen geringen positiven Zusammenhang mit dem Interesse an sozioökonomischen Problemsituationen.

8.2 Inhaltsspezifikationen

Dieses Kapitel widmet sich den Inhaltsspezifikationen des WBK-T2: Inhaltsbereiche des WBK-T2, d.h. die Auswahl der Problemsituationen (Kapitel 8.2.1) sowie Inhaltselemente für die Items der ausgewählten Problemsituationen im WBK-T2 (Kapitel 8.2.2). Die Inhaltsspezifikationen des WBK-T2 orientieren sich grösstenteils am WBK-T1, jedoch wurden einige inhaltliche Modifikationen vorgenommen, so z.B. bei Zusammensetzung der Problemsituationen und der Inhaltselemente für die Items.

8.2.1 Inhaltsbereiche des WBK-T2: Problemsituationen

Die Inhaltsbereiche des WBK-T2, d.h. die sozioökonomischen Problemsituationen wurden ausgehend von der Domänenanalyse für das Konstrukt sozioökonomische Facette der WBK (vgl. Kapitel 4.3.2.c, vgl. Anhang Tabelle A- 1) anhand fünf Kriterien ausgewählt. Diese systematische Inhaltsanalyse war aus mehreren Gründen notwendig: (1) Die Problemsituationen des WBK-T1 mussten hinsichtlich der enger ausgelegten Inhaltsdomäne des Konstrukts WBK-soek (vgl. Kapitel 8.1.1) überprüft werden; (2) die Problemsituationen des WBK-T1 mussten hinsichtlich der politischen Relevanz (Aktualität) im Schweizer Kontext neu geprüft werden (AERA et al., 2014, S. 83f.); und (3) neue Problemsituationen für den WBK-T2 sollten gefunden und hinsichtlich des Konstrukts WBK-soek geprüft werden.

8.2.1.a Kriterien

Die sozioökonomischen Problemsituationen für den WBK-T2 wurden nach folgenden Kriterien identifiziert und evaluiert (vgl. Eberle et al., 2016):

- **Repräsentativität (politische Relevanz):** Das Problem ist aktuell und authentisch und hat insofern realen Lebensweltbezug; das Problem ist im

⁷⁴ Für Schweizer Schüler*innen an gymnasialen Maturitätsschulen und Berufsmaturitätsschulen: geringe positive Korrelation zwischen ökonomischer Kompetenz (OEKOMA-Test) und Fachinteresse (Interesse am Wirtschaftsunterricht) ($r = .18, p < .01$) (Schumann & Eberle, 2014a).

gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereich situiert und ist insofern eine sozioökonomische Problemsituation.

- Wissenschaftlichkeit: Das Problem lässt sich mittels fachwissenschaftlicher Konzepte (z.B. theoretische Modelle, empirische Evidenz) bearbeiten, v.a. aus den Sozial- und Wirtschaftswissenschaften.
- Komplexität, Kontroversität und Multiperspektivität: Das Problem ist nicht-trivial und offen; das Problem lässt sich aus mehreren Perspektiven analysieren (z.B. individuell/kollektiv, kurzfristig/langfristig, lokal/global); für das Problem gibt es nicht-eindeutige, strittige Lösungen, die Lösungen erzeugen Ziel- und Interessenskonflikte bei den Akteuren und erfordert ihre Kompromissbereitschaft.
- Bekanntheit: Das Problem lässt sich mit Lernzielen/Lerninhalten aus den Bildungsplänen der Zielgruppe verknüpfen.
- Adaptierbarkeit: Das Problem ist übertragbar/anwendbar im deutschsprachigen Raum, insbesondere für Deutschland.

Das Kriterium „Repräsentativität (politische Relevanz)“ war schon für die Testentwicklung des WBK-T1 massgeblich (Eberle et al., 2016, S. 101f.). Allerdings wurde dieses Kriterium beim WBK-T1 nicht strikt auf den gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereich angewendet, da das dem WBK-T1 zugrunde gelegte Konstrukt, die WBK, weiter gefasst ist als das Konstrukt des WBK-T2, die WBK-soek (vgl. Kapitel 4.1 und 7.1). Für die Beurteilung der Repräsentativität der sozioökonomischen Problemsituationen für den WBK-T2 wurde auf die Domänenanalyse des gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereichs abgestützt. Für den WBK-T2 wird eine ausreichende Repräsentativität (politische Relevanz) angenommen, wenn (1) sich in einem Politikfeld tatsächlich sozioökonomische Problemsituationen finden und wenn (2) diese Problemsituationen in den letzten 20 Jahren durch mindestens drei Abstimmungsvorlagen die gesellschaftliche Debatte bestimmten (vgl. Anhang Tabelle A- 2).

Das Kriterium „Wissenschaftlichkeit“ wurde erst bei der Testentwicklung des WBK-T2 aufgenommen. Dies, weil bei der Kodierung des WBK-T1 festgestellt wurde, dass es für einige Problemsituationen – insbesondere STEU, JUG und EUR – zu wenig fachwissenschaftliche Konzepte bekannt waren, aufgrund derer man die der Schüler*innen-Antworten hätte eindeutig und einheitlich bewerten können. Eine solche Kodierung ist jedoch Voraussetzung für die Auswertungsobjektivität und somit für eine valide Interpretation der Testergebnisse (vgl. Kapitel 5.1.2.a). Die Wissenschaftlichkeit der sozioökonomischen Problemsituationen wurde anhand der Fachstruktur der Wirtschaftswissenschaft beurteilt. Dafür wurde auf ein differenziertes Kategoriensystem aus dem Projekt OEKOMA zurückgegriffen (Schumann et al., 2010). Das Kategoriensystem unterscheidet die beiden Fachbereiche Volkswirtschaftslehre (VWL) und

Betriebswirtschaftslehre (BWL); innerhalb jedes Fachbereichs sind Oberkategorien und Unterkategorien mit Schlüsselbegriffen festgelegt.⁷⁵ Für sozioökonomische Problemsituationen sind v.a. die VWL-Kategorien relevant. Eine ausreichende Wissenschaftlichkeit für den WBK-T2 wird angenommen, wenn die Problemsituation mindestens fünf Unterkategorien abdeckt (vgl. Anhang Tabelle A- 2). Damit wird die inhaltliche Verschränkung der sozioökonomischen Problemsituationen mit der Bezugsdisziplin Wirtschaftswissenschaft gewährleistet.

Das Kriterium „Komplexität, Kontroversität und Multiperspektivität“ war bereits für die Testentwicklung des WBK-T1 massgeblich (Eberle et al., 2016, S. 101f.). Die Problemsituationen sollten abstrakt und komplex bezüglich ihrer Ursachen und Wirkungen sein. Zudem sollten ihre Lösungsansätze kontrovers und multiperspektiv sein bezüglich individueller und kollektiver Interessenskonflikte sowie bezüglich kurzfristiger und langfristiger Zielkonflikte, also nicht zu einer richtigen Lösung führen, sondern das Abwägen von Vorteilen und Nachteilen verschiedener Lösungen erfordern (vgl. Eberle, 2006b, S. 18f.; 2015, S. 19).

Das Kriterium „Bekanntheit“ wurde erst bei der Testentwicklung des WBK-T2 aufgenommen. Es soll eine minimale Verknüpfung der Problemsituationen mit den Lerninhalten/Lernzielen aus den Lehrplänen der Zielgruppe geben, so dass der WBK-T2 – sofern im Rahmen dieser Dissertation ausreichend Evidenz für die Validität gefunden wird – für problemorientierte Interventionsstudien eingesetzt werden kann. Für die Beurteilung der Bekanntheit der sozioökonomischen Problemsituationen wurden diese mit drei Lehrplänen der Sekundarstufe II abgeglichen: gymnasiale Maturitätsschulen (Bern, 2005; EDK, 1994; St.Gallen, 2006), Berufsmaturitätsschulen (SBFI, 2012) und kaufmännische Berufsfachschulen (SKKAB, 2011a, 2011b).⁷⁶ Es wurde geprüft, ob und inwiefern sich die Problemsituationen mit Lerninhalten/Lernzielen des Fachs „Wirtschaft & Recht“ (WuR) bzw. „Wirtschaft & Gesellschaft“ (WuG) sowie der verwandten Fächer „Geschichte (inkl. Staatslehre)“ bzw. „Geschichte & Politik“, „Geografie“ sowie „Philosophie (inkl. Ethik)“ verknüpfen lassen. Beim Fach WuR bzw. WuG wurden die Teilgebiete Volkswirtschaftslehre (VWL), Betriebswirtschaftslehre (BWL) sowie Rechts- und Staatskunde (RECHT/STAAT) betrachtet. Allerdings wurden nur inhaltliche Elemente im Sinne von Schlüsselbegriffen berücksichtigt, nicht auch kognitive Elemente im Sinne

⁷⁵ Das Kategoriensystem umfasst für den Fachbereich VWL vier Oberkategorien: 1) Grundlagen, 2) Mikroökonomie, 3) Makroökonomie und 4) Internationale Wirtschaftsbeziehungen. Der Oberkategorie Mikroökonomie sind die drei Unterkategorien Angebot & Nachfrage, Wettbewerb/Marktformen, Rolle des Staates zugordnet. In der Oberkategorie Makroökonomie finden sich z.B. folgende Unterkategorien: Konjunktur, Inflation/Deflation, Arbeitsmarkt, Wirtschaftsentwicklung, Öffentliche Finanzen, Wirtschaftspolitik (Schumann et al., 2010, S. 10).

⁷⁶ Für die gymnasialen Maturitätsschulen der Rahmenlehrplan für Maturitätsschulen (Adams et al., 2015) und exemplarisch der Lehrplan für das Gymnasium im Kanton St. Gallen (EDK, 1994), für die Berufsmaturitätsschulen der Rahmenlehrplan für die Berufsmaturität (St.Gallen, 2006), für kaufmännische Berufsfachschulen der Bildungsplan und Lernzielkatalog Kauffrau/Kaufmann EFZ (SBFI, 2012).

einer vorgegebenen Taxonomie.⁷⁷ Für den WBK-T2 wird eine ausreichende Bekanntheit angenommen, wenn (1) die sozioökonomische Problemsituation durch mindestens zwei Lernziele/Lerninhalte aus der VWL abgedeckt wird, und wenn (2) mindestens ein Lernziel/Lerninhalt aus einem weiteren Teilgebiet (BWL, RECHT/STAAT) bzw. einem weiteren Fach (z.B. Geschichte, Geografie, Philosophie) abgedeckt wird (vgl. Anhang Tabelle A- 2).

Das Kriterium „Adaptierbarkeit“ wurde sowohl für die Testentwicklung des WBK-T1 (Eberle et al., 2016, S. 101f.) als auch für die Testrevision für den WBK-T2 einbezogen. Es ist vor dem Hintergrund internationaler Studien ausschlaggebend. Insbesondere wurde beim Forschungsprojekt CoBALIT mit dem WBK-T1 eine international vergleichende Studie mit kaufmännischen Lernenden in der Schweiz und in Deutschland durchgeführt (Eberle et al., 2016; Schumann et al., 2017). Dieses Kriterium ermöglicht denn auch weitere international vergleichende Studien, so z.B. ein Vergleich von Gymnasiast*innen in der Schweiz und Deutschland (Siegfried & Ackermann, under review).

Um eine erste Eingrenzung vorzunehmen, wurden die drei Kriterien „Repräsentativität (politische Relevanz)“, „Wissenschaftlichkeit“ und „Bekanntheit“ auf alle sozioökonomischen Problemsituationen angewendet, die mittels der Domänenanalyse identifiziert worden waren (vgl. Anhang Tabelle A- 1). Für eine weitere Eingrenzung wurden die zwei Kriterien „Komplexität & Kontroversität“ und „Adaptierbarkeit“ nur noch auf die sozioökonomischen Problemsituationen angewendet, die bereits die ersten drei Kriterien erfüllten (vgl. Anhang Tabelle A- 2). Allerdings wurden die Problemsituationen des WBK-T1 (vgl. Tabelle 7-2) nach allen fünf Kriterien beurteilt.

Schliesslich wurden folgende zehn Problemsituationen kriteriengeleitet beurteilt (vgl. Tabelle 8-1): Altersvorsorge (AHV), Arbeitslosigkeit, Energieversorgung (ENE), Agrarhandel (LAN), Managervergütungen (MAN), Staatsverschuldung (STA), Steuerwettbewerb/Finanzausgleich (STEU), Umwelt-/Klimaschutz, Jugendverschuldung (JUG), Eurokrise (EUR).

⁷⁷ Im Gegensatz zum kaufmännischen Bildungsplan geben die Lehrpläne der gymnasialen Maturitätsschulen keine anzustrebenden Kognitionsprozesse vor.

Tabelle 8-1. Beurteilung sozioökonomischer Problemsituationen für Testentwicklung und -revision

Sozioökonomische Problemsituation	Herkunft	Beurteilung nach Kriterien	Entscheidung
Altersvorsorge (AHV)	WBK-T1, Version CH, Version DE	▪ Repräsentativität (politischen Relevanz): erfüllt; div. Abstimmungsvorlagen zur AHV/IV, breite öffentliche Debatte. ▪ Wissenschaftlichkeit: erfüllt; 4 VWL-Kategorien, 1 BWL-Kategorie. ▪ Komplexität: erfüllt. ▪ Bekanntheit: erfüllt; Lerninhalte „Sozialversicherungen“ und „sozialer Ausgleich“. ▪ Adaptierbarkeit: teilweise erfüllt, aber institutionelle Besonderheiten berücksichtigen (z.B. AHV → gesetzliche Rentenversicherung).	repräsentativ und adäquat für Konstrukt operationalisierbar im WBK-T2 → belassen, Testrevision
Arbeitslosigkeit	neu	▪ Repräsentativität (politischen Relevanz): erfüllt; div. Abstimmungsvorlagen zur ALV, breite öffentliche Debatte. ▪ Wissenschaftlichkeit: erfüllt; 4 VWL-Kategorien, 1 BWL-Kategorie. ▪ Komplexität: erfüllt. ▪ Bekanntheit: erfüllt; Lerninhalte „Arbeitslosigkeit“, „Sozialversicherungen“ und „sozialer Ausgleich“. ▪ Adaptierbarkeit: teilweise erfüllt, aber institutionelle Besonderheiten berücksichtigen (z.B. ALV → Harz IV).	repräsentativ und adäquat für Konstrukt operationalisierbar im WBK-T2 → vorerst keine Testentwicklung
Energieversorgung (ENE)	WBK-T1, Version CH, Version DE	▪ Repräsentativität (politischen Relevanz): erfüllt; div. Abstimmungsvorlagen zu AKW-Ausstieg, breite öffentliche Debatte. ▪ Wissenschaftlichkeit: erfüllt; 5 VWL-Kategorien, 2 BWL-Kategorien. ▪ Komplexität: erfüllt. ▪ Bekanntheit: erfüllt; Lerninhalte „Wirtschaftspolitik“ und „Marktwirtschaft“. ▪ Adaptierbarkeit: erfüllt, aber institutionelle und faktische Besonderheiten berücksichtigen (z.B. CO₂-Gesetz, Energiestatistik).	repräsentativ und adäquat für Konstrukt operationalisierbar im WBK-T2 → belassen, Testrevision

Tabelle 8-1. Beurteilung sozioökonomischer Problemsituationen für Testentwicklung und -revision

Sozioökonomische Problemsituation	Herkunft	Beurteilung nach Kriterien	Entscheidung
Landwirtschaft (LAN) → Agrarhandel	WBK-T1, Version CH	▪ Repräsentativität (politischen Relevanz): teilweise erfüllt; wenige Abstimmungsvorlagen zu Agrarpolitik und Freihandelsabkommen, aber breite öffentliche Debatte. ▪ Wissenschaftlichkeit: erfüllt; 5 VWL-Kategorien. ▪ Komplexität: erfüllt. ▪ Bekanntheit: erfüllt; Lerninhalte „Globalisierung/Freihandel“ und „Wirtschaftspolitik“. ▪ Adaptierbarkeit: nicht erfüllt (EU-Politik im Agrarbereich).	repräsentativ und adäquat für Konstrukt nicht operationalisierbar im WBK-T2 (Adaptierbarkeit) → ausschliessen, keine Testrevision
Managervergütungen (MAN) → Gehaltsgerechtigkeit	WBK-T1, Version CH, Version DE	▪ Repräsentativität (politischen Relevanz): erfüllt; einige Abstimmungsvorlagen zu Managerlöhnen und Mindestlöhnen, aber breite öffentliche Debatte. ▪ Wissenschaftlichkeit: erfüllt; 4 VWL-Kategorien, 2 BWL-Kategorien. ▪ Komplexität: erfüllt. ▪ Bekanntheit: teilweise erfüllt; wenige Lerninhalte für Problemanalyse, unklare Norm für Kodierung. → erfüllt falls erweitert zu „Gehaltsgerechtigkeit“; Lerninhalte „Marktwirtschaft/Preisgrenzen“, „Personal/Lohnformen“ und „Arbeitsrecht“. ▪ Adaptierbarkeit: erfüllt, aber institutionelle und faktische Besonderheiten berücksichtigen (z.B. Arbeitsgesetz, Gehaltsstatistik).	repräsentativ und adäquat für Konstrukt operationalisierbar im WBK-T2 → belassen, Testrevision, ggf. inhaltlich ausweiten: Gehaltsgerechtigkeit
Staatsverschuldung (STA)	WBK-T1, Version CH, Version DE	▪ Repräsentativität (politischen Relevanz): erfüllt; einige Abstimmungsvorlagen zu Sparmassnahmen, aber breite öffentliche Debatte. ▪ Wissenschaftlichkeit: erfüllt; 5 VWL-Kategorien. ▪ Komplexität: erfüllt. ▪ Bekanntheit: erfüllt; Lerninhalte „Konjunkturpolitik“ und „Staatsverschuldung“. ▪ Adaptierbarkeit: erfüllt, aber faktische Besonderheiten berücksichtigen (z.B. Verschuldungsquote).	repräsentativ und adäquat für Konstrukt operationalisierbar im WBK-T2 → belassen, Testrevision

Tabelle 8-1. Beurteilung sozioökonomischer Problemsituationen für Testentwicklung und -revision

Sozioökonomische Problemsituation	Herkunft	Beurteilung nach Kriterien	Entscheidung
Steuerwettbewerb (STEU) → Steuerwettbewerb/ Finanzausgleich	WBK-T1, Version CH	▪ Repräsentativität (politischen Relevanz): erfüllt; einige Abstimmungsvorlagen zu Steuerharmonisierung und Finanzausgleich, aber breite öffentliche Debatte. ▪ Wissenschaftlichkeit: teilweise erfüllt; 3 VWL-Kategorien, 2 BWL-Kategorien. ▪ Komplexität: nicht erfüllt; von Experten als über-komplex beurteilt. ▪ Bekanntheit: teilweise erfüllt; wenige Lerninhalte für Problemanalyse, unklare Norm für Kodierung; Lerninhalte „Konjunkturpolitik“ und „Steuern“. ▪ Adaptierbarkeit: nicht erfüllt (nur Gewerbesteuer in Deutschland).	repräsentativ und adäquat für Konstrukt nicht operationalisierbar im WBK-T2 (Wissenschaftlichkeit, Adaptierbarkeit) → ausschliessen, keine Testrevision
Umwelt- und Klimaschutz	neu	▪ Repräsentativität (politischen Relevanz): teilweise erfüllt; einige Abstimmungsvorlagen zu Umweltschutz, breite öffentliche Debatte. ▪ Wissenschaftlichkeit: erfüllt; 5 VWL-Kategorien, 2 BWL-Kategorien. ▪ Komplexität: erfüllt. ▪ Bekanntheit: erfüllt; Lerninhalte „Energie/Ökologie“ und „Marktwirtschaft/Umweltsteuer“. ▪ Adaptierbarkeit: erfüllt; aber institutionelle und faktische Besonderheiten berücksichtigen (z.B. Umweltschutzgesetz, Klimaziele).	repräsentativ und adäquat für Konstrukt operationalisierbar im WBK-T2 → vorerst keine Testentwicklung
Jugendverschuldung (JUG)	WBK-T1, Version CH, Version DE	▪ Repräsentativität (politischen Relevanz): nicht erfüllt; keine Abstimmungsvorlagen, kein Politikfeld, repräsentiert nicht den „gesellschaftlichen/wirtschaftlichen Lebensbereich“ des WBK-Modells sondern den „persönlichen Lebensbereich“. ▪ Wissenschaftlichkeit: nicht erfüllt. ▪ Komplexität: nicht erfüllt; von Experten als unter-komplex beurteilt; Items von CoBALIT-Stichprobe wurden überdurchschnittlich gut gelöst. ▪ Bekanntheit: teilweise erfüllt; Lerninhalte „Umgang mit Geld“ in neueren Lehrplänen (z.B. Lehrplan 21 auf Sekundarstufe I). ▪ Kontroversität: nicht erfüllt; keine Interessenskonflikte, keine strittigen Lösungen. ▪ Adaptierbarkeit: erfüllt.	nicht repräsentativ und adäquat für Konstrukt nicht operationalisierbar im WBK-T2 (Wissenschaftlichkeit, Komplexität, Kontroversität) → ausschliessen, keine Testrevision

Tabelle 8-1. Beurteilung sozioökonomischer Problemsituationen für Testentwicklung und -revision

Sozioökonomische Problemsituation	Herkunft	Beurteilung nach Kriterien	Entscheidung
Eurokrise (EUR)	WBK-T1, Version CH, Version DE	▪ Repräsentativität (politischen Relevanz): nicht erfüllt; keine Abstimmungsvorlagen, kein Politikfeld. ▪ Wissenschaftlichkeit: teilweise erfüllt. ▪ Komplexität: nicht erfüllt; von Experten als über-komplex beurteilt, nicht einmal Experten „verstehen“ das Problem; Items von CoBALIT-Stichprobe unterdurchschnittlich gut oder gar nicht gelöst. ▪ Bekanntheit: teilweise erfüllt; wenig Lerninhalte für Problemanalyse, unklare Norm für Kodierung; Lerninhalte „wirtschaftliche Integration/Währungsraum“. ▪ Adaptierbarkeit: erfüllt.	nicht repräsentativ und adäquat für Konstrukt nicht operationalisierbar im WBK-T2 (Wissenschaftlichkeit, Komplexität) → ausschliessen, keine Testrevision

Hinweis: WBK-T1 = originale Version des WBK-Tests (Ackermann, 2018a; Eberle et al., 2016), Version CH = Testversion für die Schweiz, Version DE = adaptierte Testversion für Deutschland.

8.2.1.b Auswahl

Aufgrund der obigen kriteriengeleiteten Inhaltsanalyse erweisen sich sechs der zehn sozioökonomischen Problemsituationen als für die Inhaltsdomäne des Konstrukts repräsentativ und adäquat sowie für die Testentwicklung/-revision operationalisierbar. Von diesen sechs sind vier Problemsituationen bereits im WBK-T1 enthalten und erweisen sich auch für den WBK-T2 als geeignet (vgl. Tabelle 7-3): Altersvorsorge, Energieversorgung, Staatsverschuldung und Managergehälter.

Auf die Entwicklung der zwei Problemsituationen „Arbeitslosigkeit“ und „Umwelt-/Klimaschutz“ musste im Projekt WBKgym aus ressourcentechnischen Gründen verzichtet werden. Zum einen waren bereits vier inhaltsvalide Problemsituationen für die Testung verfügbar, zum anderen hätte es für den Einsatz von vier und mehr Problemsituationen eines rotierten *booklet designs* und einer sehr grossen Stichprobe bedurft (vgl. Frey et al., 2009). Letzteres liess sich im Rahmen eines einzelnen Dissertationsvorhabens nicht bewerkstelligen.

Die Problemsituationen „Agrarhandel“ (im WBK-T1 „Landwirtschaft“) und „Steuerwettbewerb/Finanzausgleich“ (WBK-T1 „Steuerwettbewerb“) sind spezifisch auf den politischen Kontext der Schweiz ausgerichtet, so dass eine Adaptierbarkeit für andere deutschsprachige Länder, insbesondere Deutschland, nicht möglich ist und sie deshalb für den WBK-T2 ungeeignet sind. Bei der Problemsituation „Eurokrise“ des WBK-T1 ist die Repräsentativität nicht gegeben, da sie nicht der politischen Situation der Schweiz entspricht. Zudem hat sie eine geringe Wissenschaftlichkeit und eine vergleichsweise zu hohe Komplexität; sie ist daher für den WBK-T2 ungeeignet. Die Problemsituation „Jugendverschuldung“ des WBK-T1 verfehlt die Repräsentativität, da sie nicht in den gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereich gehört. Ferner hat sie eine vergleichsweise zu geringe Komplexität und keine Kontroversität; sie ist daher für den WBK-T2 ungeeignet.

8.2.2 Inhaltselemente der sozioökonomischen Problemsituationen: Items

Für jede ausgewählte sozioökonomische Problemsituation wurde eine Inhaltsanalyse vorgenommen, um fachdidaktisch akkurate und adäquate Inhaltselemente zu bestimmen (exemplarisch für AHV und ENE vgl. Tabelle 8-2): Versicherungsprinzip, 3-Säulen-System der Vorsorge, Energiesteuer, Energiestrategie 2050. Die Inhaltselemente dienten bei der Konstruktion der Items als Orientierung für eine repräsentative Abbildung der Problemsituation hinsichtlich des Konstrukts. Die Inhaltselemente decken multiple Perspektiven der sozioökonomischen Problemsituation ab, z.B. individuelle und die aggregierte Perspektive, lokale und die globale Ebene, kurze und lange Frist.

Tabelle 8-2. Inhaltselemente der Problemsituationen AHV und ENE im WBK-T2

Sozioökonomische Problemsituation	Inhaltselemente gemäss fachdidaktischer Analyse
Altersvorsorge (AHV)	▪ Bedeutung der sozialen Vorsorge für den Einzelnen und für die Gesellschaft ▪ Gerechtigkeitskonzepte (nach Leistung/meritokratisch, nach Bedarf/demokratisch/sozial) ▪ 3-Säulen-System der Vorsorge in der Schweiz (staatlich, beruflich, privat) ▪ Zweck und Funktionsweise einer Sozialversicherung, insb. AHV: Risiken, Trägerschaft, Finanzierungsverfahren, Beiträge und Leistungen ▪ Ursachen für Unterfinanzierung der AHV: wirtschaftliche Entwicklung (Tertiärsektor/Dienstleistungen), technologische/medizinische Entwicklung (Empfängnisverhütung, Medikamente), demografische Entwicklung (Geburtenrate, Lebenserwartung) ▪ Massnahmen zur Finanzierung der AHV: Rentenalter erhöhen, Renten kürzen, Lohnbeiträge erhöhen, Bundesbeiträge (MWST) erhöhen, etc. ▪ Folgen der Massnahmen zur Finanzierung der AHV auf Arbeitsmarkt und für Wirtschaftsakteure (Haushalte, Unternehmen, Staat, etc.)
Energieversorgung (ENE)	▪ Bedeutung der Energieversorgung für den Einzelnen und für die Gesellschaft ▪ Energieformen: Primärenergie, Sekundärenergie, Endenergie ▪ Energieträger: nicht-erneuerbare Ressourcen/Energien (fossil, nuklear), erneuerbare Ressourcen/Energien ▪ Energieverbrauch der Schweiz: nach Energieträgern und nach Wirtschaftssektoren, zum heutigen Zeitpunkt und für den vergangenen Zeitraum ▪ Ursachen der Umwelt- und Klimabelastung: freie Güter, externe Kosten ▪ Folgen des Energieverbrauchs und CO₂-Ausstosses für Umwelt und Klima ▪ Ziele und Massnahmen der internationaler Klimapolitik (z.B. Kyoto-Protokoll) ▪ Ziele und Massnahmen der Schweizer Energiepolitik (z.B. Energiestrategie 2050) ▪ Massnahmen zur Reduktion des Energieverbrauchs und der CO₂-Emissionen: Lenkungsabgabe, Emissionszertifikate ▪ Massnahmen zur alternativen Energieerzeugung: Solaranlagen, Windanlagen, Biogasanlagen ▪ Folgen des Ausstiegs aus der Kernkraft auf Energiemarkt und für Wirtschaftsakteure (Haushalte, Unternehmen, Staat, Ausland, etc.)

Quelle: Eigene Entwicklung und Darstellung.

Bei der Konstruktion der Items wurden die Inhaltselemente mit dem Kognitionsprozess und dem Antwortformat verwoben, um Wissensstrukturen (v.a. Fakten- und Konzeptwissen) und Problemlösestrukturen (z.B. SWOT-Analyse) des WBK-Modells abzubilden (vgl. Kapitel 4.2). Die Spezifikationen des WBK-T2 zu Inhalt, Format und Kodierung finden sich im Anhang Tabelle A- 4.

8.3 Formatspezifikationen

Dieses Kapitel widmet sich den Formatspezifikationen des WBK-T2: Testaufbau und Testlänge des WBK-T2 (Kapitel 8.3.1), Konstruktion der Einleitungstexte (Kapitel 8.3.2) und Konstruktion der Itemtypen bzw. Items (Kapitel 8.3.3). Die Formatspezifikationen des WBK-T2 orientieren sich grösstenteils am WBK-T1, jedoch wurden einige formelle Modifikationen vorgenommen (Antwortformat, Antwortstrukturierung, Antwortkodierung).

8.3.1 Testaufbau und Testlänge

Jede Problemsituation beginnt mit einem Einleitungstext von maximaler einer A4-Seite, gefolgt von etwa acht Items und Reflexionsfragen (vgl. Abbildung 8-1). Die Einleitungstexte umreissen die jeweilige sozioökonomische Problemsituation (Fakten) und deuten mögliche Lösungsvorschläge an (Positionen, Meinungen) (vgl. Kapitel 8.3.2). Die Items beziehen sich auf verschiedene Inhaltselemente der Problemsituation und beleuchten jeweils multiple Perspektiven der Problemsituation (z.B. individuell vs. kollektiv, lokal vs. global, kurzfristig vs. langfristig) (vgl. Tabelle 8-2). Sie variieren in der Darstellungsform, im Antwortformat und im Kognitionsprozess (vgl. Kapitel 8.3.3). Die Reflexionsfragen beziehen sich jeweils gebündelt auf ein Thema: die Motivation beim Bearbeiten der Items und die Wahrnehmung der Item-Schwierigkeit.

Die Problemsituationen sind in sich thematisch abgeschlossen und nicht aufeinander aufbauend. Sie können modulartig zu einem Testheft zusammengesetzt werden (vgl. Eberle et al., 2016; Siegfried & Ackermann, under review). Allerdings empfiehlt sich aufgrund der unterschiedlichen objektiv gemessenen und der subjektiv wahrgenommen Schwierigkeit der Problemsituationen eine bestimmte Reihenfolge im Testheft, um die Testpersonen nicht zu Beginn der Testung zu demotivieren und zu frustrieren (vgl. Jonkisz et al., 2012, S. 68). Die Reihenfolge der Problemsituationen sollte nach den folgenden Kriterien festgelegt werden:

- Curricularer Bezug: Welche Lerninhalte wurden bereits bzw. wurden erst kürzlich im Unterricht behandelt?
- Politischer Bezug: Über welche Abstimmungsvorlagen wurden jüngst diskutiert bzw. entschieden?

Konkret empfiehlt es sich aufgrund der Erfahrungen im Projekt WBKgym, die Problemsituationen AHV und ENE am Anfang des Testhefts zu platzieren, die Problemsituationen STA und MAN am Ende.

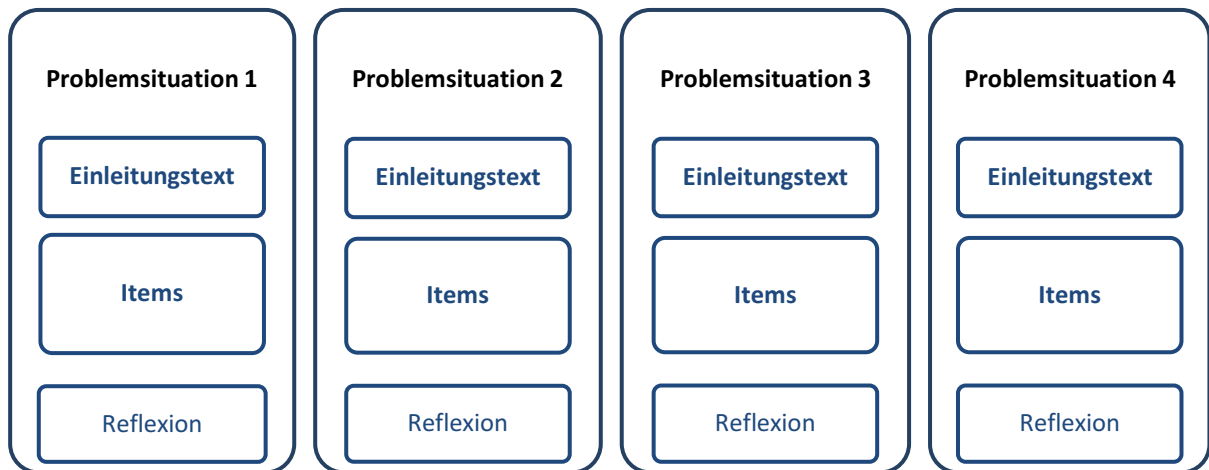


Abbildung 8-1. Aufbau des WBK-T2

Der WBK-T2 besteht aus vier Problemsituationen mit *insgesamt 32 Items* (vgl. Tabelle 7-2), und zwar in der Version CH und in der Version DE. Jede Problemsituation umfasst 7 bis 9 Items, davon 4 bis 5 mit gebundenem Antwortformat und 3 bis 4 mit freiem Antwortformat (vgl. Tabelle 8-5). In Anbetracht des Gütekriteriums Testökonomie (Döring & Bortz, 2016, S. 449f.) (vgl. Kapitel 5.1.2.d) ist die Testlänge ($m = 32$) bzw. die Anzahl Constructed-Response-Items ($m = 15$) an der oberen Grenze. Die manuelle Kodierung der Constructed-Response-Items ist zeit- und kostenintensiv und intellektuell anspruchsvoll, da sie von den Kodierer*innen ein hohes Fachwissen erfordert.

8.3.2 Einleitungstexte

Der Einleitungstext benennt und umreißt die sozioökonomische Problemsituation: Er zeigt die Aktualität und Authentizität der Problemsituation auf (Fakten) und deutet mögliche Lösungsansätze an (Positionen, Meinungen). Insofern erfüllt der Einleitungstext für den WBK-T2 mehrere Funktionen: die sozioökonomische Problemsituation kontextualisieren, Informationen für den Problemlöseprozess vorstrukturieren, Vorwissen der Testpersonen aktivieren. Exemplarisch wird der Einleitungstext zur Problemsituation „Altersvorsorge“ (AHV) gezeigt (vgl. Abbildung 8-2).

Für die Konstruktion der Einleitungstexte wurden die mittels Internetrecherchen zusammengetragenen Materialien nach Informationsquelle und Textsorte sortiert. Bevorzugt wurden bekannte und verlässliche Informationsquellen sowie verschiedene Textsorten, von objektiveren (z.B. Behörden) und subjektiveren Akteuren (z.B. Parteien und Interessengruppen) (vgl. auch Eberle et al., 2016, S. 102). Auf Basis der selektierten Materialien wurden Texte erstellt und wo möglich mit statistischen Abbildungen (Tabellen, Grafiken) ergänzt. Bei der Formulierung der Einleitungstexte wurde auf folgende Kriterien geachtet (vgl. Tabelle 8-3): Konstrukt-repräsentative Beschreibung der Problemsituation (d.h. fachlich korrekte und politisch ausgewogene Inhalte) sowie zielgruppengerechte Sprache (keine/wenige Fremdwörter und Fachbegriffe).

Tabelle 8-3. Konstruktionsregeln für Einleitungstexte

Kriterium	Beispiele
Bekannte und verlässliche Informationsquellen	▪ Behörden, z.B. Bundeskanzlei der Schweizerischen Eidgenossenschaft (admin.ch), div. Bundesämter (z.B. BFS, SECO, BSV, BFE) ▪ internationale Organisationen, z.B. OECD, WTO ▪ Politische Parteien und Interessengruppen, z.B. CVP, FDP, SPS, SVP, Grüne, Economiesuisse, Unia ▪ Medien, z.B. NZZ, Tages-Anzeiger, Bund
Verschiedene Textsorten (Genre)	▪ Fakten, z.B. Factsheets, Statistiken (als Tabellen, Grafiken), Berichte, Analysen ▪ Meinungen, z.B. Positionspapiere, Kommentare, Leserbriefe
Konstrukt-repräsentative Beschreibung der Problemsituation	▪ Fachlich korrekter Inhalt ▪ Politisch ausgewogener Inhalt: progressive und konservative Positionen
Zielgruppengerechte Sprache	▪ Semantik/Vokabular: keine Fremdwörter, keine Fachbegriffe ▪ Syntax: einfacher Satzbau

Quelle: Eigene Entwicklung (vgl. auch Eberle et al., 2016, S. 102).

8.3.3 Itemtypen

Um für die Iteminhalte eine Konstrukt-repräsentative und -adäquate Variation im Itemformat zu erreichen, wurde ein Konstruktionsschema mit sechs Itemtypen entworfen (vgl. Tabelle 8-4). Dieses Konstruktionsschema diente als Referenzrahmen für die Testentwicklung und -revision und ermöglichte eine systematische und regelgeleitete Item-Konstruktion. Die Itemtypen variieren im Kognitionsprozess, in der Darstellungsform, im Antwortformat und in der Kodierung (*scoring*) (vgl. Shavelson, 2012).

Beim Kognitionsprozess wird zwischen drei beabsichtigten bzw. erforderlichen Kognitionsstufen unterschieden (vgl. Kapitel 4.4.1): (k1) Wiedererkennen/Wiedergeben, (k2) Verstehen/Anwenden, (k3) Vergleichen/Beurteilen/Entscheiden. Die Darstellungsform kann entweder Text, Tabelle (Entscheidungsmatrix, SWOT-Matrix) oder Diagramm (Netzdiagramm, *Concept Map*) sein. Das Antwortformat ist gebundenen oder frei; es wurde eine dreistufige Einteilung gewählt: (f1) *Selected-Response* (Multiple-Choice), (f2) *Short-Constructed-Response* (Kurzantwort, „nennen“), (f3) *Extended-Constructed-Response* (ausführliche Antwort, „erklären“ oder „begründen“) (vgl. Jonkisz et al., 2012, S. 39ff.). Die Kodierung der Itemtypen ist entweder dichotom mit den Stufen 0/1 oder polytom mit den Stufen 0/1/2. Beim Itemtypus V und VI werden die dichotom kodierten Teilantworten zum Wert 2 bzw. 4 summiert.

Die Alters- und Hinterlassenenversicherung

Die Alters- und Hinterlassenenversicherung (AHV) bildet zusammen mit der Invalidenversicherung (IV) die 1. Säule im schweizerischen Vorsorgesystem. Die AHV zahlt verschiedene Renten aus:

- die Altersrente deckt den Existenzbedarf im Alter;
- die Waisenrente bzw. Halbwaisenrente verhindert eine finanzielle Notlage beim Tod der Eltern bzw. eines Elternteils;
- die Witwer-/Witwenrente verhindert eine finanzielle Notlage beim Tod des Ehegatten bzw. des eingetragenen Partners.

Die AHV wird nach dem sogenannten Umlageverfahren finanziert: Die einbezahlten Beiträge werden sofort wieder als Renten ausgezahlt. Im Unterschied zur beruflichen Vorsorge (Pensionskasse) oder zur privaten Vorsorge (z.B. Sparbuch) wird bei dieser Finanzierungsart kein Kapital angehäuft. Der Beitragssatz für die AHV beträgt 8.4 % vom Bruttolohn und wird hälftig von Arbeitgeber und Arbeitnehmer getragen. Die Ausgaben der AHV werden nicht allein durch Lohnbeiträge gedeckt. Der Bund kommt für knapp ein Fünftel der Ausgaben der AHV auf. Sein Beitrag wird unter anderem durch die Mehrwertsteuer sowie die Tabak- und Alkoholsteuer finanziert.

Die demographische Entwicklung der Schweiz zeigt, dass in der Vergangenheit die Lebenserwartung gestiegen und die Geburtenrate gesunken ist. In Zukunft dürfte die Zahl der Rentner weiterhin ansteigen, während die Zahl der Erwerbstätigen stagniert oder sogar leicht abnimmt. 1970 kamen 4.3 Erwerbstätige für einen Rentner auf, derzeit sind es 3.7, 2030 werden es noch 2.1 sein. Die steigende Zahl der Rentner kann nicht mehr durch eine steigende Anzahl Erwerbstätiger aufgefangen werden. Das stellt den sogenannten Generationenvertrag in Frage. Für das Jahr 2030 wird das Finanzierungsdefizit der AHV auf ca. 55 Milliarden Schweizer Franken geschätzt.

Massnahmen für die AHV

Der Bundesrat will die AHV mit folgenden Massnahmen finanziell sanieren:

- Rentenalter 65 für Frauen: Das Rentenalter für Frauen soll von 64 auf 65 Jahre erhöht und somit dem heutigen Rentenalter für Männer angepasst werden.
- Rentenalter 67 für alle: Das Rentenalter soll für alle Personen auf 67 erhöht werden.
- Altersrücktritt «à la carte»: Jeder soll sein Pensionsalter selber bestimmen. Bei Frühpensionierung wird die Rente gekürzt. Für die Spätpensionierung sind finanzielle Anreize vorgesehen.
- Die Mehrwertsteuer soll zugunsten der AHV erhöht werden.

Quellen (modifiziert):

„Demographie und AHV“, Bundesamt für Sozialversicherungen (BSV), 29.09.2008.

„Grundlagen der AHV“, Bundesamt für Sozialversicherungen (BSV), 10.10.2008.

„Bersets Visionen für die AHV“, Neue Zürcher Zeitung, 21.11.2012.

Abbildung 8-2. Beispiel für Einleitungstext (Problemsituation AHV)

Quelle: Ackermann (2018c).

Die ersten vier Itemtypen (I, II, III und IV) wurden bereits im Projekt CoBALIT für den WBK-T1 eingeführt und verwendet (Eberle et al., 2016, S. 102f.). Die letzten zwei Itemtypen (V und VI) wurden im Projekt WBKgyM für den WBK-T2 ergänzt. Die Erweiterung des Konstruktionsschemas um die zwei neuen Itemtypen V und VI war darauf ausgelegt, eine von Experten kritisierte Abhängigkeit von Kognitionsprozess und Antwortformat bei den Items zu durchbrechen. Zudem wurden für den WBK-T2 die Itemtypen III und IV hinsichtlich Antwortstrukturierung (Umfang) und Antwortkodierung modifiziert. Einzig die Itemtypen I und II blieben unverändert. Nachfolgend wird für jeden Typus exemplarisch ein Item abgebildet und erläutert. Die Spezifikationen des WBK-T2 zu Inhalt, Format und Kodierung finden sich im Anhang Tabelle A- 4.

Tabelle 8-4. Konstruktionsschema für Items (Itemtypen)

Typ	Beschreibung	Kognitions- prozess	Darstellungs- form	Antwortformat	Antwort- kategorien (Scoring)
I	Eine bestmögliche Antwort aus vier Antwortoptionen wählen	k1: wiedererkennen, wiedergeben	Text	f1: Selected- Response	Dichotom (0/1)
II	Eine Aussage qualifizieren (R/F) und begründen	k2: verstehen, anwenden	Text	f3: Extended- Constructed- Response	Dichotom (0/1)
III	Eine Ursache, Wirkung oder Massnahme erläutern	k2: verstehen, anwenden	Text	f3: Extended- Constructed- Response	Polytom geordnet (0/1/2)
IV	Vorteile/Nachteile von Varianten anhand Kriterien abwägen, für 1 Variante entscheiden und begründen	k3: vergleichen, beurteilen, entscheiden	Tabelle	f3: Extended- Constructed- Response	Polytom geordnet (0/1/2)
V	Begriffe (Merkmale, Beispiele) nennen	k1: wiedererkennen, wiedergeben	Text	f2: Short- Constructed- Response	Dichotom (0/1) → summiert auf 2
VI	Zusammenhänge erkennen, die Beziehung zwischen Elementen qualifizieren (+/0/-)	k2: verstehen, anwenden	Diagramm	f1: Selected- Response	Dichotom (0/1) → summiert auf 4

Quelle: Eigene Entwicklung (vgl. auch Eberle et al., 2016, S. 102f.).

Beim *Itemtypus I* sollen die Testpersonen aus vier Antwortoptionen eine bestmögliche Antwort wählen (vgl. Abbildung 8-3). Hiermit wird vornehmlich Faktenwissen und elementares Konzeptwissen geprüft. Es ist nur eine bestimmte Lösung zugelassen, d.h. nur eine der vier Antwortoptionen ist korrekt. Dieser Typus blieb von der Revision des WBK-Tests unberührt und findet sich gleichermassen im WBK-T1 und im WBK-T2.

Der *Itemtypus II* fordert die Testpersonen auf, eine Aussage mit richtig oder falsch zu qualifizieren und ihre Antwort zu begründen (vgl. Abbildung 8-4). Hierbei geht es um die Prüfung vertieften Konzeptwissens. Die Aussagen sind fachwissenschaftlich nicht eindeutig zu beantworten, insofern sind mehrere unbestimmte Lösungen als korrekte Antwort zugelassen, sofern die Antwort kohärent begründet ist. Dieser Typus blieb ebenfalls von der Revision des WBK-Tests unberührt.

Nach welchem Versicherungsprinzip funktioniert die AHV?	
(1)	☐ Arbeitnehmer zahlen Beiträge ein, woraus ihre eigene Rente finanziert wird.
(2)	☐ Arbeitnehmer sparen Kapital an, woraus ihre eigene Rente finanziert wird.
(3)	☒ Arbeitgeber und Arbeitnehmer zahlen Beiträge ein, woraus Renten finanziert werden.
(4)	☐ Arbeitgeber und Arbeitnehmer sparen Kapital an, woraus Renten finanziert werden.

Code	Erläuterung
1	(3) korrekte Antwort. Umlageverfahren, Arbeitnehmer und Arbeitgeber leisten Beiträge, um fremde Renten zu finanzieren.
0	(1) falsche Antwort. Umlageverfahren, aber Arbeitnehmer und Arbeitgeber leisten Beiträge. (2) falsche Antwort. Kapitaldeckungsverfahren, bei beruflicher und privater Vorsorge, für eigene Rente. (4) falsche Antwort. Kapitaldeckungsverfahren, bei beruflicher Vorsorge, für eigene Rente.
999	Keine Antwort.
777	Zwei Antworten.

Abbildung 8-3. Beispiel des unveränderten Itemtypus I mit Kodierungschema (Item AHV_5)

Quelle: Ackermann (2018a, 2018c).

Entscheiden Sie, ob die folgende Aussage richtig oder falsch ist. Begründen Sie Ihre Antwort. „Durch den Ausstieg aus der Kernenergie steigt der Preis der Elektrizität.“				
Antwort:	☐	richtig	☐	falsch
Begründung:				

Abbildung 8-4. Beispiel des unveränderten Itemtypus II (Item ENE_8)

Quelle: Ackermann (2018a, 2018c).

Beim *Itemtypus III* sollen die Testpersonen eine Ursache, Wirkung oder Massnahme erläutern (vgl. Abbildung 8-5). Mit diesem Typus wird ebenfalls vertieftes Konzeptwissen geprüft und es sind mehrere unbestimmte Lösungen zugelassen, sofern die Antwort korrekt und kohärent begründet ist. Der Typus III wurde bei der Testrevision für den WBK-T2 formell modifiziert, d.h. bezüglich Antwortstrukturierung und Umfang (eine Auswirkung erläutern statt zwei Auswirkungen nennen und begründen) und folglich bezüglich Kodierung (polytom 2-stufig statt polytom 4-stufig). Für diese Modifikation gab es diverse Gründe: Die konstruierte Doppelantwort (2x Nennung und Begründung) lies lokale stochastische Abhängigkeit vermuten, da Nennung 2 abhängig von Nennung 1 ist. Eine getrennte Kodierung von Nennung und Begründung erwies sich als schwierig, da teilweise die Nennung nicht explizit, sondern implizit in der Begründung enthalten war; zudem war die Kodierung zweideutig, da der Itemscore 2 entweder als zwei korrekte Nennungen oder eine korrekte Nennung mit Begründung zu interpretieren war. Die psychometrische Qualität der Items war in der CoBALIT Stichprobe eher unbefriedigend (Lösungsquote < 20 %).

AHV_11

Wie wirkt sich die Erhöhung des Rentenalters, auf z.B. 68 Jahre, auf den Arbeitsmarkt aus? Nennen und begründen Sie zwei Auswirkungen.		
Massnahme 1 <i>AHV_11a1</i>	Nennung:	
	Begründung:	
Massnahme 2 <i>AHV_11a2</i>	Nennung:	
	Begründung:	

AHV_11m

Wie wirkt sich eine Erhöhung des Rentenalters (z.B. auf 67 Jahre) auf den Arbeitsmarkt aus? Erläutern Sie eine Auswirkung.
Erläuterung:

Code	Erläuterung und Beispiele
---	aus Erläuterung muss hervorgehen, dass Erwerbspersonen gemeint sind: „Erwerbspersonen“ = Erwerbstätige + Erwerbslose; „Beschäftigung“ ≠ Erwerbspersonen.
2	Vollständige korrekte Antwort. ▪ <i>höhere Arbeitslosigkeit:</i> ○ <i>ältere AN besetzen Stellen länger → weniger offene Stellen für jugendliche AN → Unternehmen haben moralische Verpflichtung zur Weiterbeschäftigung von älteren und erfahrenen AN → wenig Chancen für jüngere unerfahrene AN → höhere Jugendarbeitslosigkeit.</i> ○ <i>ältere AN sind gegenüber jungen AN weniger konkurrenzfähig → Unternehmen stellen eher jüngere AN ein → ältere AN finden im Alter keine Stelle → höhere Altersarbeitslosigkeit.</i> ▪ <i>höhere Personalkosten für Unternehmen: Lohnkosten: ältere AN sind auf höherer Lohnstufe als jüngere AN → höhere Lohnkosten. Sozialkosten: ältere AN sind anfälliger auf Krankheit und Unfall → höhere Kosten für Lohnfortzahlung bei Krankheit und Unfall.</i> ▪ <i>geringere Personalkosten für Unternehmen: ältere AN bleiben länger im Betrieb → weniger Kosten für Personalrekrutierung und -einarbeitung. mehr Stellensuchende als offene Stellen → mehr Verhandlungsmacht der AG → geringere Löhne.</i> ▪ <i>mehr Erwerbspersonen (NICHT: Erwerbstätige!!): junge AN kommen auf den Arbeitsmarkt, ältere AN bleiben auf dem Arbeitsmarkt.</i> ▪ <i>geringere Arbeitsproduktivität: ältere AN sind schwächer, langsamer, unflexibler als junge AN → geringere Leistungsfähigkeit → geringere Arbeitsproduktivität.</i> ▪ <i>mehr Weiterbildungen: Erwerbsleben dauert länger → berufliche Qualifikation wird erhöht/angepasst → life-long-learning</i> ▪ <i>geringere Motivation: ältere AN wollen ihren Lebensabend genießen und sind nicht motiviert zusätzliche Jahre zu arbeiten.</i> ▪ <i>weniger Brain Drain: ältere AN haben Fachwissen und Erfahrung → Know-how kann länger genutzt werden und geht nicht verloren.</i> ▪ <i>... [keine abschliessende Liste]</i>
1	Unvollständige korrekte Antwort. Teilweise korrekte Antwort. ▪ <i>weniger Stellen [HINWEIS: aus Antwort/Begründung muss hervorgehen, dass „offene Stellen“ gemeint sind].</i> ▪ <i>mehr Arbeitskräfte [HINWEIS: aus Antwort/Begründung muss hervorgehen, dass „Erwerbspersonen“ gemeint sind].</i> ▪ <i>Durchschnittsalter der Erwerbspersonen steigt.</i> ▪ <i>mehr Leistung kann erbracht werden, da ältere Menschen länger arbeiten.</i> ▪ <i>... [keine abschliessende Liste]</i>
0	Falsche Antwort. ▪ <i>bessere Finanzierung der AHV: 2 Rentenjahre fallen weg [HINWEIS: kein Bezug zum Arbeitsmarkt].</i> ▪ <i>mehr Arbeitnehmer/Beschäftigte [HINWEIS: mehr Erwerbspersonen, gleich viele Stellen, c.p.].</i> ▪ <i>weniger Rentner.</i> ▪ <i>Produktion steigt.</i> ▪ <i>... [keine abschliessende Liste]</i>
999	Keine Antwort
888	Inhaltsfremde Kommentare

Abbildung 8-5. Beispiel des modifizierten Itemtypus III mit Kodierungschema (Item AHV_11 und AHV_11m)

Quelle: Ackermann (2018a, 2018c).

Beim *Itemtypus IV* sollen die Testpersonen Vorteile und Nachteile von zwei Lösungsvarianten anhand vorgegebener Kriterien abwägen, um sich für eine Variante zu entscheiden und sie zu begründen (vgl. Abbildung 8-6). Damit wird zum einen vertieftes Konzeptwissen und zum anderen Methodenwissen (spezifische Problemlöseverfahren mittels SWOT-Tabelle) geprüft. Auch hier sind mehrere unbestimmte Lösungen zugelassen, sofern die Antwort korrekt und kohärent begründet ist. Der Typus IV wurde bei der Testrevision für den WBK-T2 inhaltlich und formell modifiziert, d.h. bezüglich Informationsvorgabe, Antwortstrukturierung und Umfang (vorgegebene SWOT-Tabelle statt leere Entscheidungstabelle) sowie bezüglich Kodierung (polytom 2-stufig statt polytom 5-stufig). Für diese Modifikation gab es etliche Gründe: Die konstruierten Teilantworten liessen lokale stochastische Abhängigkeit vermuten, da die Beurteilung b abhängig von der Beurteilung a ist, was sich in der CoBALIT Stichprobe durch repetitive Antworten zeigte. Antwortformat und Antwortstrukturierung wurde von den Experten als methodisch anspruchsvoll und intuitiv unverständlich beurteilt; dies zeigte sich in der CoBALIT Stichprobe durch (zu) viele fehlende Antworten. Die psychometrische Qualität der Items war bei der CoBALIT Stichprobe eher unbefriedigend (Lösungsquote < 10 %).

Der *Itemtypus V* fordert die Testpersonen auf, Unterbegriffe zu einem Oberbegriff zu nennen (vgl. Abbildung 8-7). Damit wird vornehmlich Faktenwissen und elementares Konzeptwissen geprüft. Es sind mehrere bestimmte Lösungen zugelassen, d.h. eine beschränkte Anzahl Antworten ist korrekt. Der Typus V wurde aus zwei Gründen neu konstruiert. Erstens, um Inhaltselemente, die mit dem Itemtypus I aufgrund schlechter Distraktoren nicht funktioniert hatten (z.B. Items AHV_1, ENE_2, MAN_1, MAN_4), beizubehalten und in ein neues Format zu überführen. Zweitens, um einer Format-Kognition-Abhängigkeit entgegenzuwirken; so standen für theoretisch einfachere Inhaltselemente nun ein gebundenes (Typus I) und freies (Typus V) Antwortformat zur Verfügung.

Beim *Itemtypus VI* sollen die Testpersonen die Beziehung zwischen Elementen mit positiv, negativ oder neutral qualifizieren (vgl. Abbildung 8-8). Mit diesem Typus wird vertieftes Konzeptwissen und Methodenwissen (spezifisches Problemlöseverfahren mittels Netzwerkdiagramm) geprüft. Es ist nur eine bestimmte Lösung zugelassen, d.h. nur eine der drei Antwortoptionen ist korrekt. Der Typus VI wurde neu konstruiert, um neue Inhaltselemente aufzunehmen, die mit den bisherigen Itemtypen nicht abgedeckt werden konnten, insbesondere die Darstellung multipler Zusammenhänge anhand einer Strukturskizze.

STA_8

Wie soll das Schuldenproblem des Staates gelöst werden? Beurteilen und begründen Sie die zwei Lösungsansätze anhand der vorgegebenen Kriterien in der Tabelle.			
Kriterium	Lösungsansatz 1: Senkung der Staatsausgaben		Lösungsansatz 2: Erhöhung der Steuereinnahmen
politische Realisierbarkeit	Beurteilung: besser		Beurteilung: schlechter
	Begründung: Ausgaben senkungen eher politisch durchsetzbar		Begründung: Steuererhöhungen kaum politisch durchsetzbar
Wirtschaftswachstum	Beurteilung:		Beurteilung:
	Begründung:	STA_8a	Begründung: STA_8b
Sozialer Ausgleich	Beurteilung:		Beurteilung:
	Begründung:	STA_8c	Begründung: STA_8d
Entscheiden Sie sich für einen der obigen Lösungsansätze und begründen Sie Ihre Entscheidung.			
Entscheidung: STA_8e	☐	Lösungsansatz 1	
	☐	Lösungsansatz 2	
	☐	keiner der obigen Lösungsansätze	
Begründung:			

STA_8m

Wie soll das Schuldenproblem des Staates in Zukunft gelöst werden? In der Politik werden gegenwärtig zwei Lösungsansätze diskutiert: 1) Staatsausgaben senken, 2) Steuereinnahmen erhöhen. Die Tabelle zeigt die Beurteilung (Vorteile und Nachteile) der Lösungsansätze ausgehend von der aktuellen Situation.		
Lösungsansatz 1) Staatsausgaben senken		
Vorteile ▪ Anreiz zu mehr Effizienz der staatlichen Leistungen ▪ Priorisierung der staatlichen Ausgabegebiete ▪ evtl. Privatisierung von staatlichen Leistungen ▪ ...	Nachteile ▪ für Unternehmen: weniger staatliche Aufträge → weniger Gewinne → weniger Investitionen ▪ für Haushalte: weniger Sozialleistungen → weniger Einkommen → weniger Konsum ▪ für Gesamtwirtschaft: weniger staatlicher Konsum ▪ ...	
Lösungsansatz 2) Steuereinnahmen erhöhen		
Vorteile ▪ Steuerprogression möglich → mehr Steuereinnahmen ▪ evtl. Umverteilung von Steuereinnahmen ▪ ...	Nachteile ▪ für Haushalte: weniger Einkommen und Vermögen → weniger Konsum ▪ für Unternehmen: weniger Gewinn und Kapital → weniger Investitionen ▪ ...	
Entscheiden Sie sich für einen Lösungsansatz. Begründen Sie Ihre Entscheidung, indem Sie sich auf die Kriterien Wirtschaftswachstum und sozialer Ausgleich beziehen.		
Entscheidung:	☐	Lösungsansatz 1) Staatsausgaben senken
	☐	Lösungsansatz 2) Steuereinnahmen erhöhen
Begründung:		

Abbildung 8-6. Beispiel des modifizierten Itemtypus IV (Item STA_8 und STA_8m)

Quelle: Ackermann (2018a, 2018c).

ENE_2

Welche drei Energieträger sind erneuerbar?		
(1)	☐	Erdöl, Erdgas, Erdwärme
(2)	☐	Holz, Torf, Kohle
(3)	☐	Wasser, Wind, Uran
(4)	☐	Biomasse, Geothermie, Solarkraft

ENE_2m

Welche Energieträger sind nicht-erneuerbar? Nennen Sie zwei.	
Nennung 1:	
Nennung 2:	

Abbildung 8-7. Beispiel des neuen Itemtypus V (Item ENE_2 und ENE_2m)

Quelle: Ackermann (2018a, 2018c).

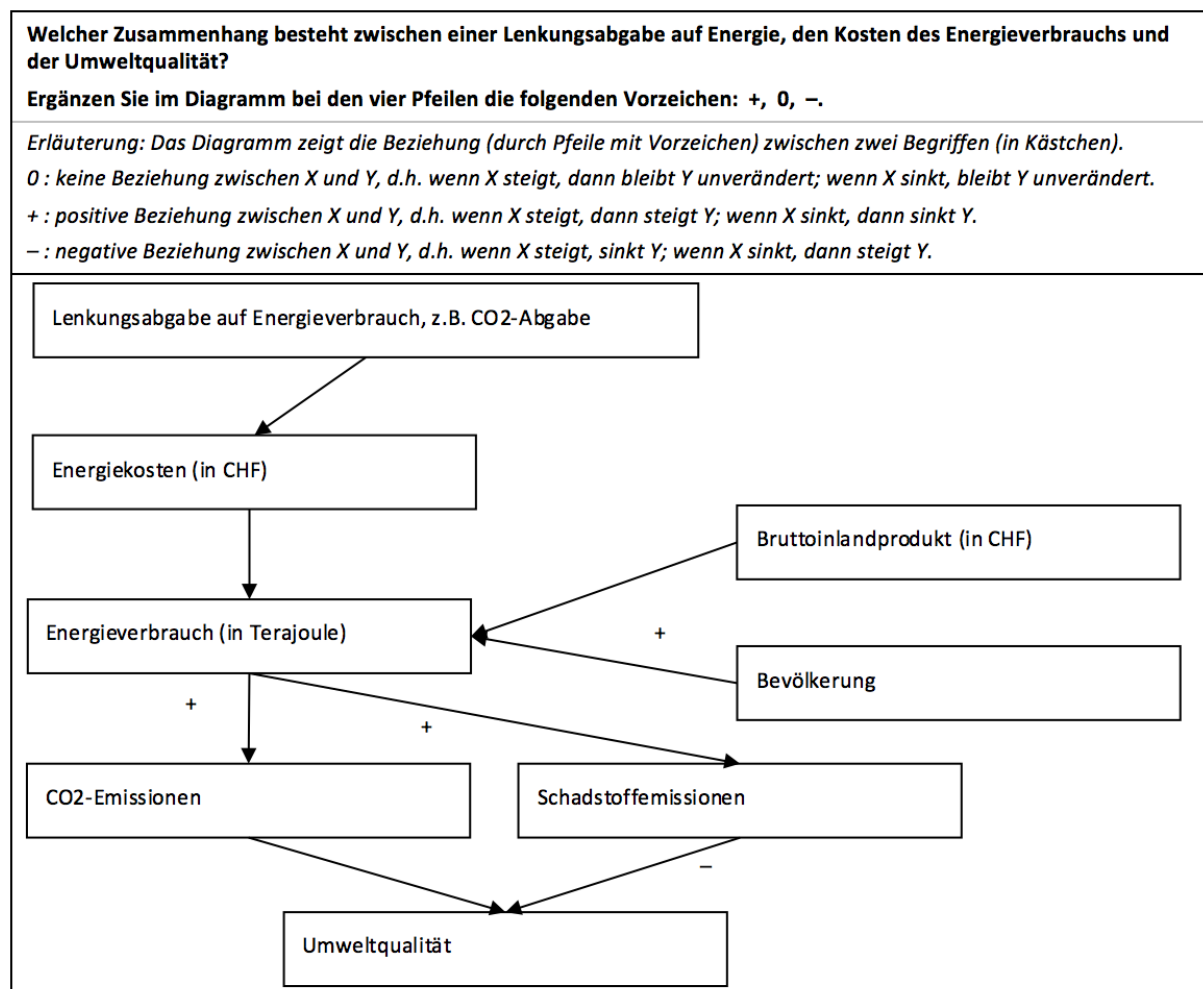


Abbildung 8-8. Beispiel des neuen Itemtypus VI (Item ENE_12n)

Quelle: Ackermann (2018a, 2018c).

Angelehnt an das Konstruktionsschema (vgl. Tabelle 8-4) verteilen sich die Items des WBK-T2 wie folgt auf Antwortformat und Kognitionsstufe (vgl. Tabelle 8-6): Insgesamt haben 53 % der Items das Format Selected-Response (f1) und 47 % Constructed-Response (f2/f3). 56 % der Items sind auf der unteren Kognitionsstufe (k1), 34 % auf der mittleren (k2) und 9 % auf der oberen (k3).

Für das gebundene Antwortformat (f1) gibt es 13 Items auf der unteren Kognitionsstufe (k1) und 4 auf der mittleren Kognitionsstufe (k2), jedoch keine auf der oberen Kognitionsstufe (k3). Beim freiem Antwortformat (f2/f3) befinden sich 5 Items auf der Stufe k1, 7 auf der Stufe k2 und 3 auf der Stufe k3. Insgesamt sind die Items zwischen den Antwortformaten gleichmässig verteilt (f1, 53 %, f2/f3, 47 %), hingegen gibt es beim Kognitionsprozess einen Überhang der unteren Stufe (k1, 56 %) gegenüber der mittleren (k2, 34 %) und oberen Stufe (k3, 9 %). Letzteres erscheint jedoch aus Gründen des Testaufbaus und der Testökonomie gerechtfertigt: Jede Problemsituation soll mehrere Items beinhalten, bei denen Wissensstrukturen bezüglich des „Problems“ aktiviert werden müssen. Es soll aber höchstens ein Item vorkommen, bei dem Problemlösestrukturen aktiviert werden müssen und das „Problem“ (abstrakt) gelöst wird.

8.4 Kodierungsspezifikationen

Für eine objektive und reliable Kodierung des WBK-T2 wurde für jedes Item ein Kodierungsschema mit Bewertungsstufen, -kriterien und -hinweisen entwickelt. Diese Kodierungsschemata wurden in einem Kodierungsmanual zusammengetragen. Das Kodierungsmanual des WBK-T2 basiert auf demjenigen des WBK-T1: Für die unveränderten und modifizierten Items wurden die bestehenden Kodierungshinweise überprüft und ggf. überarbeitet; für die neu konstruierten Items wurden Kodierungshinweise erstellt und von Experten validiert. Die Spezifikationen des WBK-T2 zu Inhalt, Format und Kodierung finden sich im Anhang Tabelle A- 4.

Das Kodierungsmanual enthält für jedes Item die Frageformulierung und Antwortstrukturierung aus dem Testheft sowie allgemeine und spezifische Kodierungshinweise. Bei den Items mit gebundenem Antwortformaten (Itemtypus I und V) beschränken sich die Kodierungshinweise auf den Code für die korrekte Antwort und die falschen Antworten (vgl. Abbildung 8-3). Bei den Items mit freiem Antwortformat (Itemtypus II, III, IV und VI) hingegen gibt es allgemeine Kodierungshinweise (z.B. Oberbegriffe und Unterbegriffe als Antwort nur einmal werten, Wiederholung und Umformulierung der Frage nicht werten) und spezifische Kodierungshinweise i.S.v. Musterlösungen für jeden Code (vgl. Abbildung 8-5).

Tabelle 8-5. Item-Anzahl im WBK-T2 nach Problemsituation und Itemtypus

Problemsituation	Typus I	Typus II	Typus III	Typus IV	Typus V	Typus VI	Total
Altersvorsorge (AHV)	4	1	1	1	1	1	9
Energieversorgung (ENE)	3	1	0	1	1	1	7
Staatsverschuldung (STA)	4	1	1	1	1	1	9
Managervergütungen (MAN)	2	1	1	0	2	1	7
Total	13	4	3	3	5	4	32

Tabelle 8-6. Item-Anzahl im WBK-T2 nach Antwortformat und Kognitionsprozess

		Antwortformat		Total
		Selected-Response (f1)	Constructed-Response (f2/f3)	
Kognitionsprozess	Wiedererkennen, wiedergeben (k1)	Typus I: 13 (41 %)	Typus V: 5 (16 %)	18 (56 %)
	Verstehen, anwenden (k2)	Typus VI: 4 (13 %)	Typus II und Typus III: 7 (22 %)	11 (34 %)
	Vergleichen, beurteilen, entscheiden (k3)	---	Typus IV: 3 (9 %)	3 (9 %)
	Total	17 (53 %)	15 (47 %)	32 (100 %)

8.5 Testadministration

Dieses Kapitel widmet sich dem Testverfahren des WBK-T2 (Kapitel 8.5.1) und der Testdauer des WBK-T2 (Kapitel 8.5.2).

8.5.1 Testverfahren

Der WBK-T2 ist ein psychologischer Test von der Art eines schriftlichen Leistungstests. Er ist innerhalb der Testteile als Power-Test konzipiert, aber er ist nicht als Speed-Test intendiert. Er kann in einem traditionellen *Paper-and-Pencil*-Verfahren oder in einem Computer-Verfahren dargeboten werden.

Das Online-Verfahren wurde im Projekt CoBALIT umgesetzt (Eberle et al., 2016): Es erwies sich in der Vorbereitung und Durchführung als aufwändiger (Programmierung und Pilotierung der Online-Testumgebung, Organisation von Erhebungsräume mit Computern und Internetanschluss, Sicherung der Testdaten etc.) und in der Nachbereitung als nicht vorteilhafter („manuelle“ Kodierung der Items mit freiem Antwortformat). Aus forschungsökonomischen Gründen wurde daher im Projekt WBKgym wiederum auf das *Paper-and-Pencil*-Verfahren zurückgegriffen, das bereits bei der Pilot-Erhebungen des Projekts CoBALIT erprobt worden war.

Da es sich beim WBK-T2 um einen Leistungstest handelt, scheint das Nebengütekriterium Unverfälschbarkeit nur in die eine Richtung gefährdet zu sein, nämlich dass die Testpersonen absichtlich schlechte Testwerte liefern (Döring & Bortz, 2014, S. 449f.) (vgl. Kapitel 5.1.2.d). Für eine solche absichtliche Verfälschung der Testwerte können bei der Testdurchführung und -auswertung des WBK-T2 verschiedene Indizien gesammelt und kumulativ erwogen werden:

- Das Testheft wird nach weniger als der Hälfte der Testzeit abgegeben (Notiz im Erhebungsprotokoll, vgl. Kapitel 10.1.2);
- im Testheft sind weniger als die Hälfte der Items nicht bearbeitet (Code 999 in Kodierungstabelle, vgl. Kapitel 8.4 und 10.2.2.a);
- der Testwert liegt weit unter dem „minimal standard“ von 50 % (deskriptive Analysen, vgl. Kapitel 8.1.2 und 11.5.1).

8.5.2 Testdauer

Im Standard-Pretesting haben sich pro Testteil (Problemsituation), je nach Zielgruppe, 15 bis 20 Minuten als ideale Testdauer herausgestellt. Bei der Stichprobe von kaufmännischen Berufsschüler*innen waren 20 Minuten pro Testteil angezeigt (vgl. Projekt CoBALIT). Bei der Stichprobe Gymnasiast*innen haben sich 15 Minuten pro Testteil bewährt (vgl. Projekt WBKgym), da diese Testpersonen i.d.R. über höhere kognitive Grundfähigkeiten verfügen und entsprechend schneller Texte lesen und interpretieren können. In dieser Zeit gelingt es ca. 90 % der Gymnasiast*innen, alle vorgelegten Items des Testteils zu bearbeiten.

Unter Berücksichtigung des Gütekriteriums Zumutbarkeit (Döring & Bortz, 2016, S. 449f.) (vgl. Kapitel 5.1.2.d) sollte die Testung für jede Testperson nicht länger als 60 Minuten dauern. Die Testdauer von ca. 60 Minuten hat sich beim Projekt WBKgym bewährt. Die Leistungsfähigkeit (Konzentration) und Leistungsbereitschaft (Motivation) lässt im Verlauf der Testung nach, so dass die Schüler*innen die Items entweder gar nicht oder nur oberflächlich lösen. Angesichts der fehlenden Werte im Datensatz wäre somit auch die Reliabilität des Tests beeinträchtigt und die valide Interpretation der Testwerte hinsichtlich des Konstrukts fragwürdig (Döring & Bortz, 2014, S. 445). Zudem können bei empirischen Studien im schulischen Kontext aus politischen und organisatorischen Gründen selten mehr als zwei bis drei Unterrichtsstunden pro Klasse für eine Datenerhebung beansprucht werden.

8.6 Zwischenfazit

In diesem Kapitel wurden die Spezifikationen des WBK-T2 beschrieben. Aus domänenlogischen und forschungspragmatischen Gründen wurde im WBK-T2 nicht alle Lebensbereiche des WBK-Modells operationalisiert sondern auf der Strukturebene Anforderungssituationen im gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereich (s3) sowie auf der Prozessebene alle drei Stufen der Informationsverarbeitung (k1, k2, k3) und die zwei Phasen (p2) und (p3) des Problemlösens. Das dem WBK-T2 zugrundeliegende Konstrukt ist also die *sozioökonomische Facette der WBK* (WBK-soek) des WBK-Modells. Für die WBK-soek wurden eine Reihe von Annahmen postuliert, die es im Rahmen der evidenzbasierten Validierung zu prüfen gilt.

Der WBK-T2 richtet sich an Schüler*innen am Ende der Sekundarstufe II. Als Zielgruppe fungieren speziell jene an kaufmännischen Berufsfachschulen, an Berufsmaturitätsschulen im Schwerpunkt WuR und an gymnasiale Maturitätsschulen im Schwerpunkt WuR vorgesehen. Als Kontrastgruppe fungieren Schüler*innen an nicht-kaufmännischen Berufsfachschulen (technisch, gewerblich, gestalterisch, gesundheitlich, sozial), an Berufsmaturitätsschulen und gymnasialen Maturitätsschulen mit anderem Schwerpunktfach.

Für die Inhaltsspezifikationen des WBK-T2 wurde der gesamtgesellschaftliche/gesamtwirtschaftliche Lebensbereich nach den Kriterien Repräsentativität (politische Relevanz), Wissenschaftlichkeit, Komplexität/Kontroversität/Multiperspektivität, Bekanntheit und Adaptierbarkeit analysiert, um geeignete sozioökonomische Problemsituationen zu identifizieren. Schliesslich erwiesen sich vier sozioökonomische Problemsituation aus dem WBK-T1 für den WBK-T2 als geeignet: Altersvorsorge, Energieversorgung, Staatsverschuldung und Managergehälter. Diese wurden fachdidaktisch analysiert, um akkurate und adäquate Inhaltelemente für die Items zu eruieren.

Für die Formatspezifikationen des WBK-T2 wurden sechs Itemtypen definiert, die in der Darstellungsform, im Antwortformat und im Kognitionsprozess variieren. Insgesamt sind 53 % der Items Selected-Response (f1) und 47 % Constructed-Response (f2/f3). 56 % der Items sind auf der unteren Kognitionsstufe (k1), 34 % auf der mittleren (k2) und 9 % auf der oberen (k3). Für die Kodierung der Items des WBK-T2 wurde ein umfangreiches Kodierungsmanual entwickelt bzw. weiterentwickelt. Für die Kodierungsspezifikationen wurde für jedes Item ein Kodierungsschema (Bewertungsstufen und -kriterien) entwickelt und in einem Kodierungsmanual zusammengetragen.

Inwiefern die Testwerte des WBK-T2 hinsichtlich des Konstrukts sozioökonomische Facette der WBK (WBK-soek) interpretiert werden können, ist Bestandteil der umfangreichen evidenzbasierten Validierung, die in dieser Dissertation dargestellt und diskutiert wird (vgl. Kapitel 9, 11 und 12).

TEIL C: EVIDENZBASIERTE VALIDIERUNG

Bei der evidenzbasierten Validierung soll aus verschiedenen Quellen Evidenz gesammelt werden für eine valide Interpretation der Ergebnisse des WBK-T2 hinsichtlich der sozioökonomischen Facette der WBK. Die evidenzbasierte Validierung für das Konstrukt WBK-soek erfolgt entlang dreier ausgewählter Evidenzquellen aus den *Standards* (AERA et al., 2014) (vgl. Tabelle 8-7).

Für die Validierung des *Testinhalts* wurden qualitative Analysen mittels Leitfaden-Interviews und kognitiven Interviews durchgeführt (vgl. Kapitel 9). Die Validierung der *internen Struktur* erfolgte durch Analysen der probabilistischen Testtheorie (IRT) mit dem Partial-Credit-Rasch-Modell (vgl. Kapitel 11). Für die Validierung der Beziehung zu externen Merkmalen wurden Mittelwertanalysen und Korrelationsanalysen gerechnet (vgl. Kapitel 12).

Tabelle 8-7. Evidenzquellen, Ziele und Vorgehen für die Validierung der WBK-soek

Evidenzquelle	Ziel	Vorgehen	Kapitel
Testinhalt (<i>evidence based on test content</i>)	Inhaltsbereiche (Problemsituationen) für den WBK-T2 bestimmen	Domänenanalyse des gesellschaftlichen/wirtschaftlichen Lebensbereichs im WBK-Modell	4.3.2, 8.2.1
	Repräsentative und adäquate Inhaltselemente (Items) für den WBK-T2 bestimmen	Fachdidaktische Analyse der Problemsituationen	4.3.2, 8.2.2
	Repräsentative und operationalisierbare Kognitionsprozesse für den WBK-T2 festlegen	Fachdidaktische Analyse der Kognitionsprozesse im WBK-Modell	4.4, 8.3.3
	Bestehende Items des WBK-T1 evaluieren und ggf. revidieren, neue Items konstruieren	Fachdidaktische Kriterien und psychometrische Indikatoren Konstruktionsschema für Itemtypen	7.2.2, 8.3.3
	Reliable Kodierung der Item-Antworten	Kodierungsmanual für Items (<i>scoring</i>); Volumen für Doppelkodierung, Kennwerte für Inter-Rater-Reliabilität (Cohens κ , Spearmans ρ)	8.4, 10.2.1
	Expertenbeurteilung der Items	Leitfaden-Interviews mit Experten (N = 6)	9.1
	Zielgruppengerechte Gestaltung des Testhefts sowie Formulierung von Instruktionen und Items	Qualitative Pilotierung: kognitive Interviews mit Probanden (<i>cognitive pretesting</i>) (N = 3) Quantitative Pilotierung: Piloterhebung (<i>standard pretesting</i>) (N = 32)	9.2 10.1.1
Interne Struktur (<i>evidence based on internal structure</i>)	Faktoriellen Struktur des WBK-T2 bzw. Dimensionalität der WBK-soek	IRT-Analysen mit Partial-Credit-Rasch-Modell und MRCML-Modell (N = 375, m = 32): Spezifikation von konkurrierenden Messmodellen	11.1
	Schätzung und Skalierung der Item- und Personenparameter	IRT-Analysen mit Partial-Credit-Rasch-Modell (N = 375, m = 32): Schätzung der Itemparameter und Personenparameter mit MML-Methode	11.2
	Schätz- bzw. Messgenauigkeit für WBK-T2	Testinformationskurve, Standardfehler der Personenparameter Probabilistische und klassische Reliabilitäts-Koeffizienten (EAP/PV, WLE, Cronbachs α)	11.3
	Gültigkeit des Testmodells bzw. Item-Homogenität im WBK-T2	IRT-Analysen mit Partial-Credit-Rasch-Modell (N = 375, m = 32): Item-Infits, Schwellenparameter, DIF bzgl. Personengruppen	11.4
	Verteilung der Testwerte und Qualität der Items im WBK-T2	Klassische Itemanalysen (N = 375, m = 32): Item-Schwierigkeit, Item-Trennschärfe	11.5

Tabelle 8-7. Evidenzquellen, Ziele und Vorgehen für die Validierung der WBK-soek

Evidenzquelle	Ziel	Vorgehen	Kapitel
	Graduierung des WBK-T2	Beurteilung der Items durch Experten (N = 5, m = 32): Ratingschema für schwierigkeitsbestimmende Merkmale Quantitative Bestimmung von Kompetenzschwellen: Regressionsanalyse von Item-Schwierigkeit auf Itemrating (m = 32) Qualitative Beschreibung von Kompetenzniveaus: Kompetenzniveaumodell	11.6
Beziehung zu externen Merkmalen	Unterschiede in WBK-soek zwischen Zielgruppe und Kontrastgruppe	Mittelwertanalysen (t-Tests) nach Schwerpunktachse: Nicht-WuR und WuR (N = 375, m = 32)	12.1
<i>(evidence based on relations to other variables)</i>	Unterschiede in WBK-soek zwischen diversen Zielsubgruppen	Mittelwertanalysen (t-Tests) nach Schulform/Ausbildung: kfmBS und gymMS (N = 792, m = 9) Mittelwertanalysen (t-Tests) nach Politiksystem/Land: gymDE und gymCH (N = 505, m = 18), kfmDE und kfmCH (N = 555, m = 13)	12.2
	Zusammenhang zwischen WBK-soek und externen Leistungs- und Interessensmerkmalen	Bivariate Korrelationsanalysen mit WBK-soek, Schulnoten WuR und Deutsch, Sachinteresse (N = 375)	12.3

Quelle: Eigene Entwicklung und Darstellung i.A.a. AERA et al. (2014).

9 Validierung des Testinhalts des WBK-T2

In diesem Kapitel wird der Leitfrage (7) nachgegangen (vgl. Kapitel 1.2):

(7) Inwiefern sind die Inhaltsbereiche und Inhaltselemente des WBK-T2 für die sozioökonomische Facette der WBK repräsentativ und adäquat?

Mit dieser Leitfrage einher gehen die Annahme I und die Annahme II zum Testinhalt des WBK-T2 (vgl. Kapitel 8.1.4.a). Evidenz zum Testinhalt kann auf zwei Arten gesammelt werden (AERA et al., 2014, S. 14f.) (vgl. Kapitel 6.1.2): Zum einen durch eine detaillierte Beschreibung des Konstrukts und der Inhaltsspezifikationen des Tests; zum anderen durch die kriteriengeleitete Einschätzung des Testinhalts durch Experten.

Unter Berücksichtigung dieser Leitfrage wurde ausgehend von der umfangreichen Domänenanalyse (vgl. Kapitel 4.3.2.c) detaillierte Inhaltsspezifikationen für den WBK-T2 vorgenommen (vgl. Kapitel 8.2), auf die an dieser Stelle nicht erneut eingegangen wird. Des Weiteren wurden zur Beantwortung dieser Leitfrage semi-strukturierte Interviews mit Experten und kognitive Interviews mit Probanden durchgeführt und qualitativ ausgewertet. Bei den Experten-Interviews soll der Testinhalt an den in den *Standards* genannten Kriterien Repräsentativität/Relevanz und Adäquanz geprüft werden (AERA et al., 2014, S. 14f.). Bei den kognitiven Interviews ist das Testheft auf sprachliche und formale Mängel zu prüfen.

Der Aufbau dieses Kapitels ist wie folgt: Als erstes werden die Erkenntnisse aus den Experten-Interviews (Kapitel 9.1) dargelegt, als zweites jene aus den kognitiven Interviews (Kapitel 9.2).

9.1 Experten-Interviews

Das Ziel der Experten-Interviews war es, die Items des WBK-T2 nach Kriterien *Relevanz* und *Adäquanz* beurteilen zu lassen. Für die Experten-Interviews wurden diese Kriterien spezifiziert und zu jedem Kriterium Leitfragen formuliert (vgl. Tabelle 9-1): domänenspezifische Relevanz (bezüglich Problemsituation), fachliche Adäquanz (fachwissenschaftlich, fachdidaktisch), sprachliche Adäquanz, gruppenbezogene Adäquanz (vgl. AERA et al., 2014, S. 14f.).

Die Experten-Interviews wurden von der Autorin als semi-strukturierte Interviews (vgl. Döring & Bortz, 2016, S. 372ff., S. 375f.) mit insgesamt sechs Expert*innen durchgeführt. Alle sechs Expert*innen arbeiteten als Fachlehrer*innen WuR an gymnasialen Maturitätsschulen bzw. kaufmännischen Berufsmaturitätsschulen, drei davon waren zudem als Fachdidaktiker*innen WuR an universitären bzw. pädagogischen Hochschulen tätig.

Tabelle 9-1. Kriterien und Interview-Leitfragen für Validierung der WBK-T2 Items

Kriterium	Leitfragen
Domänenspezifische Relevanz	▪ Inwiefern sind die Problemsituationen des WBK-T2 für die sozioökonomische Facette der WBK relevant? ▪ Inwiefern ist die Problemsituation durch die Items (Inhaltselemente) repräsentativ abgebildet? ▪ Welche Items (Inhaltselemente) müssten ggf. weggelassen bzw. aufgenommen werden?
Fachliche Adäquanz: fachwissenschaftliche Korrektheit	▪ Ist das Item fachwissenschaftlich korrekt formuliert? ▪ Welche Korrekturen müssten ggf. vorgenommen werden? ▪ Bei gebundenem Antwortformat (Selected-Response-Items): Ist die korrekte Antwort (X) fachlich korrekt? Sind die drei Antwortmöglichkeiten fachlich inkorrekt und fachdidaktisch sinnvoll? ▪ Bei freiem Antwortformat (Constructed-Response-Items): Sind die Musterlösungen im Kodierungsmanual fachlich korrekt und vollständig?
Fachliche Adäquanz: fachdidaktische Passung	▪ Ist das Item fachdidaktisch passend konstruiert, v.a. bezüglich des Inhalts und des Formats? ▪ Welches Format wäre für das Inhaltselement ggf. passender?
Sprachliche Adäquanz: allgemeine Verständlichkeit	▪ Ist das Item sprachlich verständlich formuliert? ▪ Welche sprachlichen Anpassungen müssten ggf. vorgenommen werden?
Gruppenbezogene Adäquanz	▪ Inwiefern ist das Item für die Zielgruppe adäquat? (Gymnasiale Maturitätsschule im Schwerpunktfach WuR; Berufsmaturitätsschule im Schwerpunktfach WuR; kaufmännische Berufsfachschule) ▪ Inwiefern ist das Item für die Kontrastgruppe adäquat? (Gymnasiale Maturitätsschule ohne Schwerpunktfach WuR, d.h. nur Einführungsfach WuR; Berufsmaturitätsschule ohne Schwerpunktfach WuR; andere Berufsfachschulen)

Die Ergebnisse der Experten-Interviews sind in Tabelle 9-2 zusammengefasst. Die domänenspezifische Repräsentativität befanden die Expert*innen bei 16 Items als nicht gegeben und empfahlen diese Items auszuschliessen. Die sprachliche Adäquanz wurde von Expert*innen als grösstenteils erfüllt beurteilt, empfahlen jedoch für 17 Items konkrete Modifikationen: Bei 7 Items sollte die Formulierung von Frage bzw. Instruktion vereinfacht werden, bei 10 Items sollten Fremdwörter oder Fachbegriffe durch Synonyme ersetzt werden.

Die fachliche Adäquanz beurteilten die Expert*innen ebenfalls als grösstenteils erfüllt, aber empfahlen bei 14 Items konkrete Modifikationen: Bei 3 Items sollten fachlich unpräzise Begriffe ersetzt werden, um Irritation und vorzeichenverkehrte Diskrimination zu vermeiden. Bei den Items des Typus IV sollte die Tabelle als inhaltlich angereicherte SWOT-Matrix statt als leere Entscheidungsmatrix dargestellt werden, um die formelle Komplexität zu reduzieren und auf die inhaltliche Komplexität zu fokussieren. Bei den Items des Typus VI sollte das Diagramm durch weitere Elemente und Beziehungen ergänzt werden, um die inhaltliche Komplexität zu erhöhen.

Tabelle 9-2. Ergebnisse der Experten-Interviews zum Inhalt des WBK-T2

Kriterium	Modifikation, Begründung	Itemtypus, Item-Nr.	Item #
Domänenspezifische Relevanz	Items ausschliessen, nicht modifizieren	AHV_3, AHV_4, AHV_9, AHV_10, ENE_3, ENE_5, ENE_6, ENE_7, ENE_10, STA_2, MAN_3, MAN_6, MAN_7, LAN_4, STEU_2, STEU_7	16
Fachliche Adäquanz: Fachwissenschaftliche Korrektheit	Frage, Antwortoptionen, Diagramm: fachlich unpräzise Begriffe ersetzen → reduziert Irritation und vorzeichenverkehrte Diskrimination	ENE_12n, MAN_2, MAN_10n	3
Fachliche Adäquanz: Fachdidaktische Passung	Tabelle: formell anpassen (SWOT-Matrix statt Entscheidungsmatrix), inhaltlich anreichern → reduziert formelle Komplexität, fokussiert inhaltliche Komplexität Instruktion: gemäss Tabelle anpassen	Alle Items des Typus IV: AHV_12m, ENE_11m, STA_8m, LAN_8m, STEU_9m	5
	Diagramm: inhaltlich erweitern mit Elementen/Beziehungen → erhöht inhaltliche Komplexität Instruktion: oberhalb des Diagramms platzieren → verbessert Orientierung	Alle Items des Typus VI: AHV_13n, ENE_12n, STA_10n, MAN_10n, LAN_9n, STEU_10n	6
Sprachliche Adäquanz: Allgemeine Verständlichkeit	Frage: als Aussage formulieren	AHV_2	1
	Instruktion: Leitfrage streichen, Aussage direkt formulieren → präzisiert Fragestellung	Alle Items des Typus II: AHV_8, ENE_8, STA_5, MAN_5, LAN_5, STEU_6	6
	Frage, Antwortoptionen: Fachbegriffe und Fremdwörter durch Synonyme ersetzen	ENE_4m, STA_6m, MAN_8, MAN_9n, STEU_1m	5
	→ reduziert Irritation und Nichtbearbeiten	Alle Items des Typus IV: AHV_12m, ENE_11m, STA_8m, LAN_8m, STEU_9m	5

Hinweis: N = 6 Expert*innen.

Quelle: Ackermann (2018a).

Der Ausschluss der 16 Items sowie die sprachlichen und fachlichen Modifikationen der insgesamt 26 Items wurden in der „Dokumentation der Entwicklung und Revision des WBK-Tests“ (Ackermann, 2018a) festgehalten und bei der Testrevision für den WBK-T2 berücksichtigt (vgl. Kapitel 7.2.2 und 8.3.3).

Die gruppenbezogene Adäquanz wurde von den Expert*innen je nach adressierter Zielgruppe und Kontrastgruppe unterschiedlich beurteilt. Für die Zielgruppe „Schüler*innen im Schwerpunktfach WuR an gymnasialen Maturitätsschulen“ wurden nahezu alle Items als adäquat beurteilt, für die Kontrastgruppe „Schüler*innen in anderen Schwerpunktfächern („nur“ im Einführungsfach WuR) an gymnasialen Maturitätsschulen“ jedoch nur etwa ein Drittel der Items. Dieses Ergebnis spricht dafür,

die Stichprobe des Projekts WBKgym in Abhängigkeit des Schwerpunktfachs auszuwählen, so dass für die Analysen die Gruppe WuR als Zielgruppe (*focal group*) und die Gruppe Nicht-WuR als Referenzgruppe (*reference group*) fungieren kann. Annahme I zur Repräsentativität der Items und Annahme II zur Adäquanz der Items kann nun *nach* der Testrevision für den WBK-T2 als erfüllt betrachtet werden.

9.2 Kognitive Interviews (Cognitive Pretesting)

Das Ziel der kognitiven Interviews war es, die *sprachliche Klarheit* (Formulierung von Einleitungstext und Items, Fremdwörter, Fachbegriffe) und *formelle Klarheit* (Layout, Instruktionen, Antwortstrukturierung/Antwortraum) des WBK-T2 zu überprüfen und ggf. anzupassen (vgl. Tabelle 9-3). Nicht beabsichtigte sprachliche und formelle Hürden können bei den Testpersonen Irritation verursachen und die Testleistung negativ beeinflussen.

Die kognitiven Interviews erfolgten nach der Methode des *Concurrent Think Aloud* (lautes Denken) (vgl. Döring & Bortz, 2016, S. 370f.) mit drei Schüler*innen einer kaufmännischen Handelsmittelschule. Die Testpersonen wurden aufgefordert, die allgemeinen Instruktionen, den Einleitungstext und die Itemstämme im Testheft (vgl. Anhang Abbildung A- 6) laut und langsam vorzulesen und ihre spontanen Gedanken dazu laut zu äussern. Die Autorin protokolliert die Äusserungen der Testpersonen und stellte ihnen Rückfragen.

Die Ergebnisse der kognitiven Interviews sind in Tabelle 9-3 zusammengefasst. Zur formellen Klarheit bemerkten die Proband*innen, dass das Layout des Testhefts (Deckblatt, Antwortstrukturierung) und die allgemeinen Hinweise zu Beginn des Testhefts optimiert werden sollten. Bei der sprachlichen Klarheit zeigte das Zögern und Nachfragen der Proband*innen, dass bestimmte Fremdwörter und Fachbegriffe unbekannt waren und in entsprechende Synonyme überführt werden sollten. Diese Begriffe waren deckungsgleich mit jenen, die auch die Expert*innen moniert hatten (vgl. Kapitel 9.1). Die Bemerkungen der Schüler*innen flossen sowohl in die gestalterische Überarbeitung des Testhefts als auch in die inhaltliche Modifikation der Items ein.

Tabelle 9-3. Ergebnisse der kognitiven Interviews zum Testheft des WBK-T2

Kriterium	Modifikation, Begründung	Testteil, Itemtypus, Item-Nr.
Formelle Klarheit: Layout	Anonymität: Klebeetikette auf Deckblatt mit kryptischer ID statt mit SuS-Name → Klassenliste mit Name und ID für Testleitung	Deckblatt
	Antwortstrukturierung: Quadrate statt Kreise zum Ankreuzen	Typus I, II, IV
Formelle Klarheit: Instruktionen	Testaufbau erläutern, Beispielitems weglassen → reduziert Lesezeit	Allgemeine Hinweise
Formelle Klarheit: Antwortstrukturierung, Antwortraum	---	---
Sprachliche Klarheit: Text-/Item-Formulierung; Fremdwörter, Fachbegriffe	„Status quo“ → „aktuelle Situation“	Typus IV: AHV_12m, ENE_11m, STA_8m, LAN_8m, STEU_9m
	„exzessiv“ → „überhöht“	MAN Einleitungstext, MAN_8
	„intrinsische Motivation“ → „intrinsische (d.h. die innere) Motivation“	MAN_9n
	„Fiskalpolitik“ → „Finanzpolitik“	STA Einleitungstext, STA_6m
	„Internalisierung der externen Umweltkosten“ → „Abwälzung der Umweltkosten auf den Verursacher“	ENE_4m

Hinweis: N = 3 Schüler*innen.

Quelle: Ackermann (2018a).

9.3 Zwischenfazit

In diesem Kapitel wurde Evidenz für den Testinhalt des WBK-T2 gesammelt. Als Erstes wurden Leitfaden-Interviews mit Expert*innen geführt. Als Kriterien für die Experten-Interviews dienten domänenspezifische Repräsentativität (bezüglich Problemsituation), fachliche Adäquanz (fachwissenschaftlich, fachdidaktisch), sprachliche Adäquanz und gruppenbezogene Adäquanz. Die Expert*innen beurteilten die Items des WBK-T2 als mehrheitlich sprachlich verständlich formuliert, fachwissenschaftlich korrekt und fachdidaktisch passend. Die Ergebnisse der Experten-Interviews flossen in der Form von Modifikation oder Ausschluss von Items direkt in die Revision des WBK-T1 ein. Zudem beurteilten die Expert*innen die Items des WBK-T2 für die Zielgruppe als vollumfänglich adäquat, für die Kontrastgruppe als nur teilweise adäquat.

Als Zweites wurden kognitive Interviews mit Schüler*innen geführt. Die Kriterien für die kognitiven Interviews mit Schüler*innen waren sprachliche Klarheit (Fremdwörter, Fachbegriffe) und formelle Klarheit (Layout, Instruktion, Frageformulierung, Antwortstrukturierung/Antwortraum). Die Schüler*innen haben mit ihrem „lauten Denken“ Hinweise auf formelle und sprachliche Mängel im WBK-T2. Auch diese Ergebnisse fanden Eingang in die Testrevision. Allerdings wurden die kognitiven

Interviews nicht derart durchgeführt und dokumentiert, als dass sie als Evidenzquelle für die Validierung der Denkprozesse ausreichen (vgl. Kapitel 5.2). Deshalb wäre es für die weiterführende Validierung durchaus aufschlussreich, in einem Folgeprojekt den Validierungsaspekt Denkprozesse mittels kognitiver Interviews zu untersuchen.

Die Ergebnisse dieses Kapitels liefern Evidenz dafür, dass durch die Revision des WBK-T1 bzw. mit der Entwicklung des WBK-T2 eine repräsentative und adäquate Operationalisierung des Konstrukts WBK-soek gelungen ist.

10 Durchführung und Auswertung des WBK-T2

Dieses Kapitel beschreibt die Datenerhebung, insbesondere Erhebungsdesign und Stichprobe des Forschungsprojekts WBKgym (Kapitel 10.1) sowie die Datenauswertung, insbesondere Kodierungs- und Auswertungsverfahren für den WBK-T2 (Kapitel 10.2).

10.1 Datenerhebung

In diesem Kapitel werden die Piloterhebung (Kapitel 10.1.1) und die Haupterhebung (Kapitel 10.1.2) geschildert sowie die Stichprobe des Projekts WBKgym beschrieben (Kapitel 10.1.3).

Für die Datenerhebung wurde eine Stichprobe von Schüler*innen an gymnasialen Maturitätsschulen der deutschsprachigen Schweiz ausgewählt. Diese Auswahl ist nicht beliebig, sondern folgt einer inneren Forschungslogik. Erstens bezieht sich das WBK-Modell auf die ganze Sekundarstufe II, also auf Schüler*innen an allgemeinbildenden und berufsbildenden Schulen (vgl. Kapitel 4.2). Zweitens legen die Spezifikationen des WBK-T2 als Zielgruppe Schüler*innen am Ende der Sekundarstufe II, also explizit auch Gymnasiast*innen, fest (vgl. Kapitel 8.1.3). Drittens war die WBK bis dahin erst mit dem reduzierten WBK-T1 und nur bei kaufmännischen Lernenden in der Deutschschweiz erfasst worden (Eberle et al., 2016; Holtsch & Eberle, 2018; Schumann et al., 2017) (vgl. Kapitel 7.1). Eine Validierung der WBK-soek mit dem WBK-T2 bei Gymnasiast*innen ist daher nicht nur sinnvoll sondern auch notwendig.

10.1.1 Piloterhebung (Standard-Pretesting)

Eine Piloterhebung (Standard-Pretesting) wurde im Dezember 2016 in einem deutschsprachigen Kanton der Westschweiz (Kanton Freiburg) durchgeführt. Die Stichprobe umfasste 32 Schüler*innen aus zwei Klassen der kaufmännischen Berufsmaturität. Das Ziel der Piloterhebung war zum einen den Erhebungsablauf (Moderation, Zeitbedarf, etc.) durchzuspielen und für die Haupterhebung zu vereinheitlichen. Zum andern ging es darum, die Erhebungsmaterialien (Erhebungsmanual, Erhebungsprotokoll, Klassenliste, Einverständniserklärung, Testheft, Fragebogenheft) zu erproben und ggf. für die Haupterhebung anzupassen.

Dank der Piloterhebung konnte der effektive Zeitbedarf für die einzelnen Phasen⁷⁸ der Erhebung abgeschätzt und für die Haupterhebung geplant werden. Ausgehend von den Rückfragen der Testpersonen liessen sich die schriftlichen Instruktionen in den Heften präzisieren. Schliesslich konnte aufgrund unklarer oder zweideutiger Antworten der Testpersonen, die Gestaltung der Hefte (Fragereihenfolge, Antwortstrukturierung) optimiert werden.

⁷⁸ Begrüssung und Einführung, Austeilung und Einsammeln der Hefte, Ausfüllen des Tests und des Fragebogens, Abschluss und Dank.

10.1.2 Haupterhebung/Erhebungsdesign

Die Haupterhebung fand von Januar bis März 2017 im deutschsprachigen Kanton St.Gallen statt (vgl. Tabelle 10-1); sie verteilte sich auf 18 Erhebungstermine (vgl. Anhang Abbildung A- 2). Die Stichprobe umfasst 375 Schüler*innen aus insgesamt 24 Klassen der gymnasialen Maturität (vgl. Kapitel 10.1.3). Als Instrumente wurden im *Paper-and-Pencil*-Verfahren der revidierte WBK-Test (WBK-T2) (Ackermann, 2018c) sowie ein neu entwickelte Fragebogen zu Individual- und Kontextfaktoren (WBK-Q1) (Ackermann, 2018b) eingesetzt. Der WBK-T2 wurde bereits oben ausführlich beschrieben (vgl. Kapitel 8). Allen Testpersonen wurde das gleiche Testheft des WBK-T2 mit vier Problemsituationen und 32 Items vorgelegt.⁷⁹ Der Fragebogen erfasst individuelle und kontextuelle Faktoren der Testpersonen mit insgesamt 79 Fragen.

- Leistungsmerkmale: Schulnoten in den Fächern Deutsch, Mathematik, WuR;
- Interessensmerkmale: Interesse an sozioökonomischen Problemsituationen;
- Unterrichtsmerkmale: wahrgenommene Unterrichtsqualität im Fach WuR;
- Sozialisationsmerkmale: soziodemografische, soziokulturelle und sozioökonomische Merkmale;
- Verhaltensmerkmale: politisches Interesse und Engagement.

Ein Erhebungstermin dauerte pro Klasse ca. 120 Minuten (vgl. Tabelle 10-1): 15 Minuten für Einführung, 60 Minuten für den WBK-T2, 10 Minuten Pause, 20 Minuten für den WBK-Q1, 5 Minuten für Dank und Verabschiedung sowie insgesamt 10 Minuten für Administration (Austeilen und Einsammeln der Testhefte und Fragehefte). Bei der Einführung wurde den Testpersonen eine schriftliche *Einverständniserklärung* vorgelegt, die sie über die Ziele des Forschungsprojekts, den Ablauf der Erhebung und die Konsequenzen der Testung (Vertraulichkeit, Anonymität etc.) informierte und auf der sie mit ihrer Unterschrift ihr Einverständnis zur Teilnahme bestätigten (vgl. Anhang Abbildung A- 5). Auf eine Einverständniserklärung seitens der Eltern wurde aufgrund des Alters der Testpersonen ($\pm$ 18 Jahre) verzichtet. Entsprechend wurde auf ihre Volljährigkeit (Art. 14 Zivilgesetzbuch) und behelfsmässig auf ihre situationsbezogene Urteilsfähigkeit (Art. 16 Zivilgesetzbuch) abgestützt (vgl. Hausheer & Aepli-Müller, 2016, S. 44ff.).

⁷⁹ Beim Verbundprojekt CoBALIT wurde ein rotiertes *booklet design* angewendet (Eberle et al., 2016, S. 105). *Booklet Designs* eignen sich für *Large Scale Assessments* und sind äusserst komplex (vgl. Frey et al., 2009).

Tabelle 10-1. Erhebungsdesign des Projekts WBKgym

Zeitraum	Januar bis März 2017
Region	Deutscheschweiz, Kanton St. Gallen
Schulstufe	Sekundarstufe II (11. und 12. Klassenstufe)
Schulform	Gymnasiale Maturitätsschulen, total 4 Kantonsschulen
Stichprobe	▪ Insgesamt: N = 375 ▪ 11. Schuljahr (N = 216, 58 %): Kontrastgruppe Nicht-WuR, obligatorische Einführung WuR und anderes Schwerpunktfach ▪ 12. Schuljahr (N = 159, 42 %): Zielgruppe WuR, obligatorische Einführung WuR und Schwerpunktfach WuR
Instrumente	▪ WBK-Test (WBK-T2): 32 Items zu 4 Problemsituationen, 60 min, Power-Test ▪ Fragebogen (WBK-Q1): 79 Items zu Kontext- und Individualfaktoren, 20 min
Verfahren	Paper-and-Pencil
Dauer	ca. 120 Minuten pro Klasse
Administration	Projektleiterin, z.T. geschulte Erhebungsleiter

Das *Erhebungsmanual* diente quasi als Drehbuch, um bei allen Klassen einen standardisierten Ablauf der Erhebung zu gewährleisten (vgl. Anhang Abbildung A- 3). Im *Erhebungsprotokoll* wurden die Eckdaten jeder Erhebung notiert und spezielle Vorkommnisse dokumentiert (vgl. Anhang Abbildung A- 4). Bei drei Erhebungsterminen mit Doppelklassen waren zur Unterstützung externe Erhebungsleiter dabei, die vor ihrem Einsatz eine ca. einstündige Schulung durchlaufen hatten.

10.1.3 Stichprobe

Bei einer Teilerhebung (auch: Stichprobenerhebung) werden – im Gegensatz zu einer Vollerhebung – nicht die alle Personen der Population, sondern ausgewählte Personen der Population untersucht. Die Zielpopulation (*target population*) sind alle Personen über die eine Aussage getroffen werden soll und die Auswahlpopulation (*frame population*) umfasst diejenigen Personen, die eine Chance haben in die Stichprobe zu gelangen (vgl. Döring & Bortz, 2016, S. 292ff.). Die Ausschöpfungsrate gibt das Verhältnis zwischen erreichter und geplanter Stichprobe an (vgl. Döring & Bortz, 2016, S. 296).

Die Repräsentativität (*sample representativeness*) beschreibt das „qualitative“ Verhältnis zwischen Stichprobe und Population: Eine repräsentative Stichprobe ist ein perfektes „Miniaturabbild“ der Population, d.h. die Merkmale und Effekte der Population zeigen sich unverzerrt in der Stichprobe (vgl. Döring & Bortz, 2016, S. 297ff.). Es kann zwischen merkmalspezifisch-repräsentativen und global-repräsentativen Stichproben unterschieden werden. Eine *merkmalspezifisch-repräsentative Stichprobe* liegt vor, wenn die Zusammensetzung der Stichprobe hinsichtlich *einiger* relevanter Merkmale jener der Population entspricht. Dies wird üblicherweise über einen Quotenplan erreicht, der die forschungsrelevanten Merkmale der

Population berücksichtigt. Jedoch muss bei der Interpretation der Ergebnisse bedacht werden, dass die Stichprobe hinsichtlich nicht berücksichtigter Merkmale verzerrt sein kann. Eine *global-repräsentative Stichprobe* liegt vor, wenn die Zusammensetzung der Stichprobe hinsichtlich *aller* Merkmale derjenigen der Population entspricht. Dies kann nur über eine probabilistische Stichprobenziehung (blindes Zufallsverfahren) und einer Mindeststichprobengrösse erreicht werden.

Eine repräsentative Stichprobe ist Voraussetzung für unverzerrte inferenzstatistische Kennwerte (z.B. t-Statistik, Korrelationskoeffizient). Ein statistischer Kennwert schätzt einen Populationsparameter *erwartungstreu*, wenn der Erwartungswert der Stichprobenverteilung gleich dem Populationsparameter ist (Bortz & Schuster, 2010, S. 89, S. 169f.).

10.1.3.a Ziehung der Stichprobe

Im Rahmen des Projekts WBKgym war es aus forschungsökonomischen (personelle und finanzielle Ressourcen) und politischen (Genehmigung der kantonalen Erziehungsdepartemente, Einwilligung der Schuldirektionen und Lehrpersonen) Gründen nicht möglich, eine repräsentative Stichprobe für die gesamte Deutschschweiz zu erheben. Daher wurde die Datenerhebung auf einen Deutschschweizer Kanton beschränkt, nämlich den Kanton St.Gallen.

Der Kanton St.Gallen hat einen kantonalen Lehrplan für die gymnasialen Maturitätsschulen erlassen, der die Bildungsziele und Stundentafeln einheitlich regelt. Ein kantonaler Lehrplan erlaubt es, das potenzielle curriculare Vorwissen der Stichprobe für die statistischen Analysen als gleichwertig anzunehmen. In den meisten Deutschschweizer Kanonen gibt es hingegen schulische Lehrpläne, wie z.B. in den Kantonen Luzern und Zürich. Der Kanton St.Gallen hat zudem einen „wirtschaftsfreundlichen“ Ruf, denn er hat für das Einführungsfach WuR insgesamt vier Jahreswochenstunden vorgesehen. In den meisten Deutschschweizer Kantonen sind es zwei oder drei Jahreswochenstunden, wie z.B. in den Kantonen Basel Stadt, Bern, Luzern und Freiburg, hingegen sind es im Kanton Zürich, je nach Kantonsschule, ein bis zwei Jahreswochenstunden (Eberle & Brüggengbrock, 2013). Gemäss der kantonalen Stundentafel wird das Fach WuR im Kanton St.Gallen wie folgt unterrichtet (St.Gallen, 2006, S. 10) (vgl. Tabelle 10-2): im 10. und 11. Schuljahr als obligatorische Einführung mit insgesamt vier Jahreswochenstunden, vom 9. bis 12. Schuljahr als Schwerpunktfach mit 15 Jahreswochenstunden und im 12. Schuljahr als Ergänzungsfach mit vier Jahreswochenstunden. Nicht zuletzt bestanden zu Schulleiter*innen und Lehrer*innen im Kanton St.Gallen persönliche Kontakte, die sich bei der Erhebungsorganisation und -administration als sehr nützlich und hilfreich erwiesen.

Die Zielpopulation (*target population*) des Projekts WBKgym besteht aus den Schüler*innen an gymnasialen Maturitätsschulen im Kanton St.Gallen, die sich im Schuljahr 2016/2017 im letzten Semester des Faches WuR befinden. Sie lassen sich in zwei Gruppen einteilen (vgl. Kapitel 8.1.3): Die Zielgruppe WuR ist im 12. Schuljahr und besucht das wählbare Schwerpunktfach „Wirtschaft &

Recht“. Die Kontrastgruppe Nicht-WuR ist im 11. Schuljahr, besucht das obligatorische Einführungsfach „Wirtschaft & Recht“ und belegt ein anderes Schwerpunktfach wie z.B. Biologie und Chemie (N), Physik und Anwendungen der Mathematik (P), Latein (L), Italienisch (I), Spanisch (S), Bildnerisches Gestalten (B) oder Musik (M) (St.Gallen, 2006).

Die Auswahlpopulation (*frame population*) hingegen besteht aus den Schüler*innen der oben genannten Fächer und Schuljahre an den fünf staatlichen Kantonsschulen des Kantons St.Gallen (St.Gallen, 2018): am Burggraben St.Gallen, Heerbrugg, Sargans, Wattwil, Wil. Die staatlich anerkannten privaten Mittelschulen des Kantons St.Gallen (Gymnasium Friedberg, Gymnasium Untere Waid) wurden aufgrund ihrer geringen Klassen- und Schülerzahl (i.d.R. eine Schulklasse pro Jahrgangsstufe) für das Projekt WBKgym nicht berücksichtigt. An die Schulleitungen der fünf staatlichen Kantonsschulen wurde eine schriftliche Anfrage für die Teilnahme am Projekt WBKgym gestellt (vgl. Anhang Abbildung A- 2).

Für die Stichprobenziehung wurden in einem ersten Schritt die Schulklassen nach dem Quotenverfahren ausgewählt (Quotenstichprobe, *quota sample*) (vgl. Döring & Bortz, 2016, S. 309ff.), wobei die Gruppenzugehörigkeit (WuR, Nicht-WuR) als Quotierungsmerkmal diente. Im Schuljahr 2016/2017 gab es an den staatlichen Kantonsschulen im Kanton St.Gallen in der Gruppe Nicht-WuR 25 Schulklassen (60 % aller Klassen) und in der Gruppe WuR 17 Schulklassen (40 % aller Klassen) (vgl. Tabelle 10-3). Von diesen 42 Klassen wurden 30 in einem zufälligen Verfahren gezogen und davon 24 erreicht.⁸⁰ Die nicht-erreichten 6 Klassen sind einer Pauschalabsage der fünften Kantonsschule geschuldet. Die Ausschöpfungsrate (*response rate*) auf *Klassen-Ebene* beträgt somit 57 % bzw. 80 % (vgl. Tabelle 10-3).

In einem zweiten Schritt wurden die Schulklassen nach ihrer organisatorischen Verfügbarkeit ausgewählt (Gelegenheitsstichprobe, *convenience sample*) (vgl. Döring & Bortz, 2016, S. 305ff.). Dabei spielte u.a. der Stundenplan und die Klassengrösse eine Rolle: Zur Gewährleistung der Durchführungsobjektivität i.S. einer vergleichbaren Testsituation (vgl. Kapitel 5.1.2.a) wurden Klassen bevorzugt mit einer Doppelstunde WuR am Vormittag; für eine ökonomische Testadministration wurden Klassen mit mindestens 10 Schüler*innen bevorzugt. Von den 24 erreichten Schulklassen waren gemäss den Klassenlisten 421 Schüler*innen eingeplant (vgl. Tabelle 10-3). Davon nahmen schliesslich 377 an den Erhebungen teil; 44 Testpersonen waren infolge Krankheit oder J+S-Ausbildung am Erhebungstermin abwesend. Die Ausschöpfungsrate auf *Schüler*innen-Ebene* beträgt somit 43 % bzw. 90 %.

⁸⁰ Auf Klassenebene wurden die effektiven Quoten der Stichprobe (50/50) im Vergleich zum Quotenplan der Grundgesamtheit (60/40) etwas angepasst, da die Klassen der Gruppe WuR im Erhebungsjahr 2017 durchschnittlich kleiner (sog. Halbklassen) waren als jene in der Gruppe Nicht-WuR.

Tabelle 10-2. Stundentafel für Fach WuR im Kanton St.Gallen

Fach	9. Schuljahr (1. Klasse)	10. Schuljahr (2. Klasse)	11. Schuljahr (3. Klasse)	12. Schuljahr (4. Klasse)	total
Einführung WuR	0	2	2	0	4
Schwerpunkt WuR	4	4	3	4	15
Ergänzung WuR	0	0	0	4	4

Hinweis: Angaben in Jahreswochenstunden, beim Schwerpunktfach exklusive obligatorisches Einführungsfach.

Quelle: Erziehungsdirektion des Kantons St.Gallen (2006, S. 10).

Tabelle 10-3. Grundgesamtheit und Stichprobe nach Klassen und Schüler*innen

Ebene		Nicht-WuR	WuR	Gesamt	Ausschöpfungsrate	
Klassen	Grundgesamtheit 2017 (Auswahlpopulation)	25 (60 %)	17 (40 %)	42 (100 %)	100 %	
	Stichprobe gezogen	15 (50 %)	15 (50 %)	30 (100 %)	71 %	100 %
	Stichprobe erreicht	12 (50 %)	12 (50 %)	24 (100 %)	57 %	80 %
SuS	Grundgesamtheit 2011 (Auswahlpopulation)	643 (73 %)	239 (27 %)	882 (100 %)	100 %	
	Stichprobe geplant	243 (58 %)	178 (42 %)	421 (100 %)	48 %	100 %
	Stichprobe erreicht	218 (58 %)	159 (42 %)	377 (100 %)	43 %	90 %

Hinweis: Die Grundgesamtheit auf Klassen-Ebene bezieht sich die Anzahl Klassen an den fünf Kantonsschulen im Jahr 2017. Die Grundgesamtheit auf SuS-Ebene bezieht sich auf die Anzahl Matura-Abschlüsse im Jahr 2011. Nicht-WuR = anderes Schwerpunktfach. WuR = Schwerpunktfach „Wirtschaft & Recht“. SuS = Schülerinnen und Schüler.

Quelle: Bildungsdepartement des Kantons St.Gallen (2016) und Eberle & Brüggengbrock (2013, Anhang, S. 108ff.); eigene Darstellung.

In der erreichten Stichprobe gehören 218 Schüler*innen zur Gruppe Nicht-WuR und 159 Schüler*innen zur Gruppe WuR. Das Gruppenverhältnis der erreichten Stichprobe (58 % Nicht-WuR, 42 % WuR) entspricht in etwa dem Gruppenverhältnis der Grundgesamtheit gemäss Klassenanzahl im Jahr 2017 (Quotenplan), jedoch nicht dem Gruppenverhältnis der Grundgesamtheit gemäss Matura-Abschlüssen im Jahr 2011 (73 % Nicht-WuR, 27 % WuR). Es handelt sich beim Projekt WBKgym also um eine merkmalspezifisch-repräsentative Stichprobe gemäss Quotenplan (vgl. Döring & Bortz, 2016, S. 298f., S. 307).

Dieses gewählte Verfahren der Stichprobenziehung kann dazu führen, dass die Merkmalsausprägungen in der Stichprobe im Vergleich zur Grundgesamtheit verzerrt sind und lässt womöglich nur bedingt Rückschlüsse von der Stichprobe auf die Grundgesamtheit zu (vgl. Bortz & Schuster, 2010, S. 89, S. 169f.; Döring & Bortz, 2016, S. 297f.). Diese Einschränkung muss bei der Interpretation der inferenzstatistisch generierten Ergebnisse berücksichtigt werden (vgl. Kapitel 12). Da die vorliegende Dissertation jedoch auf die evidenzbasierte Validierung abzielt und keine

Generalisierung der Ergebnisse anstrebt, ist diese Restriktion als weniger problematisch zu beurteilen.

10.1.3.b Beschreibung der Stichprobe

Bei der Datenbereinigung wurden zwei Testpersonen der Gruppe Nicht-WuR ausgeschlossen, da sie den WBK-T2 nur zur Hälfte gelöst hatten (vgl. Kapitel 10.2.2); die für die Analysen berücksichtigte Stichprobengrösse beträgt also 375. Von den 375 Testpersonen der Stichprobe sind 62 % weiblich und 38 % männlich; innerhalb der Gruppe WuR sind 47 % weiblich und 54 % männlich (vgl. Tabelle 10-4). Diese Geschlechterverhältnisse der Stichprobe entsprechen ziemlich genau denjenigen der Grundgesamtheit gemäss den Matura-Abschlüssen 2011 im Kanton St.Gallen: 60/40 für alle Gymnasiast*innen, 46/54 für die Gymnasiast*innen im Schwerpunktfach WuR (Eberle & Brüggenschrock, 2013, Anhang, S. 108ff.).

Die Schüler*innen der Stichprobe sind zum Zeitpunkt der Erhebung im Mittel 17.9 Jahre alt (SD = 0.98), wobei die Bandbreite zwischen 14.7 und 21.0 Jahren liegt. 73 % der Schüler*innen sprechen mit ihrer Familie nur Deutsch bzw. Schweizerdeutsch, 23 % Schweizerdeutsch/Deutsch und eine andere Sprache, 5 % nur eine andere Sprache.⁸¹ 90 % der Schüler*innen haben Schweizer Staatsbürgerschaft oder Doppelbürgerschaft, 10 % eine andere Staatsbürgerschaft.

Etwas mehr als die Hälfte der Eltern hat einen Bildungsabschluss auf Tertiärstufe A und B (57 %), rund zwei Fünftel auf der Sekundarstufe I und II (37 %), aber nur gerade 5 % haben einen Abschluss auf Primarstufe.⁸² Etwas weniger als die Hälfte (44 %) der Eltern übt einen akademischen Beruf aus, 18 % haben eine Führungsfunktion in Wirtschaft oder Verwaltung, je 14 % haben technische oder kaufmännische Berufe.⁸³ Der Berufsstatus Eltern, gemessen mit dem Index ISEI-08, liegt in der gesamten Stichprobe durchschnittlich bei 65.86 (SD = 17.15), schwankt aber in einer beträchtlichen Bandbreite von 16.50 bis 88.96.⁸⁴

⁸¹ Die Familiensprache wurde im WBK-Q1 mit einem Item auf einer 3-stufigen Antwortskala abgefragt (1 = nur andere Sprache, 2 = Schweizerdeutsch/Deutsch und andere Sprache, 3 = nur Schweizerdeutsch/Deutsch). Diese Variable wurde später für inferenzstatistische Analysen in zwei Kategorien (0 = Zweitsprache/Nicht-Deutsch, 1 = Erstsprache/Deutsch) rekodiert. Die Stichprobe WBKgy ist hinsichtlich der Familiensprache etwas unterschiedlich zur repräsentativen Stichprobe OEKOMA, wo bei 86 % der Gymnasiast*innen beide Elternteile Schweizerdeutsch/Deutsch sprechen und bei je 7 % ein oder kein Elternteil (Jüttler & Schumann, 2016).

⁸² Der Bildungsabschluss der Eltern wurde im WBK-Q1 mit je einem Item für Mutter und Vater auf einer 8-stufigen Antwortskala (von 1 = Primarschule bis 8 = Universität/ETH) abgefragt. Diese Variable wurde später für inferenzstatistische Analysen in drei (1 = Primarstufe, 2 = Sekundarstufe I und II, 3 = Tertiärstufe) bzw. in zwei Kategorien (0 = nicht-akademisch, 1 = akademisch) rekodiert. Die Stichprobe WBKgy ist hinsichtlich des Bildungsschlusses der Eltern vergleichbar mit der repräsentativen Stichprobe OEKOMA, wo bei 56 % der Gymnasiast*innen die Tertiärstufe, bei 38 % die Sekundarstufe I und II und bei 6 % die Primarstufe der höchste Bildungsabschluss des Vaters ist (Jüttler & Schumann, 2016).

⁸³ Der Beruf der Eltern wurde im WBK-Q1 mit je einem Item für Mutter und Vater mittels Freitext erfasst und nach der Internationalen Klassifikation für Berufe (ISCO-08) in einen 4-stelligen Code überführt (ILO, 2017).

⁸⁴ Für den Berufsstatus der Eltern wurde der ISCO-08 in den Internationalen Index für Berufsstatus (ISEI-08) umgerechnet. Der ISEI-08 hat einen Wertebereich von 0 bis 90 (Ganzeboom, 2010; ILO, 2017).

Tabelle 10-4. Stichprobe nach Sozialisationsmerkmalen

Merkmal	Ausprägung	Nicht-WuR	WuR	Gesamt
Stichprobenumfang	---	216 (100 %)	159 (100 %)	375 (100 %)
Biologisches Geschlecht	Weiblich	160 (74 %)	74 (47 %)	234 (62 %)
	Männlich	56 (26 %)	85 (54 %)	141 (38 %)
Alter (zum Erhebungszeitpunkt)	M (SD)	17.4 (0.82)	18.5 (0.76)	17.9 (0.98)
	Min Max	14.7 21.0	17.2 20.9	14.7 21.0
Familiensprache	Nur andere	5 (2 %)	13 (8 %)	18 (5 %)
	Deutsch und andere	46 (21 %)	39 (25 %)	85 (23 %)
	Nur Deutsch	165 (76 %)	107 (67 %)	272 (73 %)
Herkunftsland	Andere	21 (10 %)	12 (8 %)	33 (9 %)
	Schweiz	195 (90 %)	147 (93 %)	342 (91 %)
Staatsbürgerschaft	nur andere	21 (10 %)	16 (10 %)	37 (10 %)
	Schweiz und andere	46 (21 %)	39 (25 %)	85 (23 %)
	nur Schweiz	149 (69 %)	104 (65 %)	253 (68 %)
Bildungsabschluss Eltern ¹	Primarstufe	7 (3 %)	10 (6 %)	17 (5 %)
	Sekundarstufe I und II	72 (34 %)	67 (43 %)	139 (37 %)
	Tertiärstufe A und B	134 (63 %)	80 (51 %)	214 (57 %)
Berufsstatus Eltern ¹ (nach ISEI-08) ²	M (SD)	---	---	65.86 (17.15)
	Min Max			16.50 88.96

Hinweis: Nicht-WuR = anderes Schwerpunktfach. WuR = Schwerpunktfach Wirtschaft & Recht. Bei der Datenbereinigung wurden 2 Probanden der Gruppe Nicht-WuR ausgeschlossen, da sie den WBK-T2 nur zur Hälfte gelöst hatten.

¹⁾ Elternteil mit dem höheren Wert bzw. einem vorhandenen Wert.

²⁾ International Socio-Economic Index (ISEI-08), berechnet aus dem International Standard Classification of Occupations (ISCO-08), Wertebereich [0,90], wobei 90 der höchste ist (Ganzeboom, 2010).

10.2 Datenauswertung

Dieses Kapitel erläutert das Kodierungsverfahren (Kapitel 10.2.1), die Datenerfassung und Datenbereinigung (Kapitel 10.2.2) sowie die Analyseverfahren (Kapitel 10.2.3).

10.2.1 Kodierungsverfahren

Die Kodierung der Schüler*innen-Antworten des WBK-T2, insbesondere der 15 Constructed-Response-Items, erfolgte durch die Autorin und eine weitere fachwissenschaftlich und fachdidaktisch qualifizierte Person nach einem konventionellen und standardisierten Verfahren (Döring & Bortz, 2016, S. 558f.) (vgl. Abbildung 10-1). Am Anfang stand eine Schulung, bei welcher der zweite Kodierer mit der Zielsetzung des Forschungsprojekts WBKgym, dem Kodierungsablauf und den Kodierungsmaterialien, dem WBK-T2 und seinen Items vertraut gemacht wurden. Integriert in diese Schulung war eine Probekodierung der Testhefte aus der Piloterhebung (vgl. Kapitel 10.1.1) und eine darauf basierende Diskussion inhaltlicher und methodischer Aspekte entlang des umfangreichen Kodierungsmanuals (vgl. Kapitel 8.4).

Zur Gewährleistung der Rater-Übereinstimmung im Verlauf der dreimonatigen Kodierungsphase wurden ca. 17 % der Testhefte doppelt kodiert. Für diese Doppelkodierung wurden insgesamt 65 der 375 Testhefte zufällig ausgewählt: 30 zu Beginn der Kodierungsphase, weitere 26 in der Hälfte der Kodierungsphase. Die Nicht-Übereinstimmungen aus der Doppelkodierung bereinigten die Kodierenden in einer Konsensdiskussion. Sie ergänzten und präzisierten das Kodierungsmanual mit weiteren exemplarischen Antworten für die verschiedenen Codes (0, 1, 2, 888), insbesondere auch mit treffenden Zitaten aus den Testheften für den Code 0, z.B.:

- AHV_12: „Bietet Vorteile für jung und alt [...]“;
- ENE_8: „Die Erde ist in einem kritischen Zustand“;
- ENE_11: „Zwei Fliegen mit einer Klatsche“.

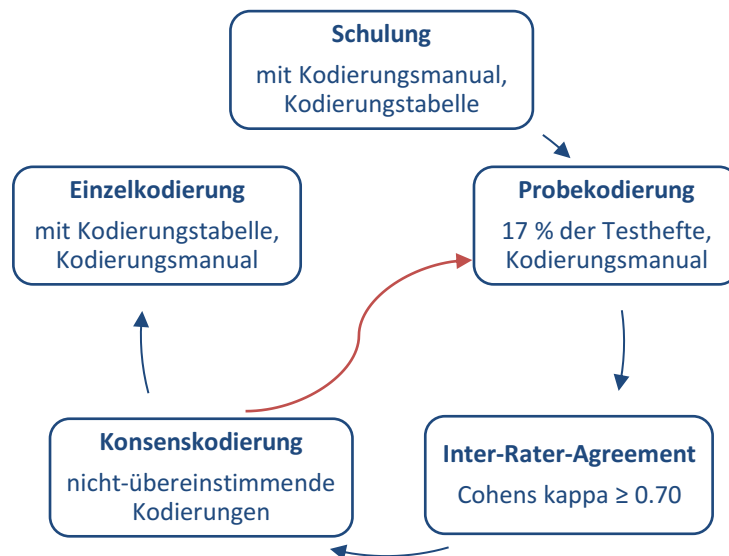


Abbildung 10-1. Kodierungsverfahren für die Constructed-Response-Items des WBK-T2

Quelle: Eigene Darstellung i.A.a. Döring & Bortz (2016, S. 558f.).

Zur Prüfung der Inter-Rater-Reliabilität wurden die zwei Koeffizienten Spearmans ρ^{85} und Cohens κ^{86} (vgl. Döring & Bortz, 2016, S. 346f.; Gwet, 2014) mit dem Programm SPSS Statistics Version 25 berechnet (IBM, 2017). Die Inter-Rater-Reliabilität war für beide Phasen der Doppelkodierung sehr gut (Spearmans $\rho = .88$ bzw. $.86$, Cohens $\kappa = .77$ bzw. $.75$). Infolgedessen wurden die Testhefte gleichmässig zwischen den Kodierenden aufgeteilt und die Antworten in Einzelkodierung kodiert. Für die zufällige Zuteilung der Testhefte wurde ein Kodierungsplan erstellt (vgl. Anhang Abbildung A- 8). Während der dreimonatigen Phase der Einzelkodierung bestand regelmässiger Austausch zwischen der Autorin und dem zweiten Kodierer, um eine einheitliche Handhabung von nicht antizipierten „schwierigen“ Fällen zu gewährleisten.

Die Kodierung der Schüler*innen-Antworten des WBK-Q1 erfolgte durch eine weitere Person. Auch hier gab es vorgängig eine Schulung der Kodiererin und einen regelmässigen Austausch, um inhaltliche und methodische Fragen zu klären. Beide Kodierer*innen mussten vorgängig eine Geheimhaltungs- und Sorgfältigkeitserklärung unterzeichnen (vgl. Anhang Abbildung A- 7).

⁸⁵ Spearmans ρ ist ein Mass für die Rangkorrelation von ordinal-skalierten Variablen. Es kann für die Rater-Übereinstimmung von ordinal-kodierten Antwortkategorien verwendet werden, aber es berücksichtigt keine Zufallsübereinstimmung. Der Wertebereich von Spearmans ρ liegt zwischen -1 und +1, wobei ein Wert von $\approx +1$ eine hohe positive Korrelation indiziert, ein Wert von ≈ -1 eine hohe negative (Döring & Bortz, 2016, S. 346f.).

⁸⁶ Cohens κ ist ein Mass für die Rater-Übereinstimmung bei nominal-skalierten Variablen und berücksichtigt Zufallsübereinstimmungen. Der Wertebereich von Cohens κ liegt zwischen -1 und +1. Nach konventionellen Normen gelten Werte über $.75$ als „sehr gut“, Werte zwischen $.60$ und $.75$ als „gut“ und Werte zwischen $.40$ und $.60$ als „mittelmässig“ bzw. „gerade noch ausreichend“ (Döring & Bortz, 2016, S. 346f.; M. Wirtz & Caspar, 2002).

10.2.2 Datenerfassung und Datenbereinigung

Die Erfassung der Schüler*innen-Antworten aus den physischen Test- und Fragebogenheften erfolgte manuell im Tabellenkalkulationsprogramm MS Excel 2011 für Macintosh. Für ein effizientes und fehlerreduzierendes Verfahren wurde eine Kodierungstabelle programmiert, bei der in jedem Tabellenfeld nur die vorgesehenen Codes eingegeben werden konnten.⁸⁷ Die Kodierungstabelle wurde mit den Daten aus dem Standard-Pretest erprobt und wo nötig angepasst. Für eine übersichtliche und arbeitsteilige Kodierung wurde für jede Schulklasse eine separate Kodierungstabelle erstellt. Nach Abschluss der Kodierung wurden die einzelnen Kodierungstabellen in einem Tabellenblatt zusammengezogen und ins Programm SPSS Statistics Version 25 importiert (IBM, 2017).

10.2.2.a Fehlende Werte

Eine *Missing-Data-Analyse* liefert Informationen über Datenstrukturen, die mit fehlenden Werten in Verbindung stehen. Sie ist stets dann durchzuführen, wenn für eine Person oder eine Variable mehr als 5 % der Werte fehlen (M. A. Wirtz, 2004). Einzelne und/oder *zufällig* auftretende fehlende Werte im Datensatz sind i.d.R. unproblematisch, jedoch können „viele“ und *systematisch* auftretende fehlende Werte zu erheblichen Verzerrungen bei der Datenauswertung und der Ergebnisinterpretation führen. Systematisch meint hierbei, dass die fehlenden Werte gehäuft bei bestimmten Personen/Personengruppen oder Variablen vorkommen (Döring & Bortz, 2016, S. 591). Treten bei einem einzelnen Fall mehr als 50 % fehlende Werte auf und/oder wurde der Test offensichtlich nicht ernsthaft ausgefüllt (z.B. mit „witzigen“ Kommentaren statt Antworten versehen), so wird der ganze Fall aus dem Datensatz ausgeschlossen (Döring & Bortz, 2016, S. 588).

Mit den Datensätzen des WBK-T2 und des WBK-Q1 wurde eine Analyse der fehlenden Werte („missing values analysis“) durchgeführt (vgl. Döring & Bortz, 2016, S. 590f.). Bei der Datenkodierung des WBK-T2 waren für fehlende Schüler*innen-Antworten absichtlich unrealistische Codes vergeben worden (vgl. Fromm, Akremi, & Baur, 2010, S. 22) (vgl. Kapitel 8.4):

- 999 für „Keine Antwort“;
- 888 für „Jux-Antwort“ bei Constructed-Response-Items;
- 777 für „Doppelte Antwort“ bei Selected-Response-Items.

Im Datensatz des WBK-T2 wurden personenbezogene fehlende Werte (*missing cases*) und variablenbezogene fehlende Werte (*missing values*) anhand deskriptivstatistischer Analysen (Häufigkeitsverteilung) bestimmt, um die Notwendigkeit einer *Missing-Data-Analyse* zu beurteilen.

⁸⁷ Für Selected-Response-Items (z.B. AHV_5) wurden die Codes 1, 2, 3, 4, 999 und 777 programmiert; für Constructed-Response-Items (z.B. AHV_1m) die Codes 0, 1, 2, 999 und 888.

Es wurden zwei Personen (*cases*) identifiziert, die weniger als die Hälfte der Items gelöst hatten; diese wurden aus beiden Datensätzen ausgeschlossen. Die effektive Stichprobengröße beträgt daher $N = 375$ (vgl. Tabelle 10-4). Des Weiteren wurden vier Items (*variables*) identifiziert, die mehr als 5 % fehlende Werte aufweisen (ENE_11m 6.4 %, ENE_12n 6.1 %, STA_7m 5.3 %, MAN_8m 5.3 %). Allerdings sind diese prozentualen Häufigkeiten noch nahe genug an der „kritischen“ Grenze von 5 %, so dass sie als unproblematisch beurteilt und auf eine umfangreiche *Missing-Data-Analyse* verzichtet wurde. Für die quantitativen Analysen wurden die fehlenden Werte im WBK-T2 (999, 888, 777) zum Wert 0 rekodiert, da ein ungelöstes Item in einem Leistungstest als „nicht wissen“ interpretiert werden kann.

Im Datensatz des WBK-Q1 wurden ebenfalls personenbezogene und variablenbezogene deskriptivstatistische Analysen durchgeführt. Alle 79 Items hatten weniger als 5 % fehlende Werte. Diese fehlenden Werte im WBK-Q1 wurden belassen, da eine unbeantwortete Frage in einem Fragebogen als „nicht wollen“ interpretiert werden kann.

10.2.2.b Neue Variablen

Nach der Datenerfassung wurde eine sorgfältige und aufwändige Bereinigung der Datensätze des WBK-T2 und WBK-Q1 vorgenommen (vgl. Döring & Bortz, 2016, S. 589ff.; Fromm et al., 2010, S. 52). Die Datenbereinigung beinhaltet Folgendes:

- Transformation von freien und halbfreien Antworten (z.B. Erst- und Zweitsprache, Staatsbürgerschaft, Beruf der Eltern) in numerische Variablen;
- Umpolung von Variablen (z.B. bei Items zu Skalen im Fragebogen);
- Rekodierung von bestehenden Variablen (z.B. Bildungsabschluss der Eltern von originaler 8-stufiger Antwortskala zu 3-stufiger Werteskala);
- Konstruktion neuer Variablen aus bestehenden Variablen (z.B. Skalenwert für WBK-T2, Indexwert für Berufsstatus der Eltern [ISEI-08]);
- Prüfung des Wertebereichs und der Wertelabels jeder Variablen;

10.2.3 Analyseverfahren und Statistik-Programme

Für die Datenauswertung wurden je nach Validierungsaspekt qualitative oder quantitative Analyseverfahren angewendet. Für die Validierung des Testinhalts (vgl. Kapitel 8) wurde eine zusammenfassende qualitative Inhaltsanalyse (Mayring, 2015) anhand des Materials aus den Experten-Interviews und den Kognitiven Interviews vorgenommen. Für die Validierung der internen Struktur (vgl. Kapitel 11) wurden folgende Analysen durchgeführt: einerseits psychometrische Analysen für ein probabilistisches Testmodell, das Partial-Credit-Rasch-Modell, mit dem Programm ACER ConQuest Version 4 (Adams et al., 2015); andererseits deskriptiv- und inferenzstatistische Analysen mit dem Programm SPSS Statistics Version 25 (IBM, 2017) durchgeführt. Für die Validierung

der Beziehung mit externen Merkmalen (vgl. Kapitel 12) wurden ebenfalls deskriptiv- und inferenzstatistische Analysen, insbesondere Mittelwert- und Korrelationsanalysen mit dem Programm SPSS Statistics Version 25 (IBM, 2017) gerechnet.

11 Validierung der Internen Struktur des WBK-T2

In diesem Kapitel wird der Leitfrage (8) nachgegangen (vgl. Kapitel 1.2):

(8) Welche psychometrische Qualität lässt sich dem gesamten WBK-T2 (Dimensionalität, Skalierbarkeit, Reliabilität, Graduierbarkeit) und den einzelnen Items (Homogenität, Trennschärfe, Schwierigkeit) nach klassischen und probabilistischen Kriterien attestieren?

Mit dieser Leitfrage wird Evidenz für die interne Struktur des WBK-T2 zusammengetragen (AERA et al., 2014, S. 16). Damit einher gehen Annahme III bis Annahme IX zur internen Struktur des WBK-T2 (vgl. Kapitel 8.1.4.a), die in der Folge überprüft werden. Für die Beantwortung dieser übergeordneten Leitfrage wurden psychometrische Analysen basierend auf der Item-Response-Theorie (IRT) mit dem Programm ACER ConQuest Version 4 gerechnet (Adams & Osses, 2016; Adams et al., 2015; Wu et al., 2007). Es wurde ein Partial-Credit-Rasch-Modell spezifiziert (Macaskill & Adams, 2016; Masters, 1982; Wu et al., 2007, S. 32ff.) (vgl. Kapitel 5.2.4) und die Item- und Personenparameter mit der *marginal maximum likelihood* Methode (MML-Methode) geschätzt (Thissen, 1996; Wu et al., 2007, S. 6) (vgl. Kapitel 5.2.3.c). Eine ausführliche Zusammenstellung der psychometrischen Kennwerte des WBK-T2 findet sich in der „Dokumentation des revidierten Tests zur Wirtschaftsbürgerlichen Kompetenz (WBK-T2)“ (Ackermann, 2018c) und ist auf Anfrage bei der Autorin erhältlich. Eine Zusammenfassung der wichtigsten psychometrischen Werte findet sich im Anhang (Tabelle A- 7 bis Tabelle A- 13).

Dieses Kapitel ist als „logisches“ Argument für die valide Interpretation der Testwerte hinsichtlich des Konstrukts WBK-soek aufgebaut; es präsentiert und diskutiert die vielfältigen Ergebnisse, die Evidenz für die interne Struktur des WBK-T2 liefern. Als Erstes wird die Dimensionalität des Konstrukts WBK-soek bzw. des WBK-T2 geprüft (Kapitel 11.1); die Eindimensionalität des Tests ist eine Voraussetzung für die Spezifikation des probabilistischen Testmodells, insbesondere die Verwendung eines Rasch-Modells. Als Zweites wird die Schätzung und Skalierung der Personen- und Itemparameter vorgenommen (Kapitel 11.2); die konsistente Schätzung der Parameter und die gemeinsame Skalierung von Personen- und Itemparameter ist ausschlaggebend für die Beurteilung der „Passung“ des WBK-T2 bei der Zielgruppe sowie für die kriteriumsorientierte Interpretation der WBK-soek von einzelnen Personen. Im dritten Schritt wird die Messgenauigkeit des gesamten WBK-T2 für die Personenparameter bzw. Personen-Testwerte untersucht (Kapitel 11.3); die Messgenauigkeit des Tests hat ebenfalls eine praktische Bedeutung für die valide Interpretation der Testwerte. Als Viertes wird die Gültigkeit des Partial-Credit-Rasch-Modells anhand der Item-Infits, der Schwellenparameter und des DIF geprüft (Kapitel 11.4); die Gültigkeit des Testmodells und insbesondere die Bestätigung der Item-Homogenität ist von zentraler Bedeutung für die inhaltliche Interpretation der Testwerte und für ihre weitere statistische

Verwendung, z.B. für Regressionsanalysen oder Strukturgleichungsmodelle. Im fünften Schritt wird die Verteilung der Testwerte sowie die Qualität der Items im WBK-T2 anhand der klassischen Item-Schwierigkeit und der klassischen Item-Trennschärfe analysiert (Kapitel 11.5); die Qualität der Items gibt wertvolle Hinweise für die Testrevision und somit die Testoptimierung. Schliesslich wird die Graduierung des WBK-T2, d.h. die quantitative Bestimmung von Kompetenzschwellen und die qualitative Beschreibung von Kompetenzniveaus versucht (Kapitel 11.6); die Graduierung ist von praktischer Bedeutung für die kriteriumsorientierte Interpretation der WBK-soek der Stichprobe.

11.1 Dimensionalität des Messmodells

Welche Evidenz findet sich für die Dimensionalität des Konstrukts sozioökonomische Facette der WBK (WBK-soek) und seine Operationalisierung mit dem WBK-T2? Für den WBK-T2 wird eine eindimensionale faktorielle Struktur postuliert (vgl. Annahme III).

Wird für einen Test eine eindimensionale faktorielle Struktur angenommen, müssen alle manifesten Indikatoren (Items) auf *einen* latenten Faktor (Personenmerkmal, Personenfähigkeit) laden, d.h. die Items des Tests werden durch *ein* Konstrukt repräsentiert (Jonkisz et al., 2012, S. 28f.). Mit anderen Worten, die Fähigkeit einer Person ist alleinig erklärend für die Item-Antworten dieser Person. Die Ergebnisse zur Dimensionalität des operationalisierten Konstrukts sind zentral für die Wahl bzw. Prüfung des Testmodells und für die eindeutige Interpretation der Testergebnisse (vgl. Moosbrugger, 2012a).

Um Annahme III zur faktoriellen Struktur des WBK-T2 zu überprüfen, wurden basierend auf der Item-Response-Theorie (IRT) ein eindimensionales Modell und mehrdimensionale Modelle mit dem Programm ACER ConQuest Version 4 gerechnet (Adams & Osses, 2016; Adams & Wu, 2010c; Adams et al., 2015; Wu et al., 2007, S. 103ff.).

Der Aufbau dieses Kapitels ist wie folgt: Zuerst werden die Modellspezifikationen der Messmodelle präsentiert (Kapitel 11.1.1). Danach werden die Modellfit-Indizes der Modelle und die Reliabilitäts-Koeffizienten diskutiert (Kapitel 11.1.2).

11.1.1 Modellspezifikationen

Für einen direkten Vergleich zweier Modelle müssen diese dieselbe Modellstruktur aufweisen, d.h. sie müssen hierarchisch geschachtelte Modelle sein (*nested models*) und durch das Fixieren bzw. Freisetzen einzelner Parameter ineinander überführbar sein (Moosbrugger & Schermelleh-Engel, 2012, S. 334f.). So muss z.B. das eindimensionale Modell als Submodell des multidimensionalen Modells spezifiziert sein. In der KTT wird die faktorielle Struktur bzw. Dimensionalität des Messmodells mittels explorativer (EFA) oder konfirmatorischer Faktorenanalyse (CFA) untersucht (vgl. Moosbrugger & Schermelleh-Engel, 2012) (vgl. Kapitel 5.2.1). In der IRT kann die faktorielle Struktur bzw. Dimensionalität des Tests durch die Spezifikation und Schätzung eines

multidimensionalen Modells überprüft werden. Dabei können zwei Typen multidimensionaler Modelle unterschieden werden (Adams, Wilson, & Wang, 1997; Adams & Wu, 2010c): *multidimensional between-item* Modelle (jedes Item lädt auf einen latenten Faktor bzw. der Test besteht aus mehreren eindimensionalen Subskalen) und *multidimensional within-item* Modelle (mindestens ein Item lädt auf mehrere latente Faktoren).

Aufgrund der ordinalen Datenstruktur des WBK-T2 wird für das eindimensionale Modell das Partial-Credit-Rasch-Modell (Adams & Wu, 2010a; Masters, 1982) und für die mehrdimensionalen Modelle das Multidimensional-Random-Coefficients-Multinomial-Logit-Modell (MRCML-Modell) verwendet. Das MRCML-Modell ist ein generalisiertes Rasch-Modell und kann folglich für den direkten Vergleich mit dem Partial-Credit-Rasch-Modell verwendet werden, da das PC-Modell ein Submodell des MRCML-Modells ist.

Aufgrund der theoretischen Annahmen zum Konstrukt WBK-soek und deren Berücksichtigung bei der Testentwicklung wird für den WBK-T2 eine unidimensionale Struktur angenommen. Es wird also ein einfaktorielles Modell (M1) mit einem latenten Faktor, einem sog. Generalfaktor, und 32 Indikatoren spezifiziert, d.h. alle 32 Items des WBK-T2 laden auf diesen einen Faktor (vgl. Abbildung 11-1). Aufgrund der inhaltlichen und formellen Testspezifikationen (Testteile, Itemtypen) (vgl. Kapitel 8.2.1 und 8.3.3) drängen sich für die empirische Prüfung der faktoriellen Struktur des WBK-T2 neben dem eindimensionalen Modell drei mehrdimensionale Modelle auf: inhaltsbezogen (Problemsituationen des WBK-T2), kognitionsbezogen (Kognitionsstufen der Items) und formatbezogen (Antwortformate der Items). Alle drei Modelle werden als *First-Order*-Modelle und als *Between-Item*-Modelle spezifiziert.

Das *Modell M2* hat vier latente Faktoren analog zu den sozioökonomischen Problemsituationen des WBK-T2 (AHV, ENE, STA und MAN), wobei die Items der jeweiligen Problemsituation auf diesen Faktor laden (vgl. Abbildung 11-2). Diese Faktoren repräsentieren die inhaltsbezogene Spezifikation des WBK-T2 auf der Ebene der Problemsituationen (vgl. Kapitel 8.2.1).

Das *Modell M3* besteht aus zwei latenten Faktoren „k1/k2: erinnern, verstehen/anwenden“ mit 29 Items und „k3: vergleichen/beurteilen/entscheiden“ mit 3 Items (vgl. Abbildung 11-3). Diese Faktoren repräsentieren die kognitionsbezogene Spezifikation des WBK-T2 auf der Ebene der Items, die im Konstruktionsschema berücksichtigt wurde (vgl. Tabelle 8-4). Mit M3 soll also geprüft werden, ob Items mit höheren Kognitionsprozessen eine andere Kompetenzstruktur haben als jene mit geringeren.

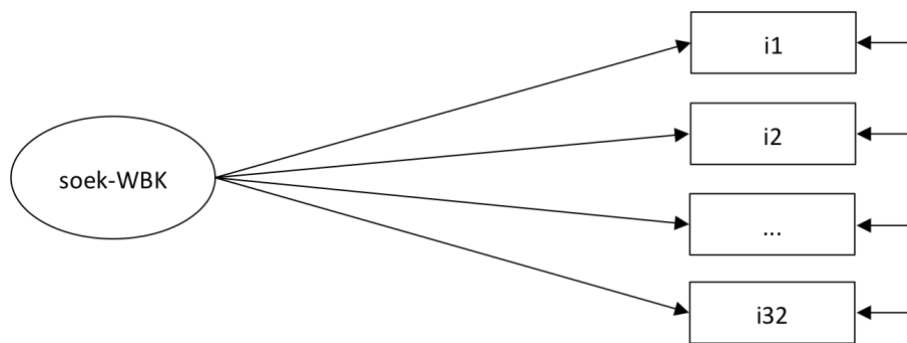


Abbildung 11-1. Messmodell M1 mit einem latenten Faktor und 32 manifesten Indikatoren

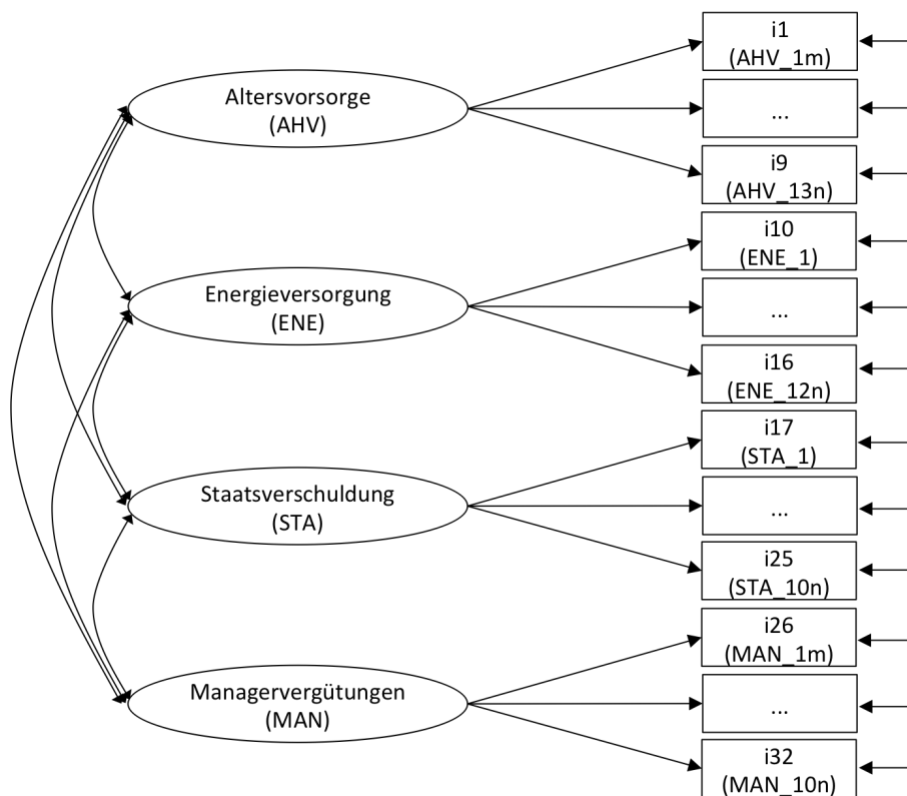


Abbildung 11-2. Messmodell M2 mit vier latenten Faktoren gemäss Inhaltsbereich

Das Modell M4 besteht aus den zwei latenten Faktoren „f1: Selected-Response“ mit 17 Items und „f2/f3: Constructed-Response“ mit 15 Items (vgl. Abbildung 11-4). Diese Faktoren bilden die formatbezogene Spezifikation des WBK-T2 auf Ebene der Items gemäss des Konstruktionsschemas ab (vgl. Tabelle 8-4). Mit M4 soll also geprüft werden, ob Items mit gebundenen und freien Antwortformaten unbesehen ihres Inhalts eine andere Kompetenzstruktur aufweisen.

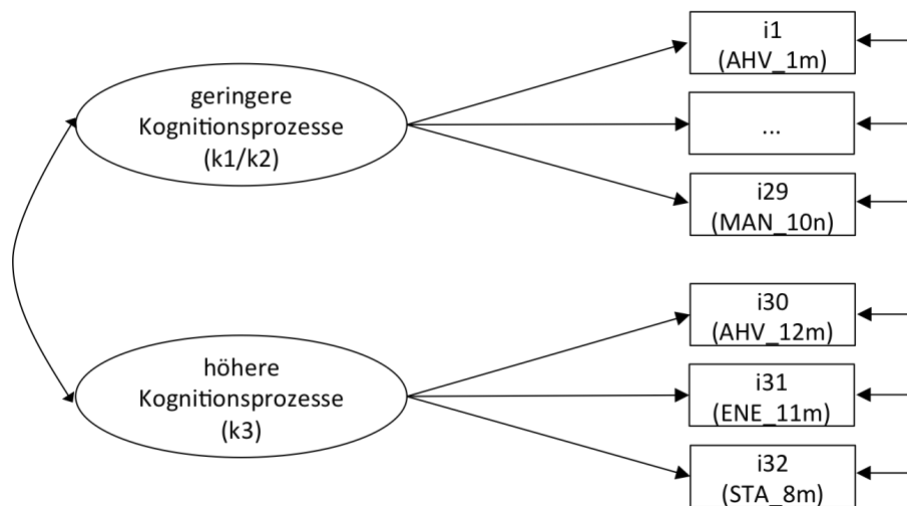


Abbildung 11-3. Messmodell M3 mit zwei latenten Faktoren gemäss Kognitionsprozess

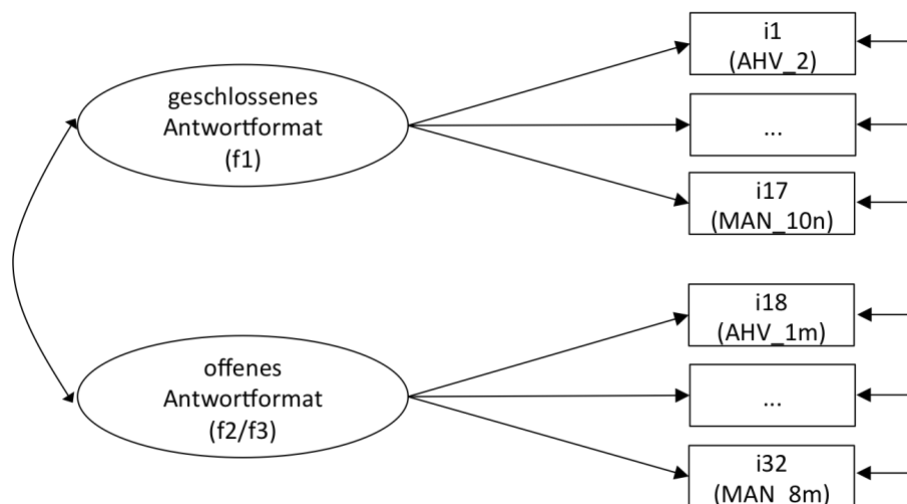


Abbildung 11-4. Messmodell M4 mit zwei latenten Faktoren gemäss Antwortformat

11.1.2 Modellfit-Indizes

Um das eindimensionale Modell zu identifizieren, wird der Mittelwert der Itemparameter auf 0 fixiert. Bei einem mehrdimensionalen Modell wird der Mittelwert der Itemparameter jeder Dimension auf 0 fixiert. Das bedeutet, dass die Personenparameter verschiedener Modelle und verschiedener Dimensionen nicht vergleichbar sind, weil sie keinen gemeinsamen Nullpunkt und keine gemeinsame Skala haben. Hingegen werden für beide Dimensionen alle Personenparameter und die jeweiligen Itemparameter geschätzt. Die Korrelation zwischen den latenten Faktoren ist um Messfehler bereinigt (Adams et al., 2015).

Die Modellgüte bzw. die Anpassungsgüte eines Modells an die Daten kann inferenzstatistisch über einen Modelltest (χ^2 -Test) und/oder deskriptivstatistisch mittels verschiedener Modellfit-Indizes (z.B. RMSEA, CFI) beurteilt werden. Der aus dem Modelltest resultierende χ^2 -Wert sollte möglichst

klein sein, jedenfalls kleiner als zweimal die Anzahl der Freiheitsgrade: $X^2 \leq 2df$. Da der X^2 -Wert abhängig vom Stichprobenumfang grösser wird, werden ergänzend Modellfit-Indizes beigezogen. Die Werte der Modellfit-Indizes werden deskriptiv an sog. Normwerten beurteilt (z.B. RMSEA < .05, CFI $\geq$.97) (Moosbrugger & Schermelleh-Engel, 2012, S. 337).

Der Unterschied in der Modellgüte von zwei hierarchisch geschachtelten Modellen kann über den Modelldifferenztest inferenzstatistisch geprüft werden. Die Differenz der X^2 -Werte beider Modelle wird wiederum an einer X^2 -Statistik geprüft: $X^2_{Modell A}(df_A) - X^2_{Modell B}(df_B) = \Delta X^2(\Delta df)$, wobei Modell A ein (restriktiveres) Submodell von Modell B ist. Wird der X^2 -Differenz-Test signifikant ($p < .05$), so passt das weniger restriktive Modell B signifikant besser auf die Daten als das restriktivere Submodell A. Bevorzugt wird dasjenige Modell, das einen „guten“ Modellfit-Index und zusätzlich einen signifikant besseren Modellfit als das Vergleichsmodell hat (Moosbrugger & Schermelleh-Engel, 2012, S. 337f.).

Für Messmodelle mit ordinalen Daten werden im Programm ACER ConQuest Version 4 für den Modellfit die Informationskriterien Deviance und AIC ausgegeben.⁸⁸ Die Modellgüte kann über die Deviance deskriptivstatistisch abgelesen werden: Je geringer seine Deviance, desto besser passt ein Modell auf die Daten (Rost, 2004, S. 339ff.). Der Unterschied in der Modellgüte kann über den X^2 -Differenz-Test für die Deviance inferenzstatistisch abgesichert werden: Mit der Differenz der Deviance (Δ Deviance) und der Differenz der Freiheitsgrade (Δ df) kann in der X^2 -Statistik die Signifikanz (p) abgelesen werden.

Die Modellfit-Statistiken der vier Modelle des WBK-T2 werden nachfolgend dargestellt (vgl. Tabelle 11-1). Die Modellfit-Indizes des restriktivsten eindimensionalen Modells M1 (Deviance = 17'262, AIC = 17'374) sind erwartungsgemäss höher als jene der weniger restriktiven mehrdimensionalen Modelle M2, M3 und M4.

⁸⁸ Informationskriterien sind informationstheoretische Masse für die Modellgüte eines Modells bzw. für den Vergleich der Modellgüte verschiedener Modelle. Je niedriger der Wert des Informationskriteriums, desto besser passt das Modell auf die Daten. Je nach Informationskriterium fließen die logarithmierte Likelihood des Modells (log-Likelihood), die geschätzten Modellparameter (k_{par}) und die Stichprobengrösse (n) mit ein. In den Statistik-Programmen werden i.d.R. drei Informationskriterien ausgegeben: AIC (*Akaike Information Criterion*) = $-2 \cdot \log(L) + 2 \cdot k_{par}$, BIC (*Bayesian Information Criterion*) = $-2 \cdot \log(L) + \log(n) \cdot k_{par}$ und cAIC (*consistent AIC*) = $-2 \cdot \log(L) + 2 \cdot \log(n) \cdot k_{par} + k_{par}$ (Eid & Schmidt, 2014, S. 190f.).

Tabelle 11-1. Modellfit-Indizes für Modelle M1, M2, M3 und M4

	M1	M2	M3	M4
Cases (n)	375	375	375	375
Items (m)	32	32	32	32
Deviance (-2*log-Likelihood)	17'262	17'167	17'257	17'247
Geschätzte Parameter (k_{par})	56	65	58	58
χ^2 -Differenz-Test (Δ Deviance Δ df)	---	95.4 -9	5.2 -2	15.6 -2
Signifikanz (p, $H_0: M_x = M_1$)	---	< .001	.073	< .001
AIC (-2*log-Likelihood + 2*k _{par})	17'374	17'297	17'373	17'363
Korrelation	---	div.	0.831	0.913
Signifikanz (p; $H_0: r = 1$)	---	div.	< .001	< .001

Das inhaltsbezogene Modell M2 zeigt von allen vier Modellen den besten Modellfit (Deviance = 17'167, AIC = 17'297) und passt signifikant besser auf die Daten als M1 (Δ Deviance = 95.4, Δ df = -9, $p < .001$). Die Korrelationen zwischen den latenten Faktoren sind – mit einer Ausnahme – hoch ($r = .47$ bis $r = .80$; im Mittel $r = .65$). Hingegen ist die Reliabilität der vier Dimensionen – aufgrund der geringen Itemzahl in jeder Dimension – durchgängig unbefriedigend ($EAP/PV_{AHV} = .566$, $EAP/PV_{ENE} = .699$, $EAP/PV_{STA} = .685$, $EAP/PV_{MAN} = .638$).

Für den besseren Modellfit des M2 gegenüber dem M1 gibt es einerseits eine technisch-statistische Erklärung: Bei M2 müssen mehr Parameter ($k_{M2} = 65$) geschätzt werden als bei M1 ($k_{M1} = 56$), daher passt es besser auf die Daten; entsprechend ist M2 weniger sparsam als M1 (vgl. Eid & Schmidt, 2014, S. 190). Andererseits kann der bessere Modellfit auch inhaltlich begründet werden: Die Faktoren des M2 beziehen sich auf verschiedene Problemsituationen, die das Konstrukt bestmöglich repräsentieren und grösstmöglich differenzieren sollen. Allerdings spricht gegen eine weitere psychometrische Verwendung des M2, dass ein multifaktorielles Messmodell ein umfangreicheres Testinstrument und ein komplexeres Testmodell erforderlich macht. Ersteres ist aus testökonomischer Perspektive nicht praktikabel, da der WBK-T2 mit einer Testlänge von 32 Items und einer Testdauer von ca. 60 Minuten bereits die zumutbare Grenze erreicht hat. Zweiteres ist aus testtheoretischer Perspektive nachteilig, da das Rasch-Modell Eindimensionalität annimmt (vgl. Kapitel 5.2.3.b) und im Vergleich zu komplexeren Modellen einfach zu skalieren ist.

Tabelle 11-2. Verteilungs- und Reliabilitäts-Indizes für die Dimensionen der Modelle M3 und M4

Modell	M1	M3: Kognitionsprozess		M4: Antwortformat	
Dimension	---	k1/k2	k3	f1	f2/f3
Cases (N)	375		375		375
Items (m)	32	29	3	17	15
Konstante (Mean, in Logits)	0.525	0.636	-0.608	0.832	0.161
Varianz	0.296	0.286	0.597	0.232	0.398
Korrelation	---		.831		.913
Signifikanz (p; H0: r = 1)	---		< .001		< .001
Reliabilität EAP/PV	.756	.745	.615	.712	.742
Reliabilität Cronbachs α	.74		---		---

Hinweis: k1/k2: erinnern, verstehen/anwenden, k3: vergleichen/beurteilen/entscheiden. f1: Selected-Response, f2/f3: Constructed-Response.

Das kognitionsbezogene Modell M3 hat einen besseren, aber nicht signifikant besseren, Modellfit als M1 (Δ Deviance = 5.2, Δ df = -2, $p = .073$). Somit passt M3 nicht signifikant besser auf die Daten als das restriktivere Modell M1. Die latenten Faktoren von M3 sind hoch korreliert ($r = .83$), aber signifikant von 1 verschieden ($p < .001$).⁸⁹ Das formatbezogene Modell M4 hat einen signifikant besseren Modellfit als Modell M1 (Δ Deviance = 15.6, Δ df = -2, $p < .001$) und passt somit signifikant besser auf die Daten als das restriktivere Modell M1. Jedoch korrelieren die latenten Faktoren von M4 sehr hoch ($r = .91$) und die Korrelation ist signifikant von 1 verschieden ($p < .001$). Eine zweidimensionale Struktur des Konstrukts nach Kognition und Format kann an dieser Stelle also nicht bestätigt werden.

Für die Modelle M3 und M4 werden zudem die Verteilungskennwerte (Mittelwert, Varianz) der Personenparameter und die Reliabilitäts-Koeffizienten der Dimensionen dargestellt (vgl. Tabelle 11-2). Bei M3 zeigt die Dimension k1/k2 mit den geringeren Kognitionsstufen eine substanziell geringere Varianz als die Dimension k3 mit der höheren Kognitionsstufe ($VAR_{k1/k2} = 0.286$ bzw. $VAR_{k3} = 0.597$). Die Reliabilität der Dimension k1/k2 ist befriedigend ($m = 29$, $EAP/PV_{k1/k2} = 0.745$), jene der Dimension k3 aufgrund der geringen Item-Anzahl entsprechend geringer und nicht befriedigend ($m = 3$, $EAP/PV_{k3} = 0.615$) (vgl. Bühner, 2011, S. 81). Bei M4 hat die Dimension f1 mit den gebundenen Antwortformaten eine geringere Varianz als die Dimension f2/f3 mit den Constructed-Response Antwortformaten ($VAR_{f1} = 0.232$ bzw. $VAR_{f2/f3} = 0.398$). Die Reliabilität ist für

⁸⁹ Die Hypothese H0, nämlich dass die Korrelation zwischen den latenten Faktoren gleich 1 ist, wurde mit einem zweiseitigen Test approximativ mittels Fischer-Z-Transformation auf Signifikanz geprüft (Eid, Gollwitzer, & Schmitt, 2011, S. 543f.; Lenhard & Lenhard, 2014).

beide Dimensionen f_1 und f_2/f_3 als befriedigend zu beurteilen ($EAP/PV_{f_1} = 0.712$, $EAP/PV_{f_2/f_3} = 0.712$) (vgl. Bühner, 2011, S. 81).

Zusammenfassend lässt sich festhalten, dass die Daten nicht genügend empirische Evidenz für ein multidimensionales Konstrukt WBK-soek bzw. eine multifaktorielle Struktur des WBK-T2 liefern. Eine kognitionsbezogene oder formatbezogene faktorielle Struktur des WBK-T2 ist aufgrund der hohen Korrelationen zwischen den latenten Faktoren auszuschliessen. Eine inhaltsbezogene Struktur ist aufgrund der teilweise hohen latenten Korrelationen und geringen Reliabilitäten empirisch fragwürdig und testtheoretisch unvorteilhaft. Unter Berücksichtigung dieser Ergebnisse zur Dimensionalitätsprüfung wird das eindimensionale Messmodell als Grundlage für die weiteren psychometrischen Analysen verwendet. Annahme III zur faktoriellen Struktur kann also als erfüllt betrachtet werden.

Diese Evidenz ist aus forschungspraktischer Perspektive als positiv zu werten, da Eindimensionalität des Tests eine Voraussetzung für die IRT-Skalierung mit dem Partial-Credit-Rasch-Modell ist (vgl. Döring & Bortz, 2016, S. 289; Eid & Schmidt, 2014, S. 129). Für die nachfolgenden Analysen wird also von einer eindimensionalen Struktur ausgegangen.

11.2 Schätzung und Skalierung der Parameter

Welche Evidenz findet sich für die Schätzung und Skalierung der Personen- und Itemparameter im WBK-T2? Die Personenparameter der Stichprobe WBKgym und die Itemparameter des WBK-T2 lassen sich konsistent schätzen und auf einer kontinuierlichen Logit-Skala skalieren (vgl. Annahme IV).

Ein Test gilt durch seine Items als umfassend operationalisiert und die Parameter des Modells als präzise schätzbar, wenn (Eid & Schmidt, 2014, S. 175f.; Strobl, 2010, S. 37; Winther, 2010, S. 216f.): (1) die Itemparameter auf der Logit-Skala ausreichend weit streuen, idealerweise im gleichen Bereich wie die Personenparameter, und (2) die Itemparameter möglichst viele verschiedene Logit-Werte aufweisen, d.h. die Items unterschiedlich schwierig sind. Das Ergebnis bezüglich der Skalierung hat eine grosse forschungspraktische Relevanz: Die Personenparameter können als metrische Variablen für weitere Analysen (z.B. Regressionsanalysen, Strukturgleichungsmodelle) verwendet werden.

Zur Prüfung von Annahme IV zur Schätzung und Skalierung der Parameter wird basierend auf der Item-Response-Theorie (IRT) ein probabilistisches Testmodell, nämlich das Partial-Credit-Rasch-Modell spezifiziert (Masters, 1982; Moosbrugger, 2012a, S. 265ff.; Strobl, 2010, S. 54ff.) (vgl. Kapitel 5.2.4). Die Item- und Personenparameter werden mit der MML-Methode (vgl. Kapitel 5.2.3.c) im Programm ACER ConQuest Version 4 geschätzt (Adams & Wu, 2010b; Adams et al., 2015; Wu et al., 2007, S. 5).

Der Aufbau dieses Kapitels ist wie folgt: Zuerst wird die *Item-Person-Map* analysiert (Kapitel 11.2.1), danach die Personenparameter und die Personen-Testwerte (Kapitel 11.2.2) sowie die Item-Schwierigkeitsparameter nach Problemsituation (Kapitel 11.2.3).

11.2.1 Person-Item-Map

Die mittels MML-Methode geschätzten Personenparameter (probabilistische Personenfähigkeit) und Item-Schwierigkeitsparameter (probabilistische Item-Schwierigkeit) werden auf eine gemeinsame Logit-Skala transformiert bzw. skaliert, somit können sie gemeinsam interpretiert werden (vgl. Kapitel 5.2.3.c). Die *Person-Item-Map* (auch: Wright-Map) zeigt die Personenparameter (θ) (Anzahl Fälle als X) und die Itemparameter (σ) (mit Item-Laufnummer) auf einer gemeinsamen Logit-Skala abgetragen (vgl. Abbildung 11-5, links). Für ordinale Antwortkategorien werden in einer separaten *Person-Item-Map* zusätzlich zu den Itemparametern auch die Kategorienparameter⁹⁰ abgetragen (vgl. Abbildung 11-5, rechts).

⁹⁰ Im Manual und im Output-File von ACER ConQuest werden die Item-Schwierigkeitsparameter als „estimate“ bezeichnet, die kumulierten Kategorien-Schwierigkeitsparameter (missverständlicherweise!) als „(Thurstonian) thresholds“ und die dekumulierten Schwellenparameter als „delta“ (Wu et al., 2007, S. 38ff., S. 159f.).

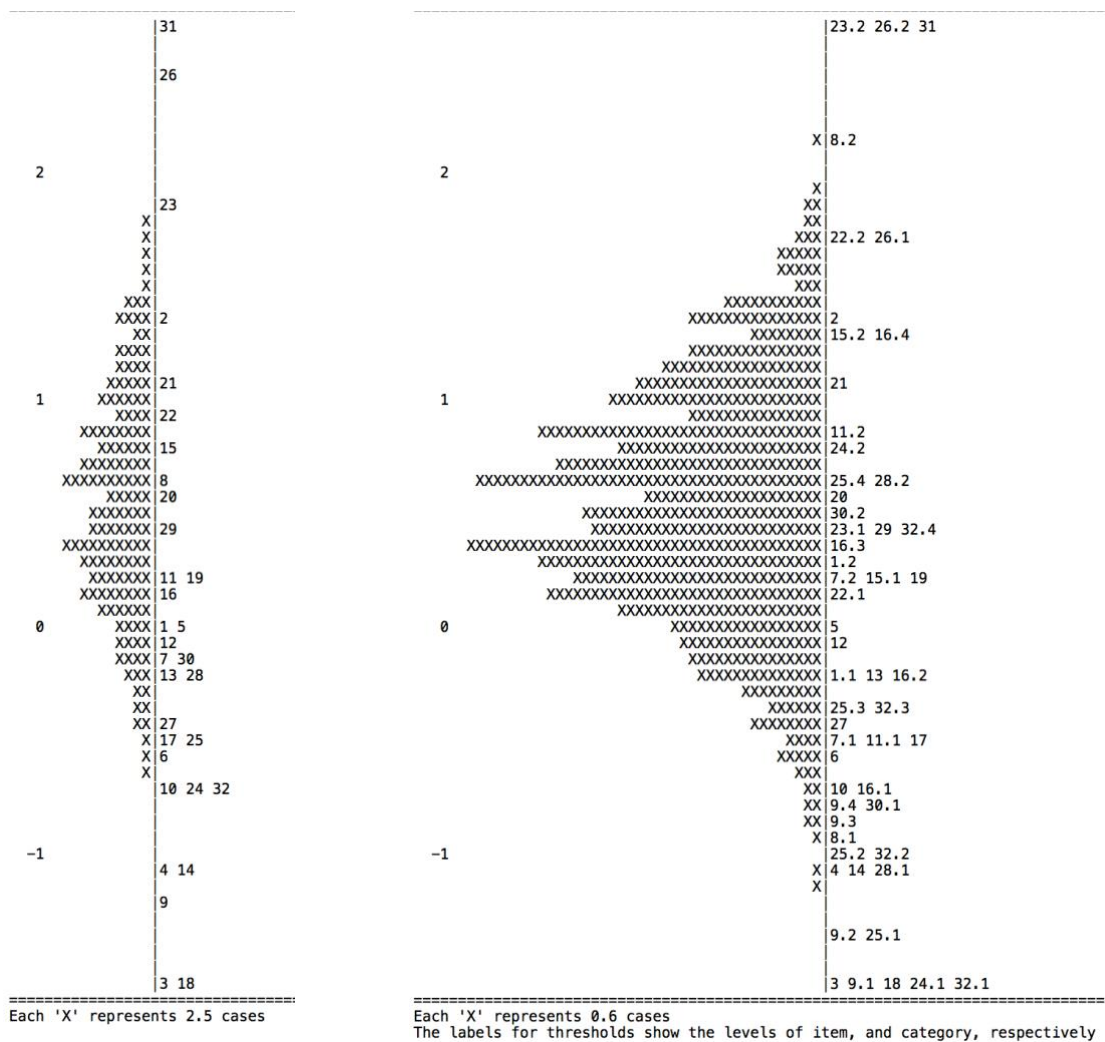


Abbildung 11-5. Person-Item-Maps für Stichprobe WBKgyM und WBK-T2

Hinweis: Personenparameter (EAP) und Item-Schwierigkeitsparameter (links) bzw. Personenparameter (EAP) und Kategorien-Schwierigkeitsparameter (rechts). Personen sind mit X markiert; Items und Antwortkategorien mit Laufnummern angegeben (z.B. 3 für Item 3, 3.1 für Kategorie 1 des Items 3). Ein positiver Wert ($\theta > 0$ bzw. $\sigma > 0$ Logits) bedeutet eine hohe Personenfähigkeit bzw. hohe Item-Schwierigkeit, ein negativer Wert ($\theta < 0$ bzw. $\sigma < 0$ Logits) eine geringe.

Prima vista zeigt die Item-Person-Map für den WBK-T2 und die Stichprobe WBKgyM, dass die Itemparameter auf der Logit-Skala breit streuen, sogar bereiter als die Personenparameter (vgl. Abbildung 11-5, links). Diese Streuung erlaubt auch für sehr niedrige und sehr hohe Personenfähigkeiten eine (zuverlässige) Schätzung. Insgesamt belegen die Lage- und Streuungsmasse der Itemparameter, dass im WBK-T2 sehr einfache und sehr schwierige Items vorzufinden sind und dass die Items des WBK-T2 das Konstrukt WBK-soek empirisch umfassend zu repräsentieren vermögen.

Die Item-Schwierigkeitsparameter streuen im Bereich von -3.213 Logits und +3.168 Logits (vgl. Abbildung 11-5, Abbildung 11-7, Anhang Tabelle A- 7). Der Mittelwert der Item-Schwierigkeitsparameter ist aufgrund der Summennormierung für die Schätzung auf 0 fixiert ($M_\sigma = 0$,

$SD_{\sigma} = 1.208$, $VAR_{\sigma} = 1.460$). Der Kolmogoroff-Smirnov-Test und der Shapiro-Wilk-Test zeigen eine Normalverteilung der Itemparameter in der Grundgesamtheit an ($K-S_{(32)} = 0.123$, $p = .200$; $S-W_{(32)} = 0.960$, $p = .266$), was für die quantitativen Analysen zur Gradierung des WBK-T2 erforderlich ist (vgl. Kapitel 11.6). 41 % der Items ($m = 19$) liegen im negativen Bereich der Logits-Skala ($\sigma_i < 0$) und sind als „eher einfach“ bzw. „sehr einfach“ zu qualifizieren, 59 % der Items ($m = 13$) im positiven Bereich ($\sigma_i > 0$) und sind als „eher schwierig“ bzw. „sehr schwierig“ zu qualifizieren. Drei Items (AHV_5, MAN_1m, MAN_9n) befinden sich ausserhalb des Logits-Bereichs $[-2, +2]$, sind also sehr einfach bzw. sehr schwierig. Die exakten Werte der Itemparameter und Schwellenparameter finden sich im Anhang Tabelle A- 7.

Die Personenparameter (WLE) haben auf der Logit-Skala den geringsten Wert bei -1.580, den höchsten bei 2.340 (vgl. Abbildung 11-5, Anhang Tabelle A- 13). Der mittlere Personenparameter beträgt 0.525 Logits ($M_{\theta} = 0.525$, $SD_{\theta} = 0.627$, $VAR_{\theta} = 0.394$). 19 % der Fälle ($N = 71$) liegen im negativen Bereich der Logit-Skala ($\theta_v < 0$), 81 % ($N = 304$) im positiven ($\theta_v > 0$). Die Personenparameter haben also eine vom Nullwert nach rechts verschobene Verteilung. Der Kolmogoroff-Smirnov-Test und der Shapiro-Wilk-Test zeigen keine bzw. knapp keine Normalverteilung für die Personenparameter in der Grundgesamtheit an ($K-S_{(375)} = 0.070$, $p < .001$; $S-W_{(375)} = 0.992$, $p = .049$). Allerdings weist der *Critical-Ratio*-Test auf eine Normalverteilung der Personenparameter hin (CR-Test Schiefe = 0.984, CR-Test Kurtosis = -0.602), so dass eine *annähernde* Normalverteilung angenommen werden kann. Die exakten Werte der Personenparameter finden sich im Anhang Tabelle A- 13).

Die Ungleichverteilung der Personenparameter auf der Logit-Skala zugunsten des positiven Logit-Bereichs deutet darauf hin, dass der WBK-T2 für die Stichprobe WBKgym zu wenige Items mit hohen Schwierigkeiten bereitstellt und dass der WBK-T2 für die vorliegende Stichprobe etwas zu einfach konzipiert ist. Zudem deutet diese Ungleichverteilung der Personenparameter darauf hin, dass – basierend auf der IRT mit dem Partial-Credit-Rasch-Modell – hohe Personenfähigkeiten weniger präzise geschätzt werden können. Diese deskriptivstatistischen Vermutungen zur Messgenauigkeit werden weiter unten wieder aufgenommen und detailliert analysiert (vgl. Kapitel 11.3.1).

11.2.2 Personenparameter und Personen-Testwerte

Wenn das Partial-Credit-Rasch-Modell gilt, dann ist der Personen-Testwert (auch: *sum score*, Rohwert) eine erschöpfende Statistik für die Personenfähigkeit (zur suffizienten Statistik vgl. Kapitel 5.2.3.b). Mit anderen Worten: Der Testwert enthält alle Informationen über die Personenfähigkeit und für jeden Testwert resultiert genau ein Personenparameter θ (Rost, 2004, S. 124; Strobl, 2010, S. 14ff.). Jedoch dürfen die Personen-Testwerte nur verwendet werden, um die Rangfolge der Personen zu bestimmen (z.B. $\theta_3 > \theta_2 > \theta_1$). Um Unterschiede zwischen einzelnen Testpersonen zu

quantifizieren, müssen die Personenparameter verwendet werden (z.B. $\theta_2 - \theta_1 = 0.5$ Logits). Numerisch identische Differenzen im Personenparameter können, unabhängig vom absoluten Wert der jeweiligen Personenparameter, als quantitativ gleiche Unterschiede in der Personenfähigkeit interpretiert werden (Rost, 2004, S. 121f.) (zur spezifischen Objektivität vgl. Kapitel 5.2.3.b).

Die *Test Characteristic Curve* (TCC) zeigt den Zusammenhang zwischen den manifesten summierten Personen-Testwerten (*sum score*) und der latenten Personenfähigkeit (*latent trait*). Der empirische Zusammenhang zwischen den Personen-Testwerten und den geschätzten Personenparametern ist i.d.R. beinahe linear, jedenfalls streng monoton und sehr hoch⁹¹ (Rost, 2004, S. 122).

Für den WBK-T2 zeigt die *Test Characteristic Curve* (TCC) einen sehr hohen ($r = .99$, $p < .001$), monotonen, aber nicht-linearen Zusammenhang zwischen den Personen-Testwerten und den Personenparametern (EAP) (vgl. Abbildung 11-6). Für eine bestimmte Differenz im Personen-Testwert (z.B. 5 Punkte) ist die Differenz im Personenparameter im mittleren Logit-Bereich geringer ($-/+ \theta = 0.526$) als im unteren ($\theta < -2.0$) und oberen ($\theta > +2.0$) Logit-Bereich. Sofern die Gültigkeit des Partial-Credit-Rasch-Modells für den WBK-T2 bestätigt werden kann (vgl. Kapitel 11.4), dürfen also für weitere Analysen die summierten Personen-Testwerte statt der geschätzten Personenparameter verwendet werden.

11.2.3 Item-Schwierigkeitsparameter nach Problemsituation

Die Verteilung der geschätzten Itemparameter auf der Logit-Skala zeigt, wie sich die Item-Schwierigkeiten innerhalb der vier Problemsituationen zusammensetzten (vgl. Abbildung 11-7).

In der Problemsituation Altersvorsorge (AHV) befinden sich mehrheitlich einfache Items, ein Item ist sehr einfach ($\sigma < -3$ Logits). Die mittlere Item-Schwierigkeit bei AHV im Vergleich zu den anderen Problemsituationen am geringsten ($\sigma_{AHV} = -0.483$). Die Problemsituationen Energieversorgung (ENE) und Staatsversorgung (STA) haben in etwa gleich viele einfache und schwierige Items, jedoch ist die Ansammlung der Item-Schwierigkeiten bei ENE „kompakter“ als bei STA. Die mittlere Item-Schwierigkeit ist bei ENE ($\sigma_{ENE} = -0.160$) etwas geringer als bei STA ($\sigma_{STA} = 0.122$). In der Problemsituation Managergehälter (MAN) gibt es zwei sehr schwierige Items ($\sigma > +2$ Logits). Die mittlere Item-Schwierigkeit bei MAN ist deutlich am höchsten ($\sigma_{MAN} = 0.625$) als bei den anderen Problemsituationen. Jedoch ist die Varianz in den probabilistischen Item-Schwierigkeiten zwischen den Problemsituationen nicht signifikant grösser als die Varianz innerhalb der Problemsituationen ($F_{(3;28)} = 1.198$, $p = 0.329$, $\eta^2 = 0.11$).

⁹¹ Die Korrelation zwischen den Personen-Testwerten und den geschätzten Personenparameter beträgt i.d.R. $.90 < r < .95$ (Rost, 2004, S. 122).

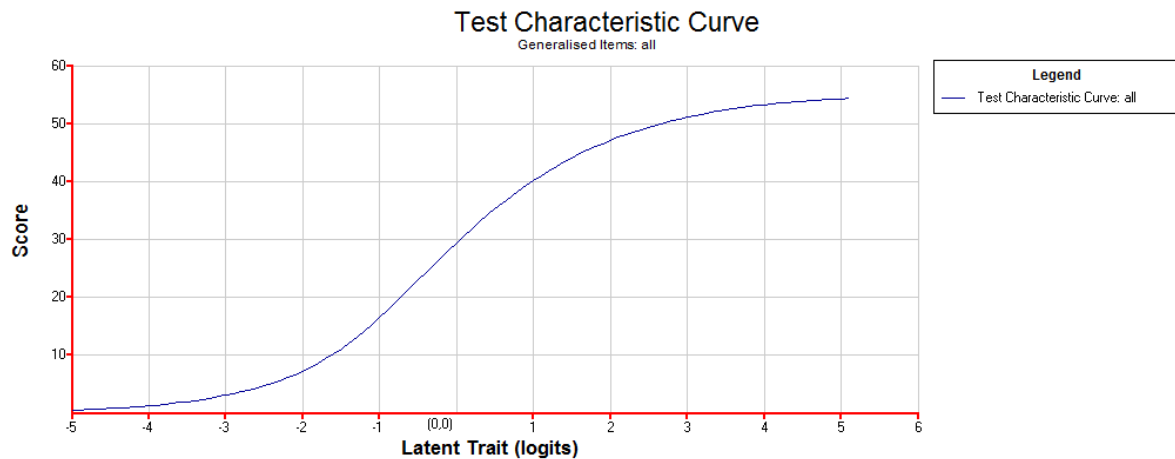


Abbildung 11-6. Test Characteristic Curve (TCC) für WBK-T2

Hinweis: Der Wendepunkt der TCC befindet sich beim Mittelwert des Personenparameters ($M = 0.526$ Logits) und des Personen-Testwertes ($M = 35.05$ Punkte).

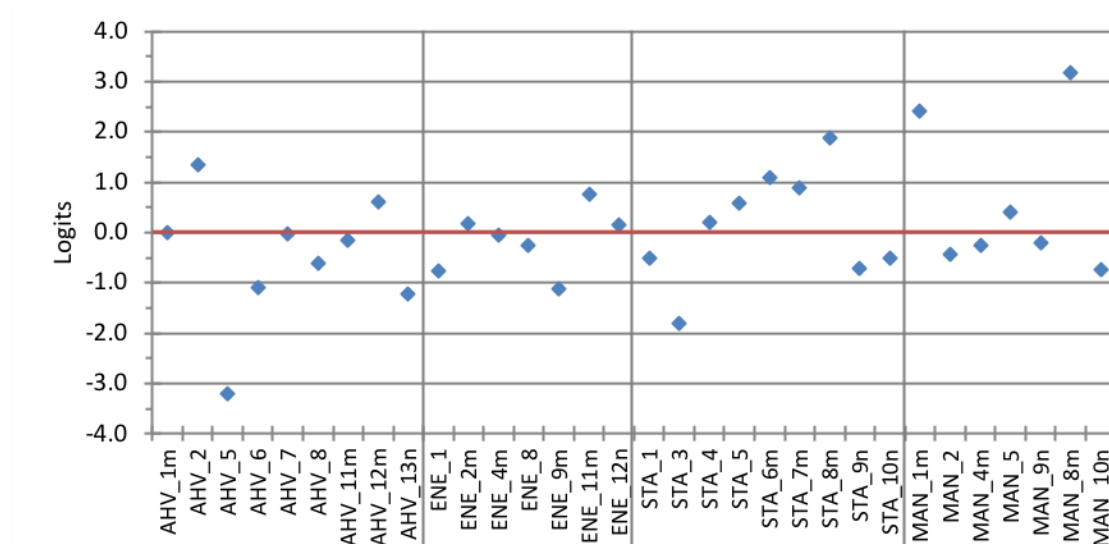


Abbildung 11-7. Itemparameter nach Problemsituation

Hinweis: Ein positiver Wert ($\sigma > 0$ Logits) bedeutet eine hohe Item-Schwierigkeit, ein negativer Wert ($\sigma < 0$ Logits) eine geringe.

Aufgrund der obigen Ergebnisse kann nun erstens festgehalten werden, dass mit den geschätzten Personenparameter und Itemparameter sowie mit der *Person-Item-Map* ausreichend Evidenz für die Skalierung des WBK-T2 vorliegt. Zweitens kann festgehalten werden, dass die Itemparameter ausreichend auf der Logit-Skala streuen, sich bezüglich der Itemtypen unsystematisch auf der Logit-Skala verteilen und insbesondere alle Personenparameter abdecken. Allerdings sind die Itemparameter im positiven Logit-Bereich für die vorliegende Stichprobe etwas unterrepräsentiert. Trotzdem kann Annahme IV an dieser Stelle als erfüllt betrachtet werden.

Ob dieses Ergebnis als ein Spezifikum des WBK-T2 oder als ein Charakteristikum der Stichprobe WBKgym zu deuten ist, kann derzeit nicht beantwortet werden. Die Autorin vermutet jedoch, dass die Stichprobe im Vergleich zur definierten Zielgruppe des WBK-T2 (vgl. Kapitel 8.1.3) eine überdurchschnittlich hohe WBK-soek aufweist. Um dies zu untersuchen bedürfte es weitere Validierungsschritte, u.a. eine genügend grosse Eichstichprobe für eine normorientierte Interpretation der Testergebnisse (vgl. Kapitel 5.1.2.d).

11.3 Messgenauigkeit des Tests

Welche Evidenz findet sich für die Messgenauigkeit des WBK-T2? Die Personenparameter und Personen-Testwerte lassen sich mit dem WBK-T2 genügend genau schätzen (Test-Reliabilität) (vgl. Annahme V).

Die Messgenauigkeit eines Tests kann anhand der Standardfehler für die einzelnen Personenparameter sowie der Reliabilitäts-Koeffizienten über alle Personenparameter bzw. Personen-Testwerte hinweg eruiert werden. Die Standardfehler für jeden Personenparameter sind informativer als die Reliabilitäts-Koeffizienten (AERA et al., 2014, S. 37ff.; Moosbrugger, 2012b, S. 108). Ein Test gilt für das zu messende Konstrukt als reliabel schätzbar, wenn (Eid & Schmidt, 2014, S. 173ff.; Moosbrugger & Kelava, 2012a, S. 11ff.; Schermelleh-Engel & Werner, 2012): (1) die Standardfehler der Personenparameter relativ gering sind und (2) die probabilistischen und klassischen Reliabilitäts-Koeffizienten relativ hoch sind ($EAP/PV \geq .70$, Cronbachs $\alpha \geq .70$). Das Ergebnis zur Messgenauigkeit ist zentral für die weitere Testanwendung: Je präziser die Personenfähigkeiten für den WBK-T2 geschätzt werden können, desto valider ist die Interpretation der Personenparameter und der Personen-Testwerte.

Zur Prüfung von Annahme V zur Messgenauigkeit des WBK-T2 werden die IRT-Analysen mit dem Partial-Credit-Rasch-Modell beigezogen (vgl. Kapitel 11.2), um die spezifische und pauschale Messgenauigkeit für die Personenparameter und Personen-Testwerte zu untersuchen.

Der Aufbau dieses Kapitels ist wie folgt: Als Erstes werden für die spezifische testwertabhängige Messgenauigkeit die *Test Information Curve* (TIC) und die Standardfehler der Personenparameter beurteilt (Kapitel 11.3.1). Als Zweites werden für die pauschale Messgenauigkeit der probabilistische Reliabilitäts-Koeffizient (EAP/PV) und der klassische Reliabilitäts-Koeffizient (Cronbachs α) beurteilt (Kapitel 11.3.2).

11.3.1 Test Information Curve (TIC) und Standardfehler

Die Messgenauigkeit eines einzelnen Personenparameters kann anhand des Standardmessfehlers, d.h. des Standardfehlers jeden geschätzten Personenparameters und des Konfidenzintervalls beurteilt sowie an der *Test Information Curve* (TIC) abgelesen werden (vgl. Kapitel 5.2.3.a).

Im WBK-T2 streuen die Personenparameter (EAP) auf der Logit-Skala von -1.580 bis 2.340 ($M = 0.526$, $SD = 0.627$). Die TIC des WBK-T2 nimmt ihre höchsten Werte im Bereich -1 bis 0 Logits an und liefert dort die maximale psychometrische Information (vgl. Abbildung 11-8). Daraus lässt sich ableiten, dass der WBK-T2 die zuverlässigste Schätzung für Testpersonen mit eher geringer Fähigkeit ($-1 < \theta < 0$; 19 % der Fälle) und die unzuverlässigste Schätzung für Personen mit sehr hoher Fähigkeit ($\theta > 1$; 22 % der Fälle) generiert. Eine immer noch zuverlässige Schätzung resultiert für Personen mit

eher hoher Fähigkeit ($0 < \theta < 1$; 59 % der Fälle) und für Personen mit sehr geringer Fähigkeit ($\theta < -1$; 0.3 % der Fälle). Die Interpretation der TIC deckt sich mit der *ad hoc* Interpretation der *Person-Item-Map*: der WBK-T2 umfasst im negativen Logit-Bereich etwas mehr Items als im positiven Logit-Bereich (vgl. Abbildung 11-5).

Im WBK-T2 nehmen die Standardfehler der Personenparameter (WLE) insgesamt relativ geringe Werte an ($M = 0.315$, $SD = 0.038$) und liegen in einem eher schmalen Bereich ($\min = 0.277$ Logits, $\max = 0.481$ Logits) (vgl. Anhang Tabelle A- 13). Die Schätzung der Personenparameter (WLE) ist für den Logit-Bereich $-0.50 < \theta < +0.55$ auf dem 5 %-Niveau signifikant und kann insofern als genügend genau angenommen werden. Die Standardfehler der Personenparameter (WLE) sind im Logit-Bereich -1 bis 0 am geringsten, entsprechend ist dort die Schätzung der Personenfähigkeit am genauesten. Umgekehrt sind die Standardfehler für den Logit-Bereich > 1 am höchsten und die Schätzung dort am wenigsten genau (vgl. Abbildung 11-9). Dieses Ergebnis ist insofern als eher kritisch zu beurteilen, als dass der Mittelwert der Personenparameter für die vorliegende Stichprobe WBKgym im Logit-Bereich $> +1$ liegt ($M_{WLE} = 0.526$, $SD_{WLE} = 0.627$), wo der WBK-T2 weniger präzise misst.

Die obigen Ergebnisse zur Test Information Curve (TIC) und den Standardfehlern für die Personenparameter liefern Evidenz für eine ausreichend genaue Schätzung der Personenparameter des WBK-T2 mit dem Partial-Credit-Rasch-Model, obschon die Schätzgenauigkeit über den oberen Logit-Bereich weniger genau ist. Annahme V kann also aus Sicht der spezifischen Messgenauigkeit bestätigt werden.

11.3.2 Reliabilitäts-Koeffizienten

Die Messgenauigkeit des Tests für die globalen Personenparameter kann anhand des probabilistischen Reliabilitäts-Koeffizienten EAP/PV beurteilt werden, für die Personen-Testwerte anhand des klassischen Reliabilitäts-Koeffizienten Cronbachs α (vgl. Kapitel 5.1.2.b). Die in den IRT-Analysen ebenfalls erhältlichen Koeffizienten der Person-Separation-Reliabilität (MLE, WLE) sind Koeffizienten für die Reproduzierbarkeit der Personenparameter: Je höher die Person-Separation-Reliabilität, desto besser kann der Test auch *tatsächlich* zwischen fähigeren und weniger fähigen Personen unterscheiden (Eid & Schmidt, 2014, S. 181f.). Die Item-Separation-Reliabilität zeigt die Reproduzierbarkeit der Itemparameter: Je höher die Item-Separation-Reliabilität, desto besser kann der Test die *tatsächlich* schwierigen und einfachen Items in ihrer Reihenfolge anordnen.

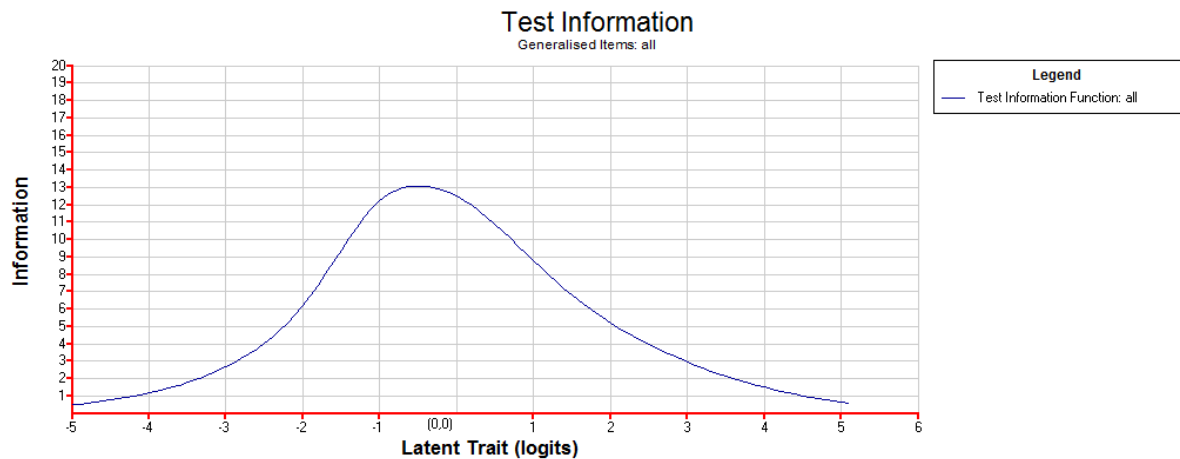


Abbildung 11-8. Test Information Curve (TIC) für Stichprobe WBKgym im WBK-T2

Hinweis: Der WBK-T2 liefert die maximale Information für Personenfähigkeiten im Logit-Bereich $-1 < \theta < 0$.

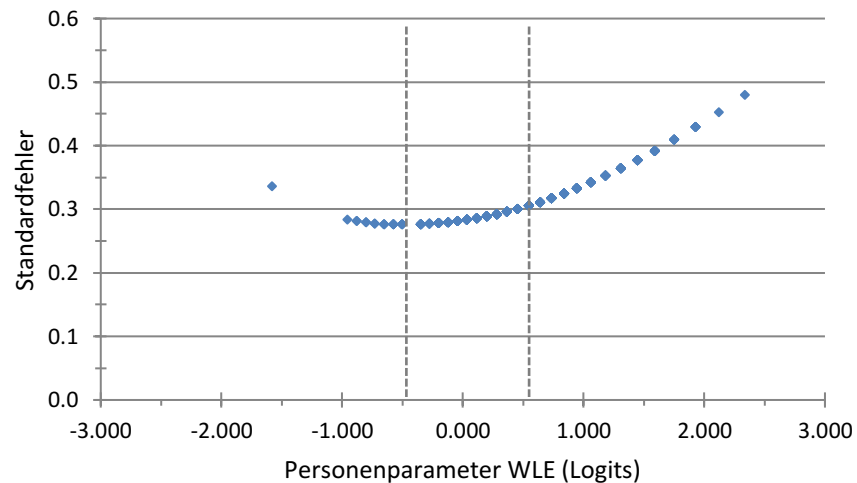


Abbildung 11-9. Personenparameter und Standardfehler für Stichprobe WBKgym im WBK-T2

Hinweis: Personenparameter (WLE). Die gestrichelten Linien markieren den Logit-Bereich, wo die Schätzung der WLE signifikant ist ($p < 0.5$). $M_{WLE} = 0.526$, $SD_{WLE} = 0.627$. $M(SE_{WLE}) = 0.315$, $SD(SE_{WLE}) = 0.038$.

Tabelle 11-3. Reliabilitäts-Koeffizienten

Probabilistische Kennwerte	Personenbezogen	Reliabilität EAP/PV	.756
		Reliabilität MLE	.752
		Reliabilität WLE	.744
Klassische Kennwerte	Itembezogen	Item-Separation Reliabilität	.989
		Reliabilität Cronbachs α	.74

Im WBK-T2 beträgt die probabilistische Reliabilität $EAP/PV = .756$, die Person-Separation-Reliabilitäten betragen $MLE = 0.752$ und $WLE = .744$ (vgl. Tabelle 11-3). Die klassische Reliabilität beträgt Cronbachs $\alpha = .74$. Diese Werte sind zwar nach den konventionellen Normen „nur“ als ausreichend zu beurteilen (vgl. Bühner, 2011, S. 81). Dennoch ist dieses Ergebnis angesichts der umfassenden Revision für den WBK-T2 mit 19 modifizierten bzw. neuen Items und 13 unveränderten Items (vgl. Kapitel 7.2.2) sowie des erstmaligen Einsatzes des WBK-T2 sehr erfreulich.

Daraus lässt sich schliessen, dass die probabilistischen und klassischen Reliabilitäts-Koeffizienten Evidenz für eine ausreichend genaue Schätzung der Personenparameter bzw. der Personen-Testwerte mit dem WBK-T2 liefern. Annahme V kann aus Sicht der pauschalen Messgenauigkeit ebenfalls bestätigt werden.

Zusammenfassend kann festgehalten werden, dass sich mit der *Test Information Curve* (TIC) und den Standardmessfehlern sowie den Reliabilitäts-Koeffizienten ausreichende empirische Evidenz für die Messgenauigkeit der Personenparameter bzw. Personen-Testwerte mit dem WBK-T2 findet. Es scheint also mit dem WBK-T2 zu gelingen, die einzelnen Personenfähigkeiten für das zu messende Konstrukt sozioökonomische Facette der WBK (WBK-soek) reliabel und trennscharf zu messen. Diese Evidenz ist massgeblich für die valide Interpretation der Testwerte hinsichtlich des Konstrukts sozioökonomische Facette der WBK.

11.4 Gültigkeit des Testmodells: Item-Homogenität

Welche Evidenz lässt sich für die Gültigkeit des probabilistischen Testmodells, mit dem die Itemparameter und Personenparameter des WBK-T2 geschätzt wurden, ableiten? Aufgrund Konstruktion des WBK-T2 mit dichotomen und ordinalen Antwortkategorien wird als Testmodell ein Partial-Credit-Rasch-Modells spezifiziert und für die Items des WBK-T2 Item-Homogenität postuliert (vgl. Annahme VI).

Die Annahmen des Partial-Credit-Rasch-Modells gelten als erfüllt, wenn (Eid & Schmidt, 2014, S. 182ff., S. 198ff.; Moosbrugger, 2012a, S. 244ff.; Strobl, 2010, 39ff., S. 54ff.): (1) die Item-Infits ihrem Erwartungswert entsprechen (ideal $wMNSQ = 1$ bzw. tolerierbar $0.80 \leq wMNSQ \leq 1.20$), (2) die Schwellenparameter, d.h. die Schnittpunkte zwischen den CC-Funktionen aufsteigend angeordnet sind, und (3) kein *Differential Item Functioning* (DIF) bezüglich der Stichprobengruppen vorliegt. Bei Konstanthalten der Personenfähigkeit dürfen zwischen den manifesten Variablen also keine Zusammenhänge mehr bestehen, abgesehen von unsystematischen Messfehlern. Die Ergebnisse zur Modellgültigkeit bzw. der Item-Homogenität sind zentral für die Testentwicklung und -evaluation sowie für die Berechnung und Interpretation der Testwerte. Zum Beispiel können die Personen-Testwerte stellvertretend für die Personenparameter als metrische Variablen verwendet werden.

Um Annahme VI zur Modellgültigkeit für den WBK-T2 zu prüfen, werden die IRT-Analysen mit dem Partial-Credit-Rasch-Modell beigezogen und die probabilistischen Item-Kennwerte (Item-Infit, Schwellenparameter) untersucht (vgl. Kapitel 11.2). Zusätzlich werden DIF-Analysen für das Partial-Credit-Rasch-Modell mit dem Programm ACER ConQuest Version 4 gerechnet (Adams & Wu, 2010b; Adams et al., 2015; Wu et al., 2007, S. 89ff.). Das DIF wird anhand einer Interaktion zwischen Itemparametern und Gruppen spezifiziert und mittels grafischer und numerischer Tests auf globaler und lokaler Ebene (X^2 -Tests, X^2 -Differenz-Test, DIF-Rasch-Tests) untersucht.

Der Aufbau dieses Kapitels ist wie folgt: Als Erstes wird für alle Items und Antwortkategorien der Toleranzbereich der Item-Infits beurteilt und mit einigen *Item Characteristic Curves* (ICC) exemplarisch illustriert (Kapitel 11.4.1). Als Zweites wird für die polytomen Items die Anordnung der Schwellenparameter beurteilt (Kapitel 11.4.2). Als Drittes wird die Homogenität der Item-Schwierigkeitsparameter mittels *Differential Item Functioning* (DIF) grafisch und numerisch für zwei Stichprobengruppen überprüft (Kapitel 11.4.3).

11.4.1 Item-Infits

Um die Gültigkeit des Partial-Credit-Rasch-Modells hinsichtlich der Item-Trennschärfe zu prüfen werden die gewichteten Item-Infits und die *Item Characteristic Curves* (ICC) herangezogen.

Für den WBK-T2 liegen die gewichteten Item-Infits für alle Items im Bereich von $0.92 \leq \text{wMNSQ} \leq 1.17$, also im Toleranzintervall von $\text{wMNSQ} \leq 1.20$ (vgl. Abbildung 11-10). Die t-Werte der wMNSQ sind im Bereich von -2.2 und 2.4, d.h. dass die Abweichung der erwarteten und beobachteten Lösungshäufigkeiten für (fast) alle Items nicht signifikant ist. Von den 32 Items zeigen drei einen geringen *Underfit* ($\text{wMNSQ} = 1.17$ für AHV_13n; $\text{wMNSQ} = 1.12$ für AHV_1m; $\text{wMNSQ} = 1.10$ für MAN_10n). Von den Kategorienparametern des WBK-T2 zeigen bloss zwei einen geringen, aber immer noch tolerierbaren *Underfit* (AHV_1m Kategorie 2, AHV_13n Kategorie 4) (vgl. Tabelle 11-4). Dieses Ergebnis zu den Item-Infits spricht für die Gültigkeit des Partial-Credit-Rasch-Modells. Es ist insgesamt sehr befriedigend, v.a. angesichts der Tatsache, dass 19 der 32 eingesetzten Items bei der umfassenden Testrevision für den WBK-T2 modifiziert oder neu entwickelt wurden (vgl. Kapitel 7.2.2).

Nachfolgend werden die *Item Characteristic Curves* (ICC) für drei ausgewählte Items aus dem WBK-T2 mit unterschiedlichen Item-Infits dargestellt (vgl. Abbildung 11-11). Die erwartete Lösungshäufigkeit pro Item oder Antwortkategorie ist jeweils mit einer durchgezogenen Linie dargestellt, die empirische beobachtete Antworthäufigkeit mit einer gestrichelten Linie zwischen 10 Beobachtungspunkten.

Item AHV_13n hat einen *Underfit* ($\text{wMNSQ} = 1.17 > 1.00$); die beobachtete ICC verläuft flacher als die erwartete. Dieses Item vermag nicht zwischen unterschiedlichen Personenfähigkeiten θ zu diskriminieren: Eine Person mit $\theta = 0$ kann dieses Item mit $p = .85$ lösen, eine Person mit $\theta = 1$ mit $p = .95$. Zum Vergleich, das Item MAN_5 hat einen *perfect fit* ($\text{wMNSQ} = 1.00$); die beiden ICC verlaufen – wenn man durch die beobachtete ICC eine Trendlinie denkt – gleich. Die Lösungswahrscheinlichkeit bei einer Personenfähigkeit von $\theta = 0$ ist $p = .40$, bei $\theta = 1$ $p = .65$. Das Item STA_9n schliesslich hat einen *Overfit* ($\text{wMNSQ} = 0.95 < 1.00$); die beobachtete ICC verläuft steiler als die erwartete. Bei einer Personenfähigkeit von $\theta = 0$ ist $p = .30$, bei jener von $\theta = 1$ ist $p = .55$.

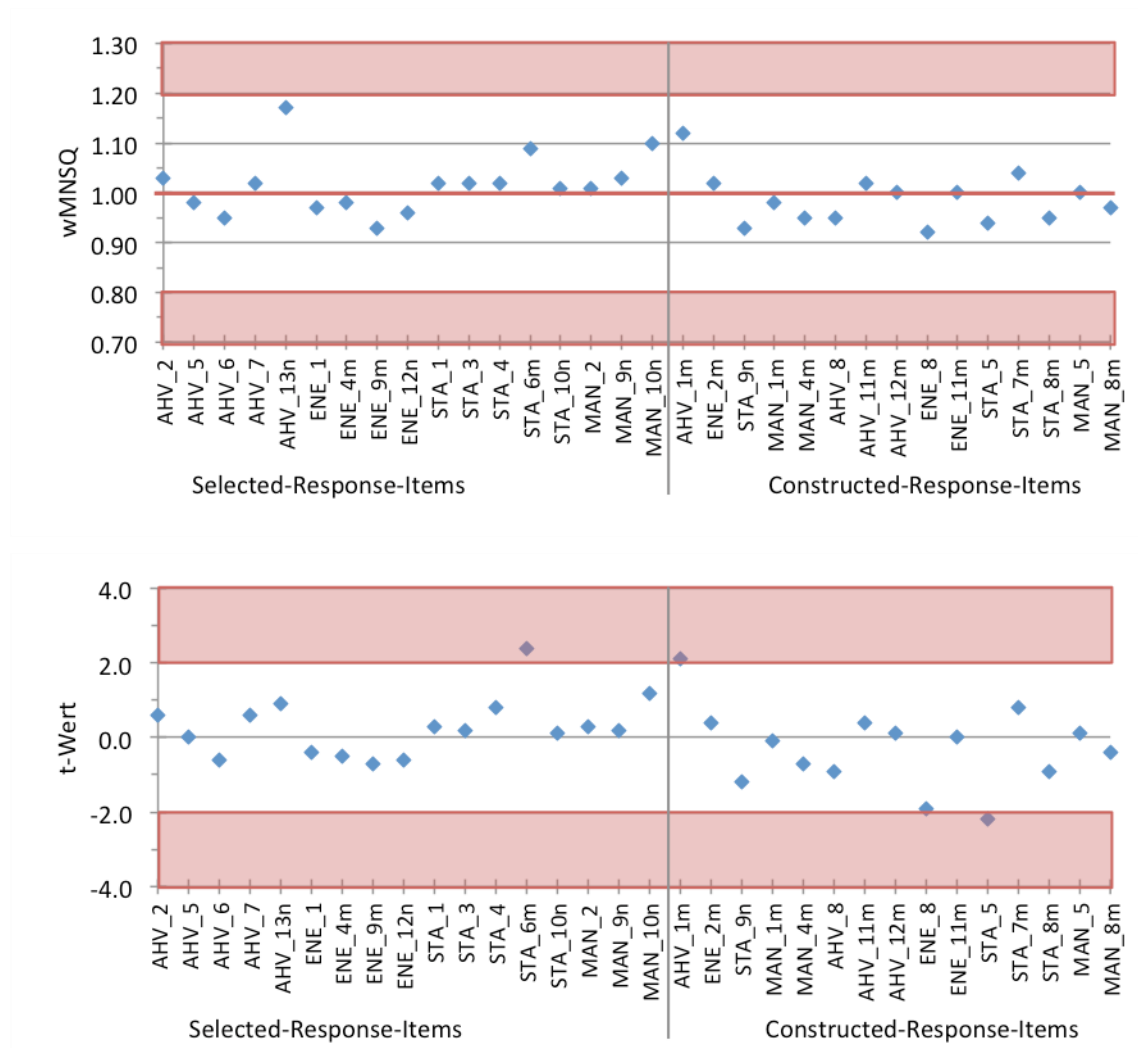


Abbildung 11-10. Item-Infits (wMNSQ) und Signifikanz (t-Statistik)

Hinweis: $m = 32$ Items. $wMNSQ > 1.20$: Underfit, $wMNSQ < 0.80$: Overfit. $wMNSQ \neq 1$ signifikant: $t\text{-Wert} > |1.96|$ bzw. $p < .05$, $wMNSQ \neq 1$ nicht signifikant: $t\text{-Wert} < |1.96|$ bzw. $p > .05$.

Tabelle 11-4. Items und Antwortkategorien mit (geringem) Underfit

Item	Itemparameter σ (in Logits) (SE σ)	Kategorie c	Kategorienparameter σ_c (in Logits)	wMNSQ (CI 95)	t-Wert
AHV_1m	-0.001 (0.048)	---	---	1.12 (0.89, 1.11)	2.1
	---	2	0.26	1.10 (0.92, 1.08)	2.5
AHV_13n	-1.229 (0.056)	---	---	1.17 (0.63, 1.37)	0.9
	---	1	-1.70	1.12 (0.73, 1.27)	0.9
MAN_10n	0.735 (0.046)	---	---	1.10 (0.84, 1.16)	1.2

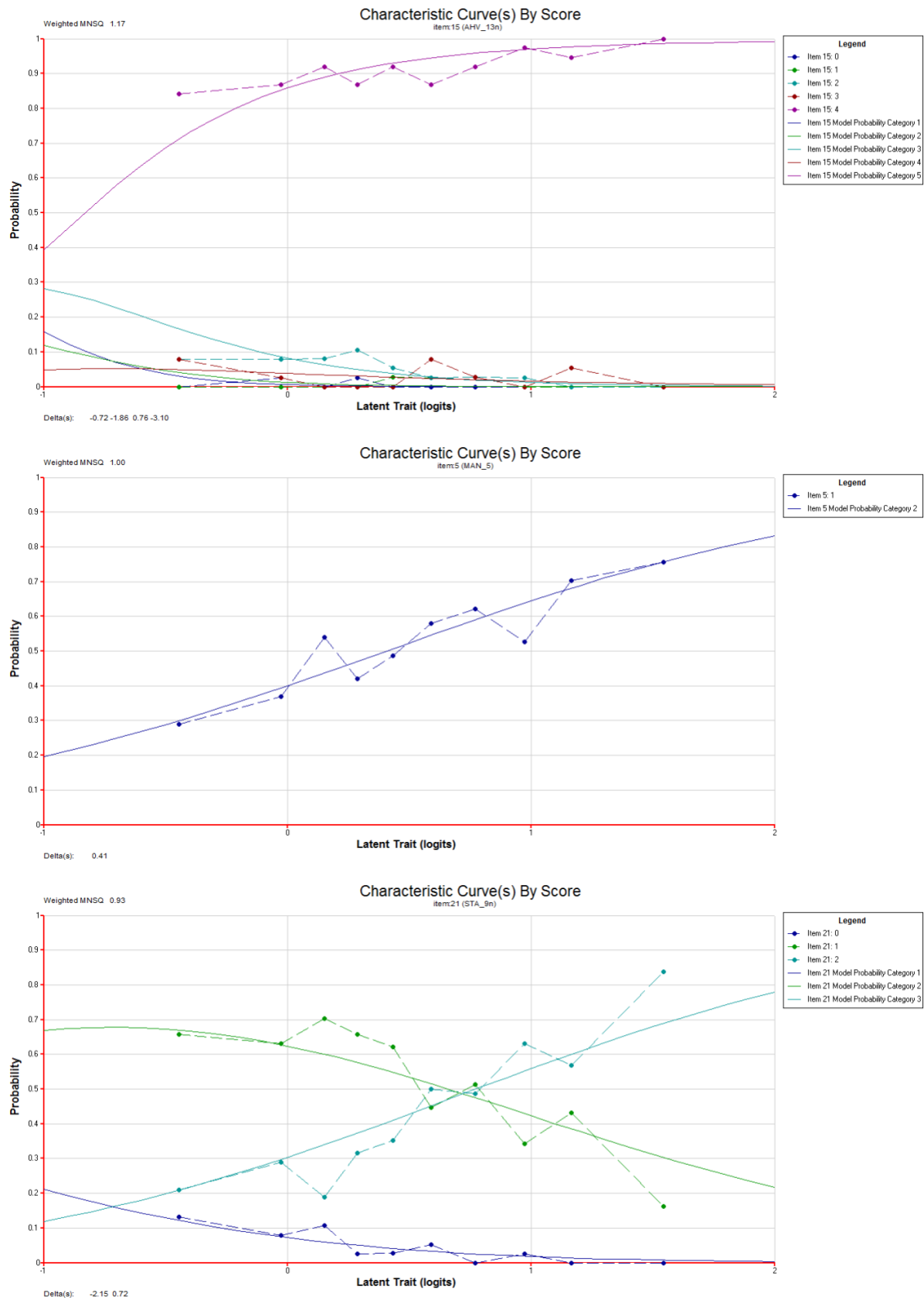


Abbildung 11-11. Item Characteristic Curve (ICC) für drei ausgewählte Items des WBK-T2

Hinweis: Item AHV_13n mit Underfit (oben), Item MAN_5 mit perfect fit (Mitte) und Item STA_9n mit Overfit (unten). Durchgezogene Linie: erwartete Lösungshäufigkeit pro Item oder Antwortkategorie. Gestrichelte Linie: empirische beobachtete Antworthäufigkeit zwischen 10 Beobachtungspunkten.

Tabelle 11-5. Items bzw. Antwortkategorien mit ungeordneten Schwellenparametern

Item	Itemparameter σ (in Logits) (SE σ)	Kategorie #	Schwellen #	Schwellenparameter τ
AHV_1m	-0.001 (0.048)	3	2	0.65 -0.65
AHV_11m	-0.162 (0.075)	3	2	0.09 -0.41
AHV_13n	-1.229 (0.056)	5	4	-0.72 -1.86 0.76 -3.10
ENE_12n	0.159 (0.058)	5	4	-0.07 -0.42 0.15 0.97
STA_10n	-0.513 (0.096)	5	4	-0.50 -1.43 -0.46 0.33

11.4.2 Schwellenparameter

Um die Gültigkeit des Partial-Credit-Rasch-Modells hinsichtlich der polytom geordneten Antwortkategorien zu prüfen, werden die Schwellenparameter herangezogen. Ein Schwellenparameter ist der Schnittpunkt von zwei CC-Funktionen: Bei zwei Antwortkategorien ergibt sich ein Schwellenparameter⁹², bei drei Antwortkategorien zwei Schwellenparameter etc. Wenn das Partial-Credit-Rasch-Modell gilt, dann müssen die Schwellenparameter aufsteigend angeordnet sein. Im Partial-Credit-Rasch-Modell wird jedoch eine ansteigende Anordnung der Schwellenparameter nicht spezifiziert und inferenzstatistisch abgesichert. Die Anordnung der Schwellenparameter kann jedoch deskriptivstatistisch untersucht werden (Strobl, 2010, S. 56).

Für den WBK-T2 zeigen fünf der 15 polytomen Items ungeordnete Schwellenparameter (vgl. Tabelle 11-5), davon gehören drei zum Typus VI (AHV_13n, ENE_12n, STA_10n). Für diese Items bzw. Antwortkategorien ist das Partial-Credit-Rasch-Modell also nicht strikt gültig. Bezeichnenderweise haben diese Items auch unbefriedigende klassische Item-Schwierigkeiten und/oder Item-Trennschärfen (vgl. Kapitel 11.5.1 und 11.5.3).

11.4.3 Differential Item Functioning (DIF)

Für die Prüfung der Item-Homogenität wird *Differential Item Functioning* (DIF) für zwei Stichprobengruppen untersucht. Die erste DIF-Analyse bezieht sich auf die Subgruppen mit unterschiedlichem curricularem Vorwissen (Nicht-WuR, WuR), die zweite auf die Subgruppen des biologischen Geschlechts (Mädchen, Jungen). Diese Auswahl der Stichprobengruppen drängt sich aufgrund der Beschreibung des Konstrukts WBK-soek auf (vgl. Kapitel 8.1.2, Annahme X und Annahme XI).

⁹² Dies ist der Fall bei dichotomen Items mit den zwei Antwortkategorien 0 und 1.

11.4.3.a DIF bzgl. Schwerpunktfach: Gruppe WuR und Gruppe Nicht-WuR

Analog zur Definition der Zielgruppe für den WBK-T2 in den Testspezifikationen (vgl. Kapitel 8.1.3), wurde für die DIF-Analyse nach dem curricularen Vorwissen bzw. dem Schwerpunktfach die Gruppe WuR als Fokusgruppe (*focal group*) und die Gruppe Nicht-WuR als Referenzgruppe (*reference group*) definiert.

Für die grafische DIF-Methode wurden die Item-Schwierigkeitsparameter in getrennten Modellen geschätzt und in einem Koordinatensystem abgetragen: die Gruppe WuR auf der Y-Achse, die Gruppe Nicht-WuR auf der X-Achse (vgl. Abbildung 11-12). Items *unterhalb* der durchgezogenen 45°-Diagonalen funktionieren für die Gruppe WuR einfacher (*DIF against reference group*): $\sigma_{i(WuR)} < \sigma_{i(Nicht-WuR)}$. Items *oberhalb* der durchgezogenen 45°-Diagonalen funktionieren für die Gruppe Nicht-WuR einfacher (*DIF against focal group*): $\sigma_{i(Nicht-WuR)} < \sigma_{i(WuR)}$. Insgesamt befinden sich sieben Items ausserhalb der beiden gestrichelten Diagonalen ($|\Delta \sigma| > 0.638$ Logits) und deuten auf DIF bezüglich Schwerpunktfach hin: vier Items zugunsten WuR (ENE_4m, STA_4, MAN_4m, MAN_9), drei Items zugunsten Nicht-WuR (AHV_1m, AHV_13n, STA_3). Alle sieben Items variieren zwischen den Problemsituationen, jedoch haben sechs Items ein gebundenes Antwortformat.

Das Ergebnis der grafischen DIF-Methode wird mit statistischen Tests der numerischen DIF-Methode abgesichert. Es wird ein restriktives Modell (M1) geschätzt, bei dem kein Interaktionseffekt zwischen den Itemparametern und der Gruppierungsvariablen Schwerpunktfach zugelassen ist, d.h. für die beiden Gruppen WuR und Nicht-WuR werden die gleichen Schwierigkeits- und Schwellenparameter unterstellt und gemeinsam geschätzt. Im weniger restriktiven Modell (M2) wird ein Interaktionseffekt zugelassen, d.h. die Schwierigkeits- und Schwellenparameter der beiden Gruppen werden als unterschiedlich unterstellt und getrennt geschätzt (Adams & Wu, 2010b; Wu et al., 2007, S. 89ff.).⁹³ Der χ^2 -Test auf Parameter-Gleichheit im M2 indiziert, dass ein signifikanter Interaktionseffekt zwischen den Itemparametern und dem Schwerpunktfach vorliegt ($\chi^2_{(31)} = 88.39$, $p < .001$). Der χ^2 -Differenz-Test von M2 und M1 bestätigt ebenfalls, dass die Gruppen WuR und Nicht-WuR systematisch unterschiedliche Itemparameter aufweisen (Δ Deviance = 233, Δ df = 55, $p < .001$) (vgl. Tabelle 11-6).

⁹³ Im Syntax-File von ACER ConQuest 4 wurde M1 spezifiziert als „item+item*step“ und M2 als „item-major+item*major+item*step*major“.

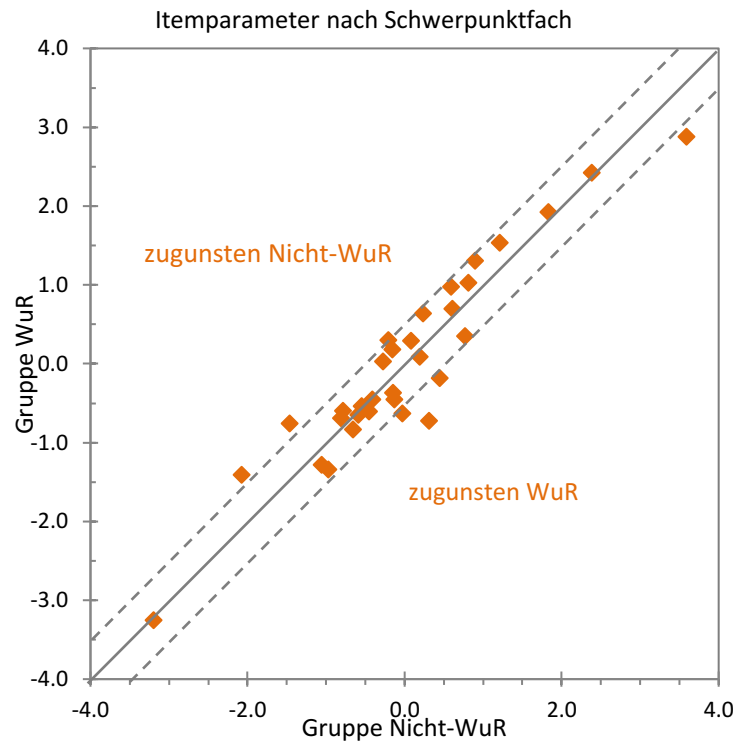


Abbildung 11-12. Grafischer Modelltest für DIF bezüglich Schwerpunktfach

Hinweis: $m = 32$ Items. Durchgezogene Diagonale: $|\Delta \sigma| = 0$ Logits. Gestrichelte Diagonale: $|\Delta \sigma| = 0.638$

Logits. Koordinatenpunkte oberhalb Diagonale: DIF zugunsten Nicht-WuR, d.h. $\sigma_{WuR} > \sigma_{Nicht-WuR}$.

Koordinatenpunkte unterhalb Diagonale: DIF zugunsten WuR, d.h. $\sigma_{Nicht-WuR} > \sigma_{WuR}$.

Tabelle 11-6. Numerischer Modelltest auf globaler Ebene für DIF bezüglich Schwerpunktfach

	M1: ohne Interaktionseffekt, gleiche Item- und Schwellenparameter	M2: mit Interaktionseffekt, unterschiedliche Item- und Schwellenparameter
X ² -Statistik	---	88.39
Freiheitsgrade (df)	---	31
Signifikanz (p, H0: kein Interaktionseffekt, H1: Interaktionseffekt)	---	< .001
Deviance (-2*log-Likelihood)	17'262	17'029
AIC (-2*log-Likelihood + 2*k _{par})	17'374	17'251
Geschätzte Parameter (k _{par})	56	111
Freiheitsgrade (df = k _{par} -1)	55	110
X ² -Differenz-Statistik (Δ Deviance Δ df)	---	233 55
Signifikanz (p, H0: M2 = M1)	---	< .001

Der DIF-Rasch-Test im Modell M2 liefert generelle Item-Schwierigkeitsparameter und Interaktionsparameter. Der Interaktionsparameter einer Gruppe (δ_{ig}) muss zum generellen Schwierigkeitsparameter (σ_i) je nach Vorzeichen addiert bzw. subtrahiert werden, um die Item-Schwierigkeit für diese Gruppe zu erhalten.⁹⁴ Die Grösse des DIF berechnet sich aus der Differenz der beiden gruppenspezifischen Item-Schwierigkeitsparameter $DIF_i = \sigma_{i(reference\ group)} - \sigma_{i(focal\ group)}$ bzw. aus der Differenz der beiden Interaktionsparameter $DIF_i = \delta_{i(reference\ group)} - \delta_{i(focal\ group)}$ (Adams & Wu, 2010b; Wu et al., 2007, S. 92ff.).

Die Interaktionsparameter δ_i für die beiden Gruppen WuR und Nicht-WuR streuen im Logit-Bereich [-0.328, 0.516]; der Mittelwert der Parameter ist aufgrund der Summennormierung auf 0 fixiert ($M = 0$ Logits) (vgl. Anhang Tabelle A- 8). Insgesamt liegen acht Items im „problematischen“ DIF-Bereich ($|\Delta \sigma| > 0.426$) (vgl. Abbildung 11-13). Die Interaktionsparameter sind jedoch nur bei vier Items signifikant ($z > 1.96$), davon hat nur eines einen grossen DIF-Effekt (vgl. Tabelle 11-7). Drei Items haben einen DIF-Effekt zugunsten der Gruppe WuR (C+, B+) (ENE_4m, STA_4, MAN_4m); ein Item zugunsten der Gruppe Nicht-WuR (B-) (AHV_1m). Von diesen vier Items fällt jedes in eine andere Problemsituation (AHV, ENE, STA, MAN).

Der X^2 -Test für die Personenparameter deutet zudem darauf hin, dass eine signifikante Differenz in der mittleren Personenfähigkeit der beiden Gruppen WuR und Nicht-WuR vorliegt ($|\Delta \theta| = 0.490$ Logits, $SE = 0.031$; $X^2_{(1)} = 60.88$, $p < .001$). Personen der Gruppe WuR haben eine systematisch höhere Fähigkeit als Personen der Gruppe Nicht-WuR ($\theta_{WuR} = 0.254$ Logits, $\theta_{Nicht-WuR} = -0.254$ Logits, $SE = 0.031$). Der Effekt ist nach den ETS-Kategorien als moderat zu beurteilen (Zwick, Thayer & Lewis, 1999; Paek & Wilson, 2011, S. 1028, S. 1039). Diese Frage wird weiter unten im Rahmen der kriterialen Validierung auch für die Personen-Testwerte untersucht (vgl. Kapitel 12.1).

⁹⁴ Beispiel anhand Item ENE_4m (vgl. Tabelle 11-7): Der positive Interaktionsparameter der Gruppe Nicht-WuR beträgt $\delta_{ENE_4m(Nicht-WuR)} = 0.516$ und wird zum generellen Schwierigkeitsparameter $\sigma_{ENE_4m} = -0.209$ addiert, um die Item-Schwierigkeit für die Gruppe Nicht-WuR zu erhalten: $\sigma_{ENE_4m(Nicht-WuR)} = 0.307$. Der negative Interaktionsparameter der Gruppe WuR $\delta_{ENE_4m(Nicht-WuR)} = -0.516$ wird vom generellen Schwierigkeitsparameter $\sigma_{ENE_4m} = -0.209$ subtrahiert, so dass $\sigma_{ENE_4m(WuR)} = -0.725$. Das Item ENE_4m ist also für die Gruppe Nicht-WuR schwieriger als für die Gruppe WuR, $\sigma_{ENE_4m(Nicht-WuR)} = 0.307 > \sigma_{ENE_4m(WuR)} = -0.725$.

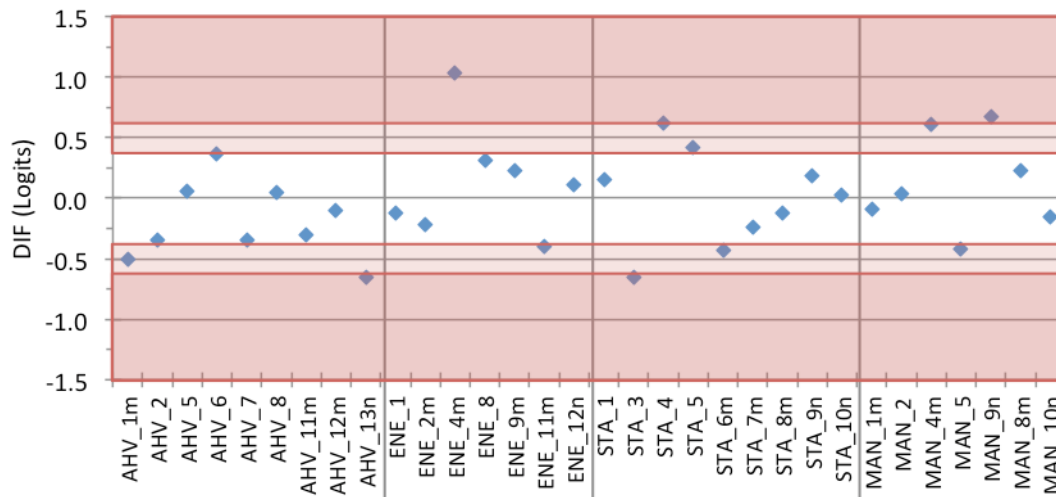


Abbildung 11-13. Numerischer Modelltest für DIF auf lokaler Ebene bezüglich Schwerpunktfach

Hinweis: $m = 32$ Items. Fokusgruppe: WuR, Referenzgruppe: Nicht-WuR.

$DIF > 0$ Logits: DIF zugunsten WuR. $DIF < 0$ Logits: DIF zugunsten Nicht-WuR. $0.426 \text{ Logits} < |DIF| < 0.638$

Logits: DIF Kategorie B+ und B– (moderater Effekt). $|DIF| \geq 0.638$ Logits: DIF Kategorie C+ und C– (grosser Effekt).

Tabelle 11-7. Numerischer Modelltest auf lokaler Ebene für DIF bezüglich Schwerpunktfach

Item	Itemparameter σ (in Logits) (SE σ)	Interaktionsparameter δ (in Logits) (SE δ)	z-Statistik	DIF (in Logits)	DIF Kategorie
AHV_1m	0.052 (0.060)	-0.254 (0.070)	-3.629	-0.508	B–
ENE_4m	-0.209 (0.123)	0.516 (0.123)	4.195	1.032	C+
STA_4	0.135 (0.113)	0.311 (0.113)	2.752	0.622	B+
MAN_4m	-0.331 (0.106)	0.304 (0.108)	2.815	0.608	B+

Hinweis: DIF-Rasch-Test. Fokusgruppe WuR, Referenzgruppe Nicht-WuR. δ der Gruppe Nicht-WuR.

$DIF = \delta_{\text{Nicht-WuR}} - \delta_{\text{WuR}}$. $DIF > 0$ Logits: DIF zugunsten WuR. $DIF < 0$ Logits: DIF zugunsten Nicht-WuR.

DIF Kategorie B+ und B– (moderater Effekt): $0.426 \text{ Logits} < |DIF| < 0.638 \text{ Logits}$ und $z > 1.96$. DIF Kategorie C+ und C– (grosser Effekt): $|DIF| \geq 0.638 \text{ Logits}$ und $z > 1.96$.

Quelle: ETS DIF Kategorien nach Paek & Wilson (2011) und Zwick et al. (1999).

11.4.3.b DIF bzgl. Geschlecht: Jungen und Mädchen

Für die DIF-Analyse bezüglich des biologischen Geschlechts wurden Jungen als Fokusgruppe und Mädchen als Referenzgruppe definiert. Für die grafische DIF-Methode wurden die in getrennten Modellen geschätzten Itemparameter der Jungen auf der Y-Achse abgetragen, jene der Mädchen auf der X-Achse (vgl. Abbildung 11-14). Items *unterhalb* der 45°-Diagonalen funktionieren für Jungen einfacher (*DIF against reference group*, $\sigma_{i(\text{male})} < \sigma_{i(\text{female})}$), jene *oberhalb* funktionieren für Mädchen einfacher (*DIF against focal group*, $\sigma_{i(\text{female})} < \sigma_{i(\text{male})}$). Insgesamt befindet sich nur ein Item ausserhalb der beiden gestrichelten Diagonalen ($|\delta| > 0.638$) und lässt DIF bezüglich des Geschlechts vermuten, und zwar zulasten der Jungen (AHV_5).

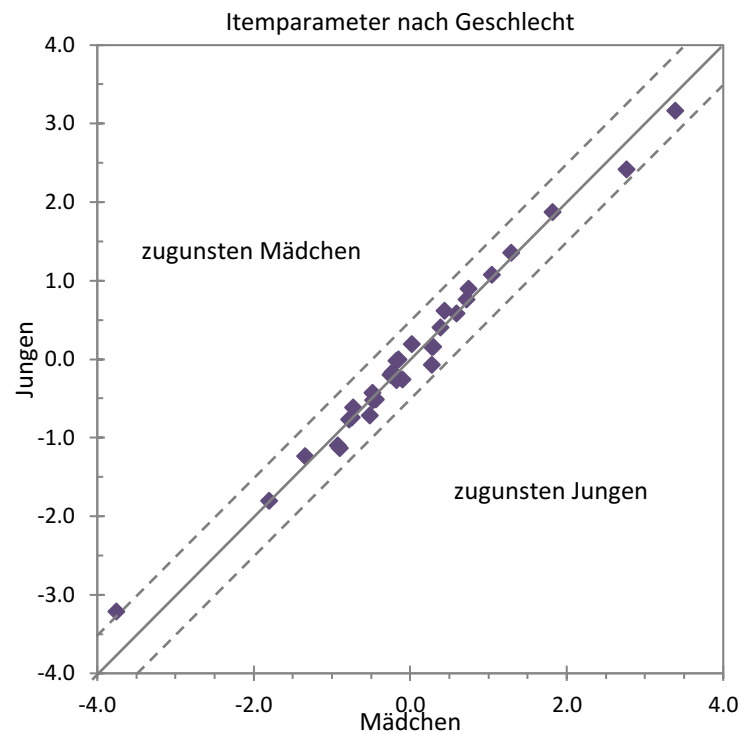


Abbildung 11-14. Grafischer Modelltest für DIF bezüglich Geschlecht (Mädchen, Jungen)

Hinweis: $m = 32$ Items. Durchgezogene Diagonale: $|\Delta \sigma| = 0$ Logits. Gestrichelte Diagonale: $|\Delta \sigma| = 0.638$ Logits. Koordinatenpunkte oberhalb Diagonale: DIF zugunsten Mädchen, d.h. $\sigma_{male} > \sigma_{female}$. Koordinatenpunkte unterhalb Diagonale: DIF zugunsten Jungen, d.h. $\sigma_{female} > \sigma_{male}$.

Tabelle 11-8. Numerischer Modelltest auf globaler Ebene für DIF bezüglich Geschlecht

	M1: ohne Interaktionseffekt, gleiche Item- und Schwellenparameter	M2: mit Interaktionseffekt, unterschiedliche Item- und Schwellenparameter
X ² -Statistik	---	94.68
Freiheitsgrade (df)	---	31
Signifikanz (p, H0: kein Interaktionseffekt, H1: Interaktionseffekt)	---	< .001
Deviance (-2*Log-Likelihood)	17'262	17'073
AIC (-2*log-Likelihood + 2*k _{par})	17'374	17'295
Geschätzte Parameter (k _{par})	56	111
Freiheitsgrade (df = k _{par} -1)	55	110
X ² -Differenz-Statistik (Δ Deviance Δ df)	---	190 55
Signifikanz (p, H0: M2 = M1)	---	< .001

Der X^2 -Test auf Parameter-Gleichheit induziert im Modell M2 mit Interaktionseffekt zwischen den Itemparametern und der Gruppierungsvariablen Geschlecht, dass ein signifikanter Interaktionseffekt zwischen den Itemparametern und dem Geschlecht vorliegt ($X^2_{(31)} = 94.68$, $p < .001$). Der X^2 -Differenz-Test von M2 und M1 bestätigt ebenfalls, dass Jungen und Mädchen systematisch unterschiedliche Itemparameter aufweisen (Δ Deviance = 190, Δ df = 55, $p < .001$) (vgl. Tabelle 11-8).

Beim DIF-Rasch-Test bezüglich des Geschlechts streuen die Interaktionsparameter δ_i der 32 Items im Logit-Bereich $[-0.619, 0.514]$ ($M = 0$ Logits) (vgl. Anhang Tabelle A- 8). Insgesamt liegen 13 Items im „problematischen“ DIF-Bereich ($|\Delta \delta| > 0.426$) (vgl. Abbildung 11-15). Die Interaktionsparameter sind bei neun Items signifikant ($z > 1.96$), davon haben aber nur drei Items einen grossen DIF-Effekt (DIF Kategorie C; AHV_5, ENE_4m, ENE_9m) (vgl. Tabelle 11-9). Drei Items haben einen DIF-Effekt zugunsten der Jungen (C+, B+) und gehören alle zur Problemsituation ENE. Sechs Items haben einen DIF-Effekt zugunsten der Mädchen (C-, B-) und gehören in die Problemsituationen AHV und STA. Aus der Problemsituation MAN zeigt kein Item ein DIF bezüglich des Geschlechts.

Der X^2 -Test für die Personenparameter ($|\Delta \theta| = 0.300$ Logits, $SE = 0.030$ Logits; $X^2_{(1)} = 24.72$, $p < .001$) deutet an, dass eine signifikante Differenz in der mittleren Personenfähigkeit bei Mädchen und Jungen vorliegt. Jungen haben eine systematisch höhere Fähigkeit als Mädchen ($\theta_{\text{male}} = 0.15$ Logits, $\theta_{\text{female}} = -0.15$, $SE = 0.03$). Allerdings ist der Effekt nach den ETS-Kategorien als vernachlässigbar zu beurteilen. Diese Frage wird weiter unten im Rahmen der kriterialen Validierung auch für die Personen-Testwerte untersucht (vgl. Kapitel 12.2.1).

An diesen Ergebnissen zum DIF bezüglich des Geschlechts sind drei Punkte auffallend. Erstens gibt der numerische DIF-Rasch-Test bezüglich des Geschlechts viel mehr Items mit DIF an als es der grafische Modelltest tut (vgl. Abbildung 11-14). Zweitens zeigt der numerische DIF-Rasch-Test mehr Items mit DIF-Effekt zugunsten der Mädchen, obschon die globale Personenfähigkeit der Mädchen geringer als jene der Jungen ist. Drittens gehören alle Items mit DIF zugunsten der Mädchen zu den Problemsituationen AHV und STA, alle Items mit DIF zugunsten der Jungen in die Problemsituation ENE. Ob unterschiedliche Interessen, i.S.v. Auseinandersetzung mit einem Gegenstand (vgl. Kapitel 8.1.4.c), von Mädchen und Jungen für „soziale“ (z.B. Altersvorsorge) und „technische“ (z.B. Energieversorgung) Themen die Lösungswahrscheinlichkeiten bzw. Item-Schwierigkeiten beeinflussen, ist ihm Rahmen dieser Dissertation nicht zu beantworten. Für weiterführende Studien mit dem WBK-T2 wurde mittlerweile eine Skala zum spezifischen Sachinteresse, d.h. zum Interesse an den konkreten Problemsituationen des WBK-T2 entwickelt und erprobt (Ackermann, 2018b).

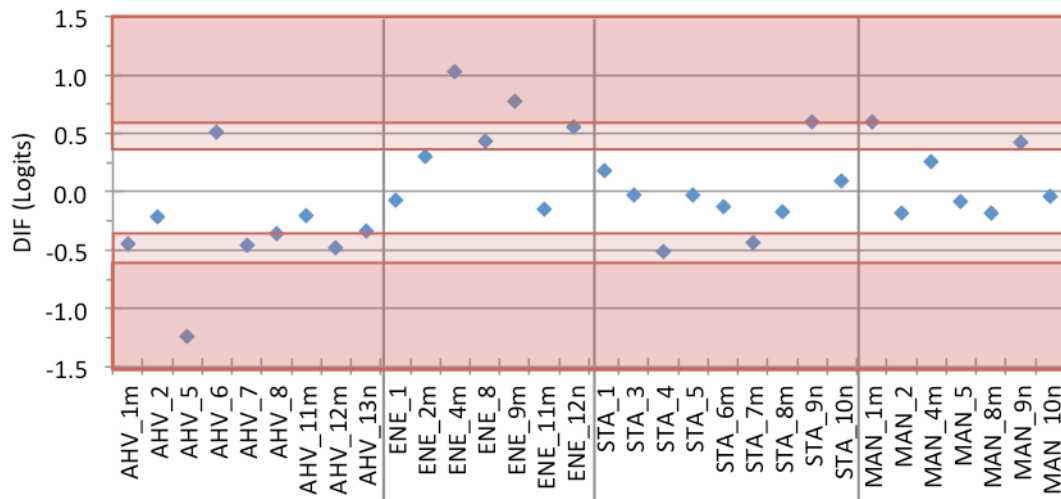


Abbildung 11-15. Numerischer Modelltest für DIF auf lokaler Ebene bezüglich Geschlecht

Hinweis: $m = 32$ Items. Fokusgruppe: Jungen, Referenzgruppe: Mädchen.

$DIF > 0$ Logits: DIF zugunsten Jungen. $DIF < 0$ Logits: DIF zugunsten Mädchen. $0.426 \text{ Logits} < |DIF| < 0.638 \text{ Logits}$: DIF Kategorie B+ und B– (moderater Effekt). $|DIF| \geq 0.638 \text{ Logits}$: DIF Kategorie C+ und C– (grosser Effekt).

Tabelle 11-9. Numerischer Modelltest auf lokaler Ebene für DIF bezüglich Geschlecht

Item	Itemparameter σ (in Logits) (SE σ)	Interaktionsparameter δ (in Logits) (SE δ)	z-Statistik	DIF (in Logits)	DIF Kategorie
AHV_1m	0.070 (0.063)	-0.223 (0.072)	-3.097	-0.446	B–
AHV_5	-3.144 (0.253)	-0.619 (0.315)	-1.965	-1.238	C–
AHV_7	0.052 (0.111)	-0.228 (0.112)	-2.036	-0.456	B–
AHV_12m	0.678 (0.092)	-0.238 (0.094)	-2.532	-0.476	B–
ENE_4m	-0.239 (0.126)	0.514 (0.126)	4.079	1.028	C+
ENE_9m	-1.292 (0.164)	0.387 (0.163)	2.374	0.774	C+
ENE_12n	-0.010 (0.096)	0.280 (0.099)	2.828	0.560	B+
STA_4	0.272 (0.109)	-0.256 (0.110)	-2.327	-0.512	B–
STA_7	0.956 (0.075)	-0.217 (0.079)	-2.747	0.434	B–

Hinweis: DIF-Rasch-Test. Fokusgruppe Jungen, Referenzgruppe Mädchen. δ der Mädchen. $DIF = \delta_{\text{female}} - \delta_{\text{male}}$. $DIF > 0$ Logits: DIF zugunsten Jungen. $DIF < 0$ Logits: DIF zugunsten Mädchen. DIF Kategorie B+ und B– (moderater Effekt): $0.426 \text{ Logits} < |DIF| < 0.638 \text{ Logits}$ und $z > 1.96$. DIF Kategorie C+ und C– (grosser Effekt): $|DIF| \geq 0.638 \text{ Logits}$ und $z > 1.96$.

Quelle: ETS DIF Kategorien nach Paek & Wilson (2011) und Zwick et al. (1999).

Die Ergebnisse zu den DIF-Analysen für den WBK-T2 zeigen, dass empirisch *keine perfekte* Item-Homogenität bezüglich der Substichproben nach Schwerpunktfach und Geschlecht vorliegt. Für die Gruppen WuR und Nicht-WuR ist das vorgefundene DIF eher als unkritisch zu beurteilen, v.a. angesichts der geringen Anzahl betroffener Items und der mehrheitlich moderaten DIF-Effekte. Für die Gruppen Mädchen und Jungen ist das DIF als etwas kritischer zu beurteilen, v.a. wegen der

Anzahl Items, ihrer ungleichen Verteilung zwischen den Gruppen und ihrer ungleichen Verteilung auf die Problemsituationen.

Die theoretischen Annahmen der Rasch-Modelle zur Item-Homogenität und insbesondere zum DIF sind sehr streng. Insofern gelten viele Rasch-Modelle nicht nur in der Stichprobe, sondern auch in der Grundgesamtheit nicht exakt, aber stellen eine gute und praktisch sinnvolle Approximation dar. Die Konsequenzen kleinerer oder grösserer Modellabweichungen müssen stets anhand des Testzwecks beurteilt werden (Eid & Schmidt, 2014, S. 201ff.).

Um eine annähernd „perfekte“ Item-Homogenität zu erreichen, könnten die drei Items mit grossem DIF oder sogar alle zehn Items mit DIF aus der Datenmatrix des WBK-T2 entfernt und – sofern sich das DIF tatsächlich verbessert – die weiteren Analysen mit dem reduzierten Datensatz gerechnet werden. Allerdings muss hier eine Abwägung zwischen formalen Anforderungen i.S.v. Modellgültigkeit und inhaltlichen Anforderungen i.S.v. Konstrukt-Repräsentativität vorgenommen werden. Das Eliminieren von derart vielen Items hätte nämlich zur Konsequenz, dass das Konstrukt WBK-soek nicht mehr repräsentativ und adäquat mit dem WBK-T2 operationalisiert wäre und die durch Domänenanalyse und Expertenbeurteilung erreichte Inhaltsvalidität nicht mehr gewährleistet wäre (vgl. Kapitel 8.2 und 9). Für den formulierten Testzweck (vgl. Kapitel 8.1) ist m.E. das vorliegende DIF im WBK-T2 „akzeptabel“ und der WBK-T2 als „nahezu“ modellkonform zu betrachten.

Annahme VI zur Modellgültigkeit kann also als erfüllt betrachtet werden. Auch diese Evidenz ist aus forschungspraktischer Perspektive als positiv zu werten, da Item-Homogenität des Tests eine Voraussetzung für weitere statistische Analysen und die valide Interpretation der Testrohwerte ist.

11.5 Verteilung der Testwerte und Qualität der Items

Welche Evidenz ergibt sich für die Verteilung der Testwerte im WBK-T2? Die Testwerte haben eine Normalverteilung (vgl. Annahme VII). Welche Evidenz ergibt sich für die Qualität der Items im WBK-T2? Die Items des WBK-T2 decken den mittleren Schwierigkeitsbereich ab (klassische Item-Schwierigkeit) und unterscheiden zwischen Testpersonen mit hohen und geringen Testwerten (klassische Item-Trennschärfe) (vgl. Annahme VIII).

Die Item-Qualität gilt als gut, wenn (Döring & Bortz, 2016; Kelava & Moosbrugger, 2012): (1) die klassischen Item-Schwierigkeiten möglichst breit im Wertebereich streuen und keine extremen Werte annehmen ($0.20 \leq x \leq .80$), (2) die klassischen Item-Trennschärfen genügend hoch sind ($x \geq .30$) und (3) die Distraktoren-Trennschärfen negativ sind. Die Ergebnisse der klassischen Itemanalyse sind zentral für die Testentwicklung und -evaluation. Dabei sollen die Item-Schwierigkeiten und die Item-Trennschärfen simultan berücksichtigt werden.

Um Annahme VII zur Testwerte-Verteilung zu prüfen, wird der summierte Testwert in einem Histogramm dargestellt und statistische Tests auf Normalverteilung gerechnet. Normalverteilung ist eine zentrale Voraussetzung für multivariate Analysen mit der Variable WBK-soek. Um Annahme VIII zur Item-Qualität des WKB-T2 zu prüfen, werden die in den IRT-Analysen (vgl. Kapitel 11.2) mit gelieferten klassischen Item-Kennwerte analysiert.

Der Aufbau dieses Kapitels ist wie folgt: Zuerst wird die Verteilung der Testwerte dargestellt und geprüft und eine kriteriumsorientierte Interpretation der Testwerte vorgenommen (Kapitel 11.5.1). Danach werden die klassischen Item-Schwierigkeiten berechnet und beurteilt (Kapitel 11.5.2). Schliesslich werden die klassischen Item-Trennschärfen aller Items und die Kategorien-Trennschärfen der Selected-Response-Items berechnet und beurteilt (Kapitel 11.5.3).

11.5.1 Testwerte-Verteilung

Die Testwerte des WBK-T2 wurden als Summenwerte aus den korrekt gelösten Item Scores gebildet (Moosbrugger, 2012b, S. 107f.). Je höher der Testwert, desto mehr Items wurden von der Testperson in der verfügbaren Zeit (ca. 60 Minuten, vgl. Kapitel 8.5.2) korrekt gelöst. Ein geringer Testwert bedeutet, dass die Testperson bestimmte Items nicht korrekt oder gar nicht gelöst hat. Im WBK-T2 konnten maximal 55 Punkte erzielt werden. Für einen Vergleich der Testwerte nach Problemsituationen und Itemtypen werden zudem die adjustierten Testwerte aus den korrekt gelösten Item Scores gebildet (Kelava & Moosbrugger, 2012, S. 89ff.).

11.5.1.a Prüfung auf Normalverteilung

Die Häufigkeitsverteilung der Testwerte im WBK-T2 wird durch ein Histogramm dargestellt (vgl. Abbildung 11-16), die entsprechenden Kennwerte der Häufigkeitsverteilung mittels deskriptiver Analysen berechnet (vgl. Tabelle 11-10). Der Kolmogoroff-Smirnov-Test und der Shapiro-Wilk-Test zur Prüfung der Normalverteilung in der Grundgesamtheit zeigen für die Testwerte keine Normalverteilung an ($K-S_{(375)} = 0.064$, $p < .001$; $S-W_{(375)} = 0.984$, $p < .001$). Da die Stichprobengrösse des Projekts WBKgym ($N = 375$) grösser als 120 und „nahe“ bei 400 liegt, wurde für die Prüfung der Normalverteilung zusätzlich der *Critical-Ratio*-Test durchgeführt. Für die vorliegende Stichprobe ergibt der *Critical-Ratio*-Test für die Schiefe eine horizontale bzw. axialsymmetrische Verzerrung ($z = -2.960$, linksschief), für die Steilheit keine vertikale Verzerrung ($z = -0.175$) an. Die direkten Grenzwerte implizieren weder für Schiefe noch für Steilheit eine Verzerrung. Daher werden die Testwerte für die nachfolgenden inferenzstatistischen Analysen (vgl. Kapitel 12.1, 12.2 und 12.3) als annähernd normalverteilt angenommen.

Bei der Gruppe Nicht-WuR indiziert der CR-Test sowohl für Schiefe ($z = -0.886$) als auch für Steilheit ($z = 0.494$) keine Verzerrung, so dass von einer Normalverteilung ausgegangen werden kann. Bei der Gruppe WuR indiziert der CR-Test für die Schiefe eine horizontale bzw. axialsymmetrische Verzerrung ($z = -2.385$, linksschief), aber für die Steilheit keine vertikale Verzerrung ($z = -1.235$), so dass von einer annähernden Normalverteilung ausgegangen wird.

11.5.1.b Testwerte nach Problemsituation

Im WBK-T2 mit den insgesamt 32 Items aus den vier Problemsituationen Altersvorsorge (AHV), Energieversorgung (ENE), Staatsverschuldung (STA) und Managervergütungen (MAN) konnten maximal 55 Punkte erzielt werden. Der mittlere Testwert lag bei 35 Punkten ($M = 35.05$, $SD = 6.60$), die geringste bei 10 ($\min = 10$) und die höchste bei 49 ($\max = 49$) (vgl. Tabelle 11-11). Die Lösungsquote im gesamten WBK-T2 beträgt 64 % (vgl. Tabelle 11-11).

Beim Thema AHV ist die Lösungsquote am höchsten ($LQ_{AHV} = 74\%$), entsprechend ist hierbei die Kompetenz der Testpersonen am höchsten. Das Thema Managervergütung hat die geringste Lösungsquote ($LQ_{MAN} = 58\%$) und die Kompetenz der Testpersonen am geringsten. Die Lösungsquoten der Themen Staatsverschuldung ($LQ_{STA} = 60\%$) und Energieversorgung ($LQ_{ENE} = 64\%$) liegen dazwischen.

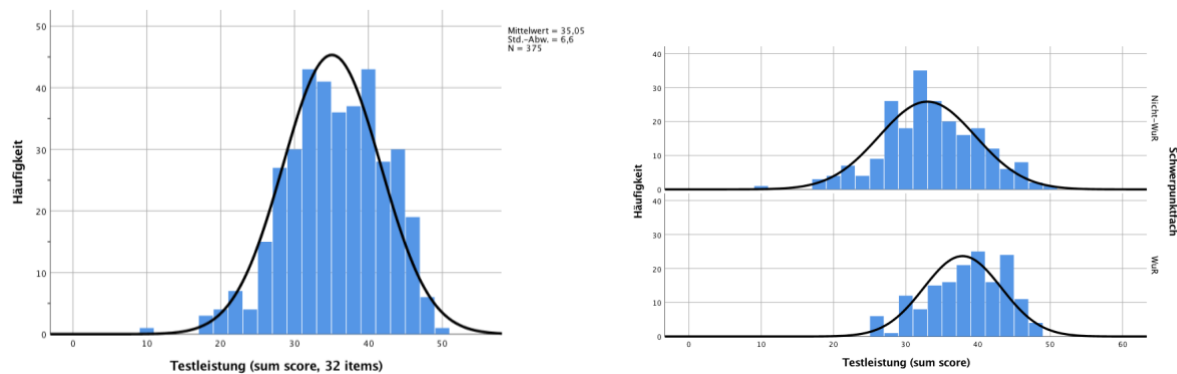


Abbildung 11-16. Häufigkeitsverteilung der Testwerte für Stichprobe WBKgym

Hinweis: $N = 375$ Cases, $m = 32$ Items. Gesamte Stichprobe (oben): $N = 375$, $M = 35.05$, $SD = 6.60$, $\min = 10$, $\max = 49$. Gruppe Nicht-WuR (mitte): $N = 216$, $M = 32.98$, $SD = 6.67$, $\min = 10$, $\max = 49$. Gruppe WuR (unten): $N = 159$, $M = 37.86$, $SD = 5.36$, $\min = 25$, $\max = 47$.

Tabelle 11-10. Deskriptive Kennwerte der Testwerte

Gruppe	N	M	SE M	SD	Schiefe (SE Schiefe)	z-Wert Schiefe	Steilheit (SE Steilheit)	z-Wert Steilheit
Total	375	35.05	0.34	6.60	-0.373 (0.126)	-2.960	-0.044 (0.251)	-0.175
Nicht-WuR	216	32.98	0.45	6.67	-0.147 (0.166)	-0.886	0.163 (0.330)	0.494
WuR	159	37.86	0.43	5.36	-0.458 (0.192)	-2.385	-0.473 (0.383)	-1.235

Tabelle 11-11. Testwerte und Lösungsquote nach Problemsituation

	N	maxScore	sumScore	min	max	LQ	M _{adj}	SD
Altersvorsorge (AHV)	375	15	11.03	4	15	74 %	0.70	0.16
Energieversorgung (ENE)	365	12	7.69	0	12	64 %	0.68	0.19
Staatsverschuldung (STA)	375	15	9.01	1	15	60 %	0.59	0.16
Managervergütung (MAN)	375	13	7.53	1	11	58 %	0.53	0.16
Total (WBK-soek)	375	55	35.05	10	49	64 %	0.63	0.12

Hinweis: maxScore = maximal mögliche Punktezahl. sumScore = effektiv erreichte Punktezahl. min = geringste erreichte Punktezahl. max = höchste erreichte Punktezahl. LQ = Lösungsquote, d.h. erreichte Punkte/maximale Punkte*100. M_{adj} = adjustierter Mittelwert für dichotome und polytome Items, Wertebereich [0,1].

Der adjustierte Mittelwert (zur Item-Schwierigkeit vgl. Kapitel 5.2.1.b) des gesamten WBK-T2 liegt bei $M = 0.63$ ($SD = 0.12$) (vgl. Tabelle 11-11). Die Problemsituationen AHV ($M_{AHV} = 0.70$) und ENE ($M_{ENE} = 0.68$) haben einen überdurchschnittlichen Mittelwert, die Problemsituationen STA ($M_{STA} = 0.59$) und MAN ($M_{MAN} = 0.53$) einen leicht überdurchschnittlichen. Die Varianz zwischen den Problemsituationen gemessen an den adjustierten Item-Mittelwerten ist nicht signifikant grösser als die Varianz innerhalb der Problemsituationen ($F_{(3;28)} = 1.039$, $p = 0.391$, $\eta^2 = 0.10$). Interessanterweise

korrespondieren die mittleren Item-Mittelwerte der Problemsituationen mit den mittleren probabilistischen Item-Schwierigkeitsparameter der Problemsituationen (vgl. Kapitel 11.2.3).

Eine Erklärung für diese Rangfolge könnte in der Repräsentativität und Bekanntheit (vgl. Kapitel 8.2.1, vgl. Anhang Tabelle A- 2) der Problemsituationen begründet liegen. Die Problemsituationen AHV und ENE scheinen bei den Testpersonen der Stichprobe einen stärkeren Lebensweltbezug⁹⁵ und Lehrplanbezug (VWL: Sozialpolitik, Sozialversicherungen. VWL: Wirtschaftspolitik, Ökonomie & Ökologie) zu haben. Umgekehrt hat die Problemsituation MAN einen schwächeren Lebensweltbezug⁹⁶ und Lehrplanbezug (BWL: Personalwesen, Lohnformen; VWL: Marktwirtschaft), insbesondere für das Einführungsfach WuR.

11.5.1.c Kriteriumsorientierte Interpretation der Testwerte

Für die kriteriumsorientierte Interpretation der Personen-Testwerte im WBK-T2 wird die Lösungsquote bzw. der adjustierte Testwert verwendet (vgl. Döring & Bortz, 2016, S. 433; Kelava & Moosbrugger, 2012, S. 94f.). Ein adjustierter Testwert von 50 % wird als *minimal standard* interpretiert; unter 50 % gilt als geringe WBK-soek (*low competence level*), zwischen 50 % und 70 % als mittlere WBK-soek (*moderate competence level*) und über 70 % als hohe WBK-soek (*high competence level*) (vgl. Kapitel 8.1.2).

In der Stichprobe WBKgym zeigen 31 % der Testpersonen (N = 118) im WBK-T2 eine hohe WBK-soek, 55 % (N = 207) eine mittlere WBK-soek und 13.3 % (N = 50) eine geringe WBK-soek (vgl. Tabelle 11-12). Mit anderen Worten, 86.1 % der Testpersonen (N = 325) erreichen den „minimal standard“ von ≥ 50 %, 13.3 % (N = 50) erreichen ihn nicht. Der mittlere Testwert der gesamten Stichprobe ($M = 35.05$, $M_{adj} = 0.63$) liegt im mittleren Kompetenzbereich, ebenfalls die mittleren Testwerte der Gruppe Nicht-WuR ($M = 32.98$, $M_{adj} = 0.59$) und der Gruppe WuR ($M = 37.86$, $M_{adj} = 0.67$) (vgl. Abbildung 11-16, Abbildung 12-1).

⁹⁵ Volksabstimmung vom 24.09.2017 über die Rentenreform 2020 (Eidgenössische Bundeskanzlei, <https://www.bk.admin.ch/ch/d/pore/rf/cr/2014/20141890.html>, 31.03.2019). Volksabstimmung vom 21.05.2017 über das Energiegesetz bzw. die Energiestrategie 2050 (Eidgenössische Bundeskanzlei, <https://www.bk.admin.ch/ch/d/pore/rf/cr/2012/20121295.html>, 31.03.2019).

⁹⁶ Volksabstimmung vom 03.03.2013 über die Abzocker-Initiative (Eidgenössische Bundeskanzlei, <https://www.bk.admin.ch/ch/d/pore/vi/vis348.html>, 31.03.2019). Volksabstimmung vom 24.11.2013 über die 1:12 Initiative (Eidgenössische Bundeskanzlei, <https://www.bk.admin.ch/ch/d/pore/vi/vis375.html>, 31.03.2019).

Tabelle 11-12. Kompetenzkategorien nach Lösungsquote im WBK-T2 (kriteriumsorientierte Interpretation)

Kompetenzkategorie	Kriterium (summierter Testwert)	Kriterium (adjustierter Testwert)	N	% gültig	% kumuliert (Prozentrang)
Geringe WBK-soek	$0.0 \leq x < 27.5$	$0.00 \leq x < 0.50$	50	13	13
Mittlere WBK-soek	$27.5 \leq x < 38.5$	$0.50 \leq x < 0.70$	207	55	69
Hohe WBK-soek	$38.5 \leq x \leq 55.0$	$0.70 \leq x \leq 1.00$	118	31	100

Hinweis: Adjustierter Testwert aus dem adjustierten Mittelwert (M_{adj}) für dichotome und polytome Items, Wertebereich $[0,1]$. Minimal standard der WBK-soek: $M_{adj} = 0.50$.

11.5.2 Klassische Item-Schwierigkeiten

Die Schwierigkeit eines Items (*item difficulty*) ermöglicht eine Aussage darüber, wie gut das Item die Differenzierung im Antwortverhalten der Testpersonen erfassen kann. Ein Item muss also derart konstruiert sein, dass nicht alle Testpersonen dieselbe Antwort geben, d.h. es darf weder zu einfach noch zu schwierig sein (Kelava & Moosbrugger, 2012, S. 76). Für die Beurteilung der klassischen Item-Schwierigkeiten im WBK-T2 werden die Item-Mittelwerte herangezogen. Die Item-Mittelwerte wurden entsprechend den Itemtypen (gebundenes/freies Antwortformat, dichotome/polytome Kodierung) (vgl. Tabelle 8-4) als adjustierte Item-Mittelwerte und/oder korrigierte Item-Mittelwerte berechnet (vgl. Kapitel 5.2.1.b, Item-Schwierigkeit): Für Constructed-Response-Items (Typus II, III, IV und V) ein adjustierter Item-Mittelwert (M_{adj}); für Selected-Response-Items (Typus I und VI) ein adjustierter (M_{adj}) und zusätzlich ein korrigierter Item-Mittelwert (M_{cor}).

Die adjustierten Item-Mittelwerte streuen im Bereich von 0.07 und 0.97 ($M = 0.63$, $SD = 0.22$), die ratekorrigierten adjustierten Item-Mittelwerte im Bereich von 0.00 und 0.96 ($M = 0.58$, $SD = 0.23$) (vgl. Abbildung 11-17, Anhang Tabelle A- 7). Beide Schwierigkeits-Indizes streuen ausreichend über den gesamten Wertebereich. Jedoch haben bei beiden Schwierigkeits-Indizes die Selected-Response-Items eine grössere Streuung als die Constructed-Response-Items.

In der Kategorie „zu einfach“ gibt es beim adjustierten Mittelwert ($M_{adj} \geq 0.80$) sieben Items und beim ratekorrigierten Mittelwert ($M_{cor} \geq 0.80$) vier Items – dabei handelt es sich ausschliesslich um Selected-Response-Items. In der Kategorie „zu schwierig“ gibt es beim adjustierten Mittelwert ($M_{adj} \leq 0.20$) zwei Items und beim ratekorrigierten Mittelwert ($M_{cor} \leq 0.20$) drei Items. Insgesamt erweisen sich die adjustierten Item-Mittelwerte von drei Items als kritisch (AHV_5, AHV_13n, MAN_9n), da sie ihre Werte im oberen oder unteren 5 %-Bereich haben ($M_i \geq 0.95$, $M_i \leq 0.50$) (vgl. Kelava & Moosbrugger, 2012, S. 87) (vgl. Tabelle 11-13). Sofern auch die anderen psychometrischen Indizes, v.a. der Item-Infit (vgl. Tabelle 11-4) und die klassische Item-Trennschärfe (vgl. Tabelle 11-14) dafür sprechen, sollten diese Items bei der nächsten Testrevision modifiziert oder ggf. aus dem Itempool eliminiert werden.

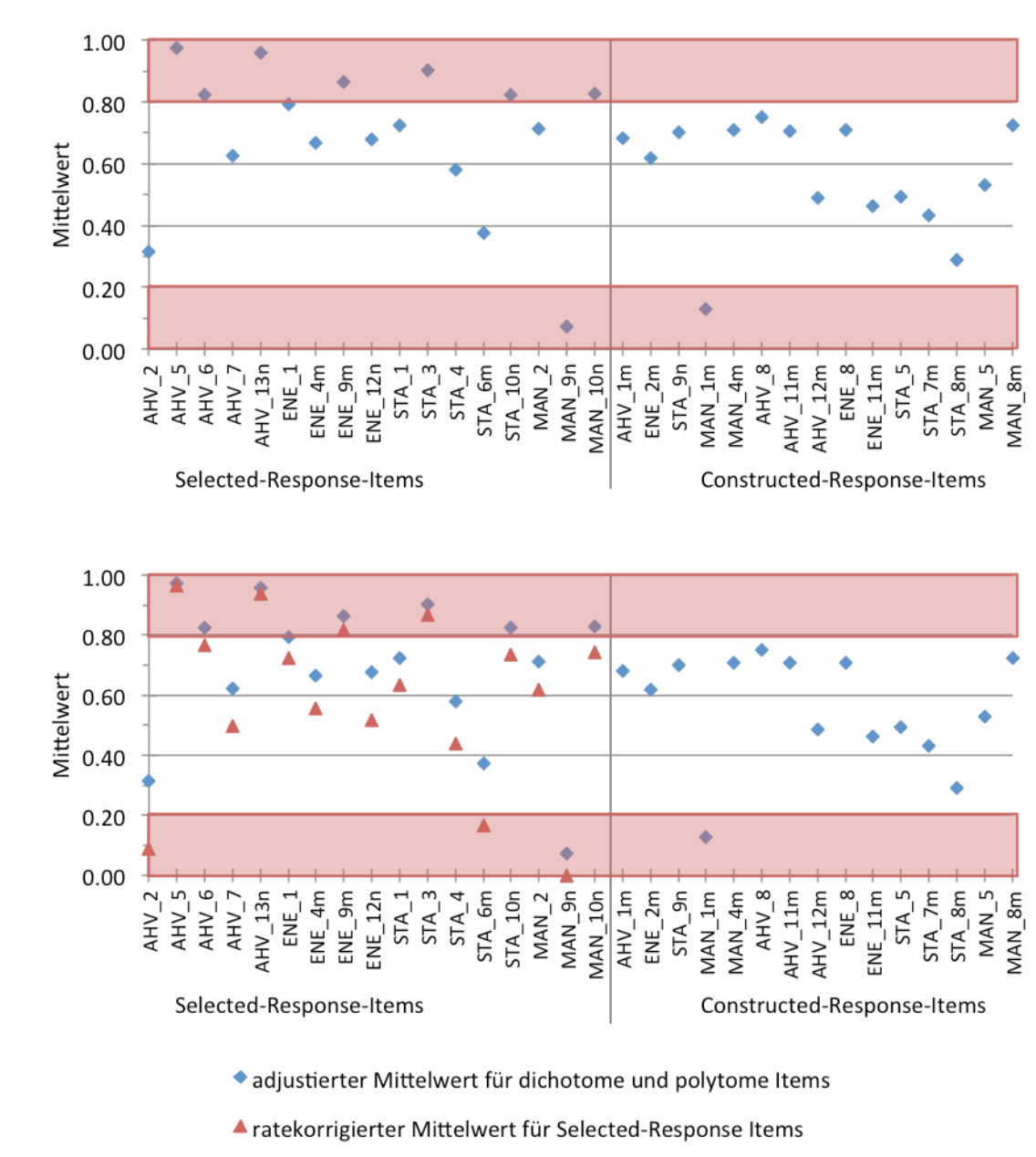


Abbildung 11-17. Klassische Item-Schwierigkeiten (Item-Mittelwerte)

Hinweis: $m = 32$ Items. Adjustierte Mittelwerte (M_{adj}) für dichotome und polytome Items (blau), ratekorrigierte Mittelwerte (M_{cor}) für Selected-Response-Items (rot). $M \approx 1$: einfaches Item, $M \approx 0$: schwieriges Item.

Tabelle 11-13. Items und Antwortkategorien mit zu geringer bzw. zu hoher klassischer Schwierigkeit

Item	Adjustierter Item-Mittelwert: M_{adj} (SD_{adj})	Adjustierter und ratekorrigierter Item-Mittelwert: M_{cor}
AHV_2	0.32 (0.62)	0.09
AHV_5	0.97 (0.22)	0.96
AHV_13n	0.96 (0.22)	0.94
MAN_9n	0.07 (0.35)	0.00

11.5.2.a Item-Schwierigkeit nach Antwortformat

Um die klassischen Item-Schwierigkeiten im WBK-T2 nach Aufgabenformat zu vergleichen, werden die adjustierten Item-Mittelwerte herangezogen (vgl. Abbildung 11-17). Die adjustierte Item-Mittelwerte der Selected-Response-Items sind durchschnittlich einfacher und streuen weiter ($M = 0.69$, $SD = 0.12$) als jene der Constructed-Response-Items ($M = 0.56$, $SD = 0.15$). Der Mittelwertunterschied zwischen den beiden Antwortformaten ist nicht signifikant, aber der Effekt des Formats als Gruppierungsvariable ist gross ($\Delta M = 0.13$, $p = .104$, Cohens $|d| = 0.93$).

11.5.2.b Item-Schwierigkeit nach Itemtypen

Um die klassischen Item-Schwierigkeiten im WBK-T2 nach den sechs Itemtypen zu vergleichen, werden die adjustierten und korrigierten Item-Mittelwerte für jeden Itemtypus zusammengefasst (vgl. Abbildung 11-18). Wie oben erläutert, fallen aufgrund der Ratekorrektur die korrigierten Mittelwerte geringer aus als die unkorrigierten adjustierten. Die Varianz in den adjustierten Item-Mittelwerten zwischen den Itemtypen ist nicht signifikant grösser als die Varianz innerhalb der Typen ($F_{(5;26)} = 1.348$, $p = 0.276$, $\eta^2 = 0.21$).

- Der Typus IV ($M = 0.41$) ist der schwierigste von allen Itemtypen (vgl. Abbildung 11-18). Dieses Ergebnis ist konform mit dem Konstruktionsschema, wonach er auf der Kognitionsstufe k3 einzuordnen ist (vgl. Tabelle 8-4). Erklärungen dafür könnten in der komplexen Darstellungsform (SWOT-Matrix), im komplexen Inhalt (politisch kontroverse und ungelöste Problemstellung) und in der grossen Lösungsoffenheit liegen. Letzteres macht das Lösen der entsprechenden Items für Testpersonen mit fundiertem Vorwissen/Wissen und differenzierter Meinung schwieriger.
- Die Typen I und VI sind ohne Rate-Korrektur am einfachsten von allen Itemtypen ($M = 0.65$ bzw. $M = 0.82$), mit Rate-Korrektur werden sie substantiell schwieriger ($M = 0.56$ bzw. $M = 0.73$). Der Typus I ist mit Rate-Korrektur sogar am zweit schwierigsten von allen Itemtypen. Dieses Ergebnis scheint ein Indiz dafür sein, dass inhaltlich anspruchsvolle Items, die z.B. vertieftes Konzeptwissen erfordern, durch Auswahlantworten bzw. Ratewahrscheinlichkeit einfacher zu lösen sind.
- Der Typus V ($M = 0.57$) befindet sich „im Mittelfeld“, was nicht mit dem Konstruktionsschema konform ist, wo er auf der Kognitionsstufe k1 anzusiedeln ist (vgl. Tabelle 8-4). Eine Erklärung dafür könnte sein, dass das Benennen von Begriffen in einem standardisierten, nicht primär curricular konstruierten und nicht auf unmittelbar vorangehende Lernprozesse ausgerichteten Leistungstest keine eigentliche Informationserinnerung darstellt und somit also anspruchsvoller ist.
- Zwischen den Typen II und III ($M = 0.62$) lassen sich keine Unterschiede im Mittelwert feststellen.

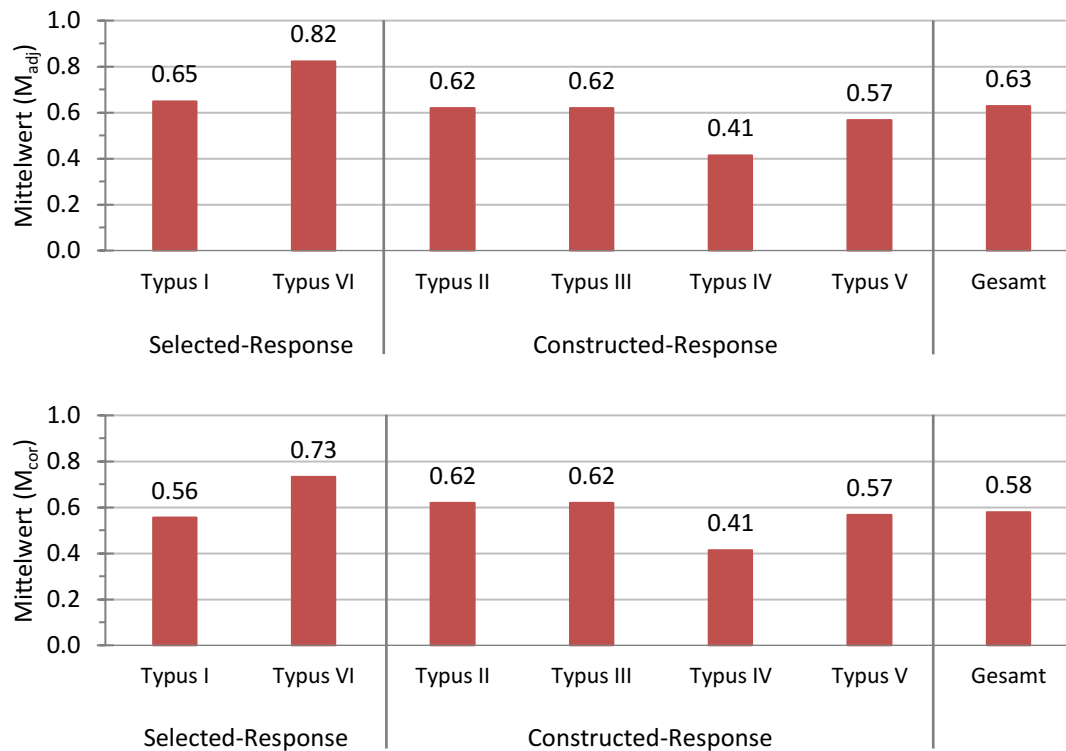


Abbildung 11-18. Klassische Item-Mittelwerte nach Itemtypen

Hinweis: $m = 32$ Items. Adjustierte Item-Mittelwerte (M_{adj}) für dichotome und polytome Items (oben), ratekorrigierte Item-Mittelwerte (M_{cor}) für Selected-Response-Items (unten).

11.5.3 Klassische Item-Trennschärfen

Die Trennschärfe eines Items (*item discriminability*) ermöglicht eine Aussage darüber, wie gut die Differenzierung zwischen den Testpersonen bei einem Item mit derjenigen des gesamten Tests übereinstimmt. Die Item-Trennschärfe (r_{it}) gibt den Zusammenhang zwischen dem Itemwert und dem Testwert an (Kelava & Moosbrugger, 2012, S. 84), die Kategorien-Trennschärfe (r_{ptbis}) den Zusammenhang zwischen der Antwortkategorie und dem Testwert.

Für die Beurteilung der klassischen Item-Trennschärfen im WBK-T2 werden die Item-Skala Korrelationen und die Kategorien-Skala-Korrelationen herangezogen.

11.5.3.a Item-Trennschärfen (Item-Skala-Korrelationen)

Im WBK-T2 variieren die unkorrigierten Item-Skala-Korrelationen im Bereich von .06 und .60; kein Item weist eine negative Trennschärfe auf. Bei allen 15 Constructed-Response-Items und bei 7 Selected-Response-Items sind die Item-Trennschärfen sehr befriedigend ($r_{it-total} \geq .30$) (vgl. Abbildung 11-19, vgl. Anhang Tabelle A- 9 bis Tabelle A- 12). Bei weiteren 6 Selected-Response-Items sind die Item-Trennschärfen knapp befriedigend ($.20 \leq r_{it-total} < .30$) und bei 4 Selected-Response-Items deutlich unbefriedigend ($r_{it-total} < .20$; AHV_13n, STA_3, STA_6m, MAN_9m).

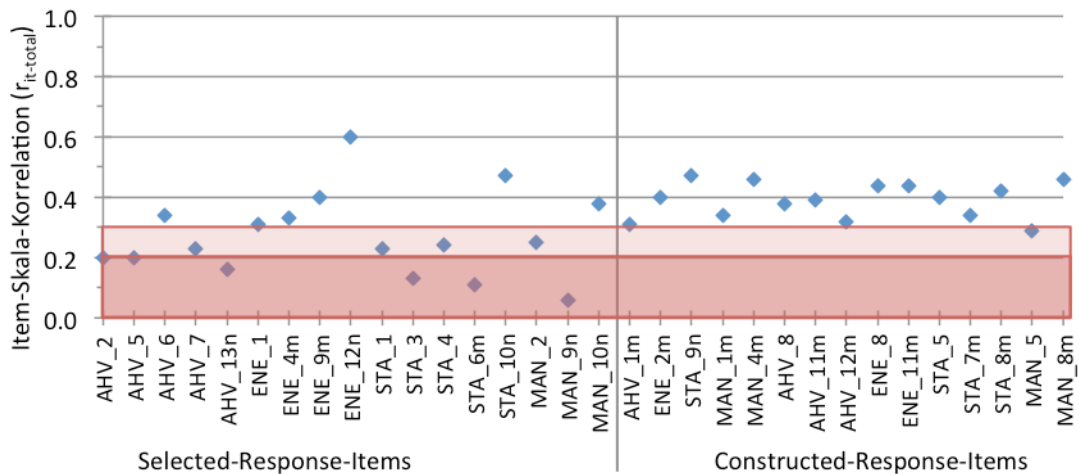


Abbildung 11-19. Klassische Item-Trennschärfen (Item-Skala-Korrelation)

Hinweis: $m = 32$ Items. Item-Skala-Korrelation $r_{it} < .30$: geringe Trennschärfe, $.30 \leq r_{it} \leq .50$: mittlere Trennschärfe, $r_{it} > .50$: hohe Trennschärfe.

Tabelle 11-14. Items und Antwortkategorien mit geringer klassischer Trennschärfe (Typus I)

Item	Item-Trennschärfe unkorrigiert ($r_{it-total}$)	Item-Trennschärfe korrigiert ($r_{it-rest}$)	Kategorien- Trennschärfe richtige Antwortkategorie (r_{ptbis})	Kategorien- Trennschärfe falsche Antwortkategorien (Distraktoren) (r_{ptbis})
AHV_2	.20	.13	.13	-.17 .03 .10
ENE_4	.33	.26	.26	-.08 -.21 .08
STA_4	.24	.17	.17	-.17 -.10 .02
STA_6m	.11	.03	.03	.15 -.23 .00
MAN_2	.25	.18	.18	.03 -.15 -.11
MAN_9n	.06	.02	.02	.14 .01 -.17

Tabelle 11-15. Items und Antwortkategorien mit geringer klassischer Trennschärfe (Typus V und VI)

Item	Kategorien #	Item-Trennschärfe unkorrigiert ($r_{it-total}$)	Item-Trennschärfe korrigiert ($r_{it-rest}$)	Kategorien-Trennschärfe (r_{ptbis})
AHV_13n	5	.16	.07	-.02 .00 -.10 .01 .07
STA_7	3	.34	.23	-.10 -.15 .31
MAN_1m	3	.34	.27	-.27 .23 .14

11.5.3.b Kategorien-Trennschärfen (Kategorie-Skala-Korrelationen)

Im WBK-T2 zeigt sich bei sechs Selected-Response-Items (Typus I) eine positive Kategorien-Trennschärfe für die falsche Kategorie ($r_{ptbis} > .0$), jedoch nur bei zwei Items eine höhere Trennschärfe der falschen Kategorie gegenüber der richtigen Kategorie (STA_6m, MAN_9n) (vgl. Tabelle 11-14). Bei drei polytom geordneten Items sind die Kategorien-Trennschärfen nicht ansteigend angeordnet (AHV_13n, STA_7, MAN_1m) (vgl. Tabelle 11-15). Diese Items (AHV_13n, STA_6m, MAN_9n) sollten bei der nächsten Testrevision modifiziert oder, falls eine Modifikation aus inhaltlichen und formellen Gründen nicht sinnvoll erscheint, aus dem Itempool eliminiert werden.

Zusammenfassend kann festgehalten werden, dass die Testwerte des WBK-T2 annähernd normalverteilt sind. Annahme VII zur Verteilung der Testwerte im WBK-T2 kann also bestätigt werden.

Des Weiteren kann festgehalten werden: Die Items des WBK-T2 decken den mittleren Schwierigkeitsbereich genügend ab und erscheinen für die vorliegende Stichprobe WBKgym als angemessen. Die Items des WBK-T2 unterscheiden genügend gut zwischen Testpersonen mit geringen und hohen Testwerten. Annahme VIII zur Qualität der Items des WBK-T2 kann also als erfüllt betrachtet werden.

11.6 Graduierung des Tests

Welche Evidenz liefert der WBK-T2 für die Graduierung der Item-Schwierigkeiten und für eine kriteriumsorientierte Interpretation der Personenfähigkeiten in einem Kompetenzniveaumodell? Für den WBK-T2 lassen sich anhand von schwierigkeitsbestimmenden Merkmalen und den probabilistischen Item-Schwierigkeiten Kompetenzschwellen quantitativ festlegen und Kompetenzniveaus qualitativ beschreiben (vgl. Annahme IX).

Die Graduierung des Tests zeigt sich dadurch, dass sich die Logit-Skala in Abschnitte unterteilen und jeder Skalenabschnitt anhand der in ihm lokalisierten Items bzw. Item-Schwierigkeiten beschreiben lässt. Analog geht man bei einem Kompetenzniveaumodell⁹⁷ davon aus, dass sich für Anforderungssituationen unterschiedliche Niveaus der Kompetenzerreichung, sog. Kompetenzniveaus hierarchisch abbilden lassen (Hartig & Jude, 2007, S. 34ff.; Hartig & Klieme, 2006, S. 133ff.).

Für die Etablierung eines Kompetenzniveaumodells sind folgende Schritte notwendig (Hartig, 2007, S. 86f., S. 96f.; Hartig & Jude, 2007, S. 34ff.; Hartig & Klieme, 2006, S. 134ff.): (1) die Itemparameter und Personenparameter gemeinsam schätzen und skalieren; (2) die Itemparameter durch schwierigkeitsbestimmende Merkmale vorhersagen (z.B. mittels eines linearen Regressionsmodells oder eines Linear-Logistischen-Testmodells (LLTM))⁹⁸; (3) die Logit-Skala in sog. Kompetenzniveaus unterteilen und jedes Kompetenzniveau anhand seiner charakteristischen Items fachlich beschreiben.

Dem Ergebnis zur Graduierung kommt zweierlei Bedeutung zu. Erstens für die weitere Testentwicklung, die schwierigkeitsbestimmenden Merkmale können als Kriterien bei der Konstruktion neuer Items und bei der Modifikation bestehender Items berücksichtigt werden (Hartig, 2007, S. 89). Zweitens für die Testanwendung, den Personenfähigkeiten (numerische Personenparameter) können spezifische fachliche, d.h. inhalts- und kognitionsbezogene Anforderungen zugeordnet werden (Hartig, 2007, S. 88).

Die Prüfung von Annahme IX zur Graduierung des WBK-T2 erfordert einen *Mixed-Methods-Approach* mit qualitativen und quantitativen Analysen (vgl. Hartig, 2007; Hartig & Jude, 2007, S. 34ff.; Hartig & Klieme, 2006, S. 133ff.). (1) Die fachlichen Anforderungen der Items des WBK-T2 anhand

⁹⁷ In anderen Studien findet sich die synonyme Bezeichnung *Kompetenzstufenmodell* (vgl. OECD, 2013). Nach Hartig (2007, S. 86) ist die Bezeichnung *Kompetenzstufe* als Übersetzung des Englischen *proficiency level* jedoch irreführend. Im pädagogischen und psychologischen Kontext werden die Stufen eines Stufenmodells anhand qualitativer Kriterien voneinander abgegrenzt (wie z.B. im Modell der kognitiven Entwicklung nach Piaget). Die „Stufen“ bzw. Niveaus in Leistungstests sind jedoch bloss eine Verbalisierung und Veranschaulichung von quantitativ erfassten Daten auf einer kontinuierlichen Skala.

⁹⁸ Die beiden Verfahren führen zu inhaltlich sehr ähnlichen Ergebnissen (vgl. z.B. Guggemos & Schönlein, 2015; Hartig, 2004), jedoch ist das LLTM viel aufwändiger in der Datenaufbereitung und Parameterschätzung. Beim LLTM wird simultan mit der Skalierung der Itemparameter auch eine Dekomposition der schwierigkeitsbestimmenden Merkmale in sog. Basisparameter vorgenommen. Jeder Basisparameter repräsentiert ein Merkmal und kann – analog zum Regressionskoeffizienten – zur Definition der Schwellen bzw. Beschreibung der Kompetenzniveaus verwendet werden.

schwierigkeitsbestimmender Merkmale beurteilen. Dafür wird ein Ratingschema für die Items des WBK-T2 entwickelt und externe Experten für die Beurteilung beigezogen. (2) Den regressionsanalytischen Zusammenhang zwischen den fachlichen Anforderungen der Items (Itemrating) und der probabilistischen Item-Schwierigkeiten (Itemparameter) (vgl. Kapitel 11.2) rechnen und Kompetenzschwellen quantitativ bestimmen. Die Regressionsanalyse und mit dem Programm SPSS Statistics Version 25 gerechnet (z.B. Fromm, 2012, S. 83ff.; IBM, 2017). (3) Die Kompetenzniveaus zwischen den Schwellen qualitativ beschreiben und die Personenfähigkeiten kriteriumsorientiert interpretieren.

Der Aufbau dieses Kapitels ist wie folgt: Zuerst wird die Beurteilung der fachlichen Anforderungen anhand schwierigkeitsbestimmender Merkmale vorgestellt (Kapitel 11.6.1). Danach wird der Zusammenhang zwischen Itemparametern und Item-Merkmalen quantitativ bestimmt (vgl. Kapitel 11.6.2) und schliesslich die Kompetenzniveaus qualitativ beschrieben und kriteriumsorientiert interpretiert (vgl. Kapitel 11.6.3).

11.6.1 Qualitative Beurteilung der theoretischen Item-Schwierigkeiten anhand schwierigkeitsbestimmender Merkmale

Schwierigkeitsbestimmende Merkmale haben sowohl bei der Konstruktion von Lern- und Prüfungsaufgaben für den Unterricht (vgl. Dubs, 2004, 2007; H. E. Fischer & Draxler, 2007; Kauertz, Löffler, & Fischer, 2015; Metzger, Waibel, Henning, Hodel, & Luzi, 1993) als auch bei der Konstruktion von Items für psychologische Leistungstests (vgl. Kauertz, 2007; Winther & Achtenhagen, 2009b; Witt, 2006) eine zentrale Bedeutung. Für Leistungstest zur Erfassung von Kompetenzen kommen schwierigkeitsbestimmende Merkmale bei zweierlei Verfahren zur Anwendung: *ex ante* für die fachdidaktisch fundierte, systematische und regelgeleitete Item-Konstruktion (z.B. Hartig, 2007; Schumann & Eberle, 2011); *ex post* für die quantitative Bestimmung und qualitative Beschreibung von Kompetenzniveaus auf einer kontinuierlichen Skala und eine fachdidaktisch sinnvolle Interpretation von Personenfähigkeiten (z.B. Beaton & Allen, 1992; Klieme, 2000). In der Forschungspraxis sind die Grenzen zwischen den beiden Verfahren allerdings fließend. Auf jeden Fall gilt: Je differenzierter die Annahmen über die zu erfassende/erfasste Kompetenz und die Prozesse beim Lösen der spezifischer Anforderungen, desto differenzierter können die Merkmale spezifiziert werden und desto nutzbarer sind die Kompetenzniveaus für die Praxis (Hartig, 2007, S. 98).

11.6.1.a Schwierigkeitsbestimmende Merkmale

Anhand anforderungsrelevanter bzw. schwierigkeitsbestimmender Merkmale kann eine Beurteilung der theoretischen Item-Schwierigkeit vorgenommen werden. Die Merkmale fungieren als Annahmen darüber, welche spezifischen fachlichen Anforderungen beim Lösen eines Items bewältigt werden müssen und insofern zur Schwierigkeit des Items beitragen (Hartig, 2007, S. 88).

Bei der Auswahl der schwierigkeitsbestimmenden Merkmale für die Kompetenzniveaumodellierung gilt es folgendes zu beachten (Schumann & Eberle, 2011, S. 83): (1) Es sollten höchstens fünf Merkmale mit jeweils drei bis vier Ausprägungen verwendet werden; (2) die Merkmale sollten disjunkt sein, d.h. keinen oder einen möglichst kleinen inhaltlichen Überschneidungsbereich haben;⁹⁹ (3) die Merkmale sollten intern graduierbar sein, d.h. in ordinale Ausprägungen i.S.v. Schwierigkeitsstufen unterteilbar sein.

Im Allgemeinen können zwei Hauptkategorien von schwierigkeitsbestimmenden Merkmalen unterschieden werden, formale Merkmale und inhaltsbezogene/kognitionsbezogene Merkmale (vgl. Hartig, 2007; Kauertz, 2007; Schumann & Eberle, 2011). Zu den *formalen Merkmalen* zählen z.B. die grafische Visulisierung des Items, der Textumfang des Itemstamms (Anzahl Wörter), das sprachliche Niveau des Itemstamms (Fremdwörter, Fachbegriffe) und das Antwortformat. Für das Merkmal „Antwortformat“ werden entweder zwei Stufen (gebundenes Format, freies Format) oder drei Stufen (Auswahlantwort, Kurzantwort, Essay-Antwort) vorgeschlagen (Draxler, 2005; H. E. Fischer & Draxler, 2001). Aus theoretischer Perspektive wird an den formalen Merkmalen kritisiert, dass sie keine Rückschlüsse darauf zulassen, *wie* die Testpersonen das Item lösen (Schumann & Eberle, 2011, S. 79). Die empirischen Befunde zum schwierigkeitsbestimmenden Charakter von formalen Merkmalen, insbesondere zum Merkmal „Antwortformat“ sind denn auch nicht eindeutig: Je nach Domäne (Erstsprache/Lesen, Mathematik, Naturwissenschaften, Beruflich/Kaufmännisch) und je nach Testform (Antwortformate mit/ohne Itemstamm-Äquivalenz) ist das Antwortformat schwierigkeitsbestimmend oder nicht (Draxler, 2005; H. E. Fischer & Draxler, 2007; Kauertz, 2007).

Für *inhaltsbezogene Merkmale*, wozu i.d.R. auch kognitionsbezogene gezählt werden, finden sich in der Literatur etliche Ansätze. Das am weitesten verbreitete Merkmal ist „kognitive Prozesse“, das nach der klassischen bzw. revidierten Taxonomie für Lernziele mit sechs Stufen (L. W. Anderson & Krathwohl, 2001; Bloom et al., 1956) oder mit vier Stufen (Marzano & Kendall, 2007) operationalisiert wird (vgl. Kapitel 4.4.1). Für wirtschaftliche Lernziele, Lern- und Prüfungsaufgaben wird eine Taxonomie mit drei Hauptprozessen vorgeschlagen (Metzger et al., 1993): (1) Informationserinnerung: Wiedererkennen, Wiedergeben; (2) Informationsverarbeitung: Sinnerfassen, Anwenden; (3) Informationserzeugung: Analyse, Synthese, Beurteilen. Aus theoretischer Sicht wird jedoch kritisiert, dass die Stufen all dieser Taxonomien nicht disjunkt sind, sondern dass die obere Stufe die jeweils unteren Stufen einschliesst. Entsprechend sind auch die empirischen Befunde zu diesen kognitiven Taxonomien nicht eindeutig (vgl. Guggemos & Schönlein, 2015; Winther, 2010; Winther & Achtenhagen, 2009b; Witt, 2006).

⁹⁹ Haben die Merkmale grössere Überschneidungsbereiche, so verletzt dies in der Regressionsanalyse die Voraussetzung der Multikollinearität.

Ebenfalls häufig wird das Merkmal „Offenheit der Lösungswege“ verwendet, aber je nach Domäne anders definiert. Für naturwissenschaftliche Lernaufgaben wird folgende Operationalisierung vorgeschlagen (Draxler, 2005; H. E. Fischer & Draxler, 2001): (1) Die Aufgabe lässt mehrere Lösungswege zu, aber gibt weder direkt noch indirekt einen bestimmten Lösungsweg vor; (2) die Aufgabe lässt mehrere Lösungsmöglichkeiten zu und schlägt einige alternative Lösungswege vor; (3) die Aufgabe schreibt einen bestimmten Lösungsweg vor. Für wirtschaftliche Lernaufgaben wird eine umgekehrte Operationalisierung vorgeschlagen (Dubs, 2004, 2007): (1) der Lösungsweg wird durch die Aufgabe vorgegeben und nur eine Lösung ist möglich; (2) mehrere Lösungswege sind durch Aufgabe zugelassen, aber nur eine konsistente Lösung ist möglich; (3) mehrere Lösungswege sind durch Aufgabe zugelassen, aber mehrere konsistente und begründbare Lösungen sind möglich.

Das Merkmal „inhaltliche Komplexität“ bezieht sich auf die Vernetzung von Inhaltselementen und Zusammenhängen. Für naturwissenschaftliche Lernaufgaben wird folgende Stufung vorgeschlagen (H. E. Fischer, Glemitz, Kauertz, & Sumfleth, 2007): (1) ein einzelnes Inhaltselement vorhanden, kein Zusammenhang, (2) mindestens zwei verbundene Inhaltselemente vorhanden, ein Zusammenhang, (3) mehrere untereinander verbundene Zusammenhänge vorhanden. Für wirtschaftliche Lernaufgaben wird eine andere Stufung vorgeschlagen (Metzger et al., 1993): (1) < 7 Informationseinheiten auszählen, (2) 8-14 Informationseinheiten auszählen; (3) > 15 Informationseinheiten auszählen. Aus theoretischer Sicht ist hier zu kritisieren, dass der Umfang (Anfang und Ende) eines Inhaltselements bzw. einer Informationseinheit schwierig abzustecken ist.

11.6.1.b Auswahl von Merkmalen und Ausprägungen

Für den WBK-T2 wurden zunächst zwei schwierigkeitsbestimmende Merkmale *ex ante* für die Itemkonstruktion festgelegt (vgl. Kapitel 8.3.3):

- Das Merkmal „Antwortformat“ wurde entsprechend dem Item-Konstruktionsschema definiert (i.A.a. Draxler, 2005; H. E. Fischer & Draxler, 2001) (vgl. Tabelle 8-4): (f1) *Selected-Response*, (f2) *Short-Constructed-Response* und (f3) *Extended-Constructed-Response*.
- Das Merkmal „Kognitionsprozess“ wurde ebenfalls entsprechend dem Item-Konstruktionsschema definiert (i.A.a. Marzano & Kendall, 2007; Metzger et al., 1993) (vgl. Tabelle 8-4): (k1) wiedererkennen/wiedergeben, (k2) verstehen/anwenden und (k3) vergleichen/beurteilen/entscheiden.

Später wurden vier weitere schwierigkeitsbestimmende Merkmale mit jeweils drei Ausprägungen für eine *ex post* Itemanalyse ausgewählt:

- Das Merkmal „Lösungsoffenheit“ bezieht sich auf die Offenheit der zugelassenen Lösung (i.A.a. Dubs, 2004, 2007): (1) genau eine bestimmte Lösung zugelassen, (2) mehrere bestimmte Lösungen zugelassen, und (3) mehrere unbestimmte Lösungen zugelassen.
- Das Merkmal „Wissensart“ bezieht sich auf die Art des kognitiven Wissens, das für die Lösung beigezogen werden muss (i.A.a. L. W. Anderson & Krathwohl, 2001): (1) *factual knowledge*, (2) *conceptual knowledge* und (3) *procedural knowledge*.
- Das Merkmal „inhaltliche Komplexität“ bezieht sich auf die vertikale Vernetzung der Inhalte (i.A.a. H. E. Fischer et al., 2007): (1) keine fachlichen Konzepte verbinden, (2) 2 fachliche Konzepte verbinden, (3) > 2 fachliche Konzepte verbinden.
- Das Merkmal „Informationsvorgabe“ wurde spezifisch für den WBK-T2 definiert, bei dem sich sowohl im Einleitungstext als auch im Itemstamm (vgl. Kapitel 8.3) direkte oder indirekte Hinweise für die Lösung eines Items finden: (1) Information vollständig vorgegeben, (2) Information teilweise vorgegeben, (3) Information nicht vorgegeben.

11.6.1.c Itemrating und Inter-Rater-Reliabilität

Nach der Benennung und Beschreibung der schwierigkeitsbestimmenden Merkmale muss durch unabhängige und geschulte Fachexperten eine Beurteilung jedes Items im Test hinsichtlich dieser Merkmale vorgenommen werden. Die Beurteilung kann entweder quantitativ (Liegt das Merkmal beim Item vor? Ja/nein) oder qualitativ (In welcher Ausprägung liegt das Merkmal beim Item vor? Stufe 1 bis x) erfolgen (Hartig, 2007, S. 89).

Die Beurteilung der theoretischen Schwierigkeit der 32 Items der WBK-T2 erfolgte durch fünf Expert*innen: drei Expert*innen waren als Fachdidaktiker*innen WuR an universitären bzw. pädagogischen Hochschulen tätig, zwei als Forscher*innen im Bereich der Wirtschaftspädagogik an universitären Hochschulen. Den Experten wurde ein Ratingschema vorlegt, das aus vier schwierigkeitsbestimmenden Merkmalen mit jeweils drei Ausprägungen besteht (vgl. Tabelle 11-16). Die zwei Merkmale „Antwortformat“ und „Kognitionsprozess“ wurden von den Experten nicht beurteilt, da diese Merkmale und ihre Ausprägungen durch das Konstruktionsschema definiert sind und bereits bei der Testentwicklung evaluiert wurden (vgl. Tabelle 8-4).

Tabelle 11-16. Ratingschema für Beurteilung der theoretischen Schwierigkeit der Items im WBK-T2

Stufe	Lösungsoffenheit	Wissensart	Inhaltliche Komplexität	Informationsvorgabe
3	mehrere unbestimmte Lösungen zugelassen (nicht abschliessend)	<i>procedural knowledge</i> : Handlungswissen, Problemstellungen bearbeiten	mehr als 2 fachliche Konzepte verbinden (hohe Komplexität)	Information nicht in Einleitungstext oder Itemstamm
2	mehrere bestimmte Lösungen zugelassen (abschliessend)	<i>conceptual knowledge</i> : Konzeptwissen, Konzepte verstehen und verbinden	nur 2 fachliche Konzepte verbinden (moderate Komplexität)	Information teilweise in Einleitungstext oder Itemstamm
1	genau eine bestimmte Lösung zugelassen	<i>factual knowledge</i> : Faktenwissen, Fakten kennen	keine fachlichen Konzepte verbinden (geringe Komplexität)	Information vollständig in Einleitungstext oder Itemstamm

Hinweis: 1 = geringster Wert, 2 = mittlerer Wert, 3 = höchster Wert.

Tabelle 11-17. Inter-Rater-Agreement und Inter-Rater-Reliabilität nach schwierigkeitsbestimmenden Merkmalen

Merkmal	k	M	SD	Inter-Rater-Agreement (Holsti-Formel)	Inter-Rater-Reliabilität (Krippendorffs α)
Antwortformat	32	1.78	0.91	---	---
Kognitionsprozess	32	1.53	0.67	---	---
Lösungsoffenheit	32	1.44	0.50	0.96	.86
Wissensart	32	1.56	0.67	0.88	.83
Inhaltliche Komplexität	32	1.81	0.78	0.81	.76
Informationsvorgabe	32	2.28	0.77	0.83	.71
Mittelwert	32	---	---	0.87	.79

Hinweis: $N = 5$ Rater, $m = 32$ Items. Mittelwert (M) und Standardabweichung (SD) von Rater e0. Inter-Rater-Übereinstimmung nach Holsti-Formel und Inter-Rater-Reliabilität nach Krippendorffs α für jedes Merkmal zwischen fünf Ratern.

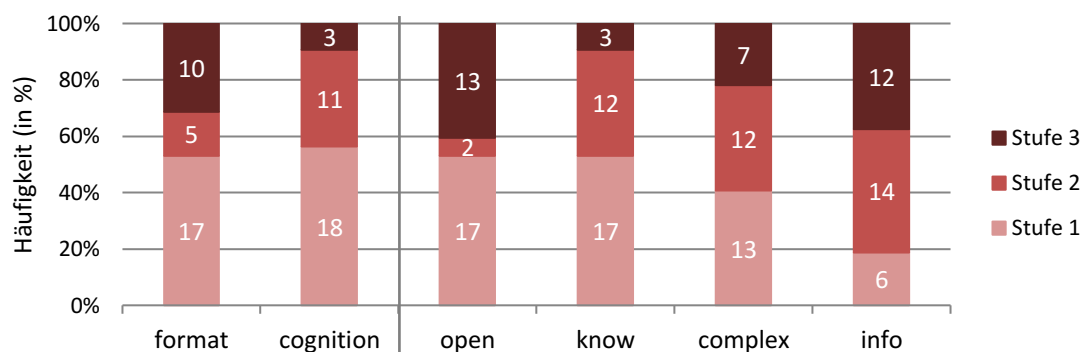


Abbildung 11-20. Finales Itemrating: Item-Anzahl nach Merkmal und Ausprägung

Hinweis: $m = 32$ Items. Merkmale „format“ und „cognition“ nach Konstruktionsschema. Alle anderen Merkmale gemäss Itemrating der Autorin (Experte e0). Inter-Rater-Reliabilität der 5 Rater nach Krippendorffs $\alpha = .79$.

Die Inter-Rater-Reliabilität zwischen den fünf Experten wurde für jedes Merkmal mit Krippendorffs α berechnet (Hayes & Krippendorff, 2007; Krippendorff, 1970).¹⁰⁰ Die durchschnittliche Inter-Rater-Übereinstimmung (Holsti-Formel = .87) und die durchschnittliche Inter-Rater-Reliabilität (Krippendorffs α = .79) war zufriedenstellend (vgl. Tabelle 11-17, Anhang Tabelle A- 14). Die Merkmale „Lösungsoffenheit“ und „Wissensart“ zeigten gute Werte ($\alpha > .83$), die Merkmale „Informationsvorgabe“ und „inhaltliche Komplexität“ zufriedenstellende Werte ($\alpha > .71$). Das Merkmal „inhaltliche Komplexität“ erwies sich im Itemrating als anspruchsvoll und „übereinstimmungsfeindlich“, weil der Umfang eines fachlichen Konzepts *per se* unklar ist (z.B. Konjunkturpolitik vs. fiskalische Konjunkturpolitik).

Das finale Itemrating für die Merkmale „Antwortformat“ und „Kognitionsprozess“ erfolgte gemäss dem Konstruktionsschema (Itemtypen, vgl. Tabelle 8-4), für die anderen vier Merkmale gemäss dem Itemrating der Autorin (Experte e0). Die Item-Anzahl auf jeder Ausprägung verteilt sich für jedes Merkmal unterschiedlich (vgl. Abbildung 11-20). Bei den Merkmalen „Antwortformat“ und „Lösungsoffenheit“ finden sich am meisten Items auf Stufe 1 und am zweitmeisten auf Stufe 3. Bei den Merkmalen „Kognitionsprozess“ und „Wissensart“ sind die meisten Items auf Stufe 1 und Stufe 2, beim Merkmal „Informationsvorgabe“ sind die meisten Items auf Stufe 2 und Stufe 3.

11.6.2 Quantitative Bestimmung von Kompetenzschwellen

Um auf der Logit-Skala eines Tests Kompetenzniveaus zu bilden, stehen verschiedene Methoden zur Verfügung. In jedem Fall aber müssen Schwellen zwischen den Niveaus definiert werden. Die inhaltliche Differenzierung der Personenfähigkeiten passiert nämlich an den Schwellen und nicht innerhalb eines Niveaus (Hartig, 2007, S. 86f.).

Die kontinuierliche Logit-Skala der Personen- und Itemparameter kann willkürlich oder regelgeleitet in einzelne Logit-Intervalle, sog. Kompetenzniveaus unterteilt werden. Mit den schwierigkeitsbestimmenden Merkmalen lässt sich eine qualitativ und quantitativ begründete Unterteilung erreichen. Jedes Kompetenzniveau wird durch eine untere und obere Schwelle auf der Logit-Skala definiert. Diese Schwellen werden quantitativ bestimmt, und zwar durch die Regression der Itemparameter auf die schwierigkeitsbestimmenden Merkmale. Die Kompetenzniveaus werden qualitativ beschrieben anhand der in der Stufe lokalisierten Items und ihren inhalts- und kognitionsbezogenen Anforderungen.

Für die inferenzstatistischen Analysen des Zusammenhangs zwischen Itemparameter und schwierigkeitsbestimmenden Merkmalen werden die Ausprägungen der Merkmale numerisch

¹⁰⁰ Krippendorffs α ist ein Mass für die Inter-Rater-Reliabilität von mehr als zwei Ratern für Variablen aller Skalenniveaus (Hayes & Krippendorff, 2007; Krippendorff, 1970). Der Wertebereich von Krippendorffs α liegt zwischen 0 und 1. Nach konventionellen Normen wird ein Krippendorffs $\alpha > .67$ als "satisfactory" und $\alpha > .80$ als "good" beurteilt (De Swert, 2012).

kodiert. Die Kodierung muss derart sein, dass die Merkmale einen positiven Zusammenhang mit der Item-Schwierigkeit aufweisen, d.h. je höher der Wert desto höher die Schwierigkeit des Items (Hartig, 2007, S. 90).

11.6.2.a Bedeutung der einzelnen Merkmale (Korrelationsanalysen)

Für die Bedeutung der einzelnen schwierigkeitsbestimmenden Merkmale hinsichtlich der probabilistischen Item-Schwierigkeiten (Itemparameter) im WBK-T2 wurden bivariate Korrelationsanalysen gerechnet (z.B. Kuckartz, Rädiker, Ebert, & Schehl, 2013, S. 207ff.) (vgl. Tabelle 11-18). Die Korrelation zwischen den Itemparametern und den schwierigkeitsbestimmenden Merkmalen ist moderat positiv und signifikant für die Merkmale „Informationsvorgabe“ ($r = .41$, $p = .019$) „Antwortformat“ ($r = .37$, $p = .036$) und „Lösungsoffenheit“ ($r = .36$, $p = .046$). Sie ist gering positiv und nicht signifikant für die Merkmale „Wissensart“ ($r = .28$, $p = .123$), „inhaltliche Komplexität“ ($r = .27$, $p = .139$) und „Kognitionsprozess“ ($r = .20$, $p = .273$).

Die Merkmale „Antwortformat“ und „Lösungsoffenheit“ sind untereinander hoch korreliert ($r = .97$, $p < .001$). Dieses Ergebnis ist bei näherer Betrachtung des Ratingschemas, insbesondere der Beschreibung der Merkmalsausprägungen nicht überraschend: Ein gebundenes Antwortformat verlangt nach einer bestimmten Lösung, ein freies Antwortformat nach mehreren (bestimmten oder unbestimmten) Lösungen. Ebenso sind die Merkmale „Wissensart“ und „Kognitionsprozess“ ($r = .86$, $p < .001$) untereinander hoch korreliert. Dieses Ergebnis ist in Anbetracht des Ratingschemas erwartungskonform: Fakten- und Konzeptwissen bedingt Informationserinnerung (Wiedererkennen/Wiedergeben), Informationsverarbeitung (Verstehen/Anwenden), Handlungswissen bedingt Informationserzeugung (Vergleichen/Beurteilen/Entscheiden). Zu einem ähnlichen Befund kommen auch vergleichbare Studien im Bereich der Wirtschaftspädagogik zu schwierigkeitsbestimmenden Merkmalen für Leistungstests (vgl. Förster, Brückner, Happ, Beck, & Zlatkin-Troitschanskaia, 2017).

Für das Ratingschema zu den schwierigkeitsbestimmenden Merkmalen (vgl. Kapitel 11.6.1.b) legen die obigen Ergebnisse nahe, in Zukunft entweder „Antwortformat“ oder „Lösungsoffenheit“ einzubeziehen sowie entweder „Kognitionsprozess“ oder „Wissensart“. Bezüglich des Konstruktionsschemas für Items (vgl. Tabelle 8-4) legen die obigen Ergebnisse nahe, das schwierigkeitsbestimmende Merkmal „Antwortformat“ darin zu belassen und das Merkmal „Informationsvorgabe“ darin aufzunehmen.

Tabelle 11-18. Korrelation zwischen schwierigkeitsbestimmenden Merkmalen und probabilistischen Item-Schwierigkeiten (Itemparameter)

	Item- parameter	format	cognition	open	know	complex	info
Probabilistische Item-Schwierigkeit (Itemparameter)	1	.37* (.036)	.20 (.273)	.36* (.046)	.28 (.123)	.27 (.139)	.41* (.019)
Antwortformat (format)		1	.64*** ($< .001$)	.97*** ($< .001$)	.59*** ($< .001$)	.32 (.074)	.29 (.198)
Kognitionsprozess (cognition)			1	.54** (.002)	.86*** ($< .001$)	.79*** ($< .001$)	.06 (.760)
Lösungsoffenheit (open)				1	.49** (.005)	.22 (.231)	.33 (.067)
Wissensart (know)					1	.80*** ($< .001$)	.18 (.318)
Inhaltliche Komplexität (complex)						1	.22 (.234)
Informationsvorgabe (info)							1

Hinweis: $m = 32$ Items. Nicht-parametrische Korrelationsanalyse. Bivariate Korrelationskoeffizienten als Spearmans ρ . Signifikanzniveaus (2-seitig): *** $p < .001$, ** $p < .01$, * $p < .05$.

11.6.2.b Bedeutung der Merkmalsausprägungen (Varianzanalysen)

Für die Bedeutung der Merkmalsausprägungen hinsichtlich der Itemparameter im WBK-T2 wurden Varianzanalysen gerechnet (vgl. Abbildung 11-21, Tabelle 11-19) (z.B. Fromm, 2012, S. 12ff.; Kuckartz et al., 2013, S. 185ff.). Bereits anhand der deskriptiven Ergebnisse ist zu erkennen, dass sich die Graduierung der Merkmalsausprägungen (Stufe 3 > Stufe 2 > Stufe 1) auf dem Itemparameter für (fast) alle Merkmale abbilden lässt, aber dass sich die Ausprägungen nicht bei allen Merkmalen gleich deutlich unterscheiden.

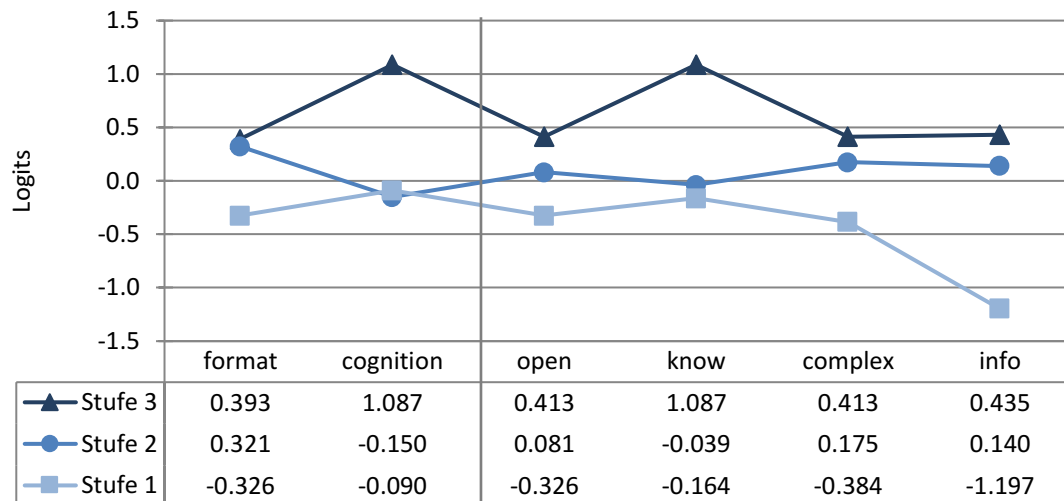


Abbildung 11-21. Deskriptive Analyse der Itemparameter nach Merkmal und Ausprägung

Hinweis: $m = 32$ Items. Mittelwerte der Merkmalsausprägungen in Logits. Merkmale „format“ und „cognition“ gemäss Konstruktionsschema. Alle anderen Merkmale gemäss Itemrating der Autorin (Experte e0).
 format = Antwortformat, cognition = Kognitionsprozess, open = Lösungsoffenheit, know = Wissensart, complex = inhaltliche Komplexität, info = Informationsvorgabe.

Tabelle 11-19. Varianzanalyse für Itemparameter nach schwierigkeitsbestimmenden Merkmalen

Merkmal	k	F	df1;df2	p	η^2
Antwortformat (format)	32	1.35	2;29	.274	0.09
Kognitionsprozess (cognition)	32	1.38	2;29	.222	0.09
Lösungsoffenheit (open)	32	1.42	2;29	.258	0.09
Wissensart (know)	32	1.41	2;29	.260	0.09
Inhaltliche Komplexität (complex)	32	1.21	2;29	.313	0.08
Informationsvorgabe (info)	32	4.73	2;29	.017	0.25

Hinweis: Alle Merkmale mit 3-stufiger Ausprägung.

Die deutlichste Abstufung und signifikante Varianzunterschiede zeigen sich beim Merkmal „Informationsvorgabe“ ($F_{(2;29)} = 4.73$, $p = .017$, $\eta^2 = 0.25$). Allerdings zeigt der *Post-hoc*-Test GT2, dass die Varianzunterschiede nur zwischen Stufe 1 und 2 ($\Delta M_{s2-s1} = -1.337$, $p = .050$) sowie zwischen Stufe 1 und 3 ($\Delta M_{s3-s1} = 1.632$, $p = .016$) signifikant sind. Ebenfalls deutliche Abstufungen, aber nicht signifikante Varianzunterschiede haben die Merkmale „Lösungsoffenheit“ ($\Delta M_{s3-s1} = 0.739$, $p > .258$, $\eta^2 = 0.09$) und „inhaltliche Komplexität“ ($\Delta M_{s3-s1} = 0.798$, $p > .313$, $\eta^2 = 0.08$). Beim Merkmal „Antwortformat“ sind die Varianzunterschiede nicht signifikant ($F_{(2;29)} = 1.35$, $p = .274$, $\eta^2 = 0.09$), insbesondere unterscheiden sich Stufe 2 (*Short-Constructed-Response*) und 3 (*Extended-Constructed-Response*) kaum ($\Delta M_{s3-s2} = 0.072$). Bei den Merkmalen „Wissensart“ ($F_{(2;29)} = 1.41$, $p = .260$, $\eta^2 = 0.09$) und „Kognitionsprozess“ ($F_{(2;29)} = 1.38$, $p = .222$, $\eta^2 = 0.09$) sind die Varianzunterschiede ebenfalls nicht signifikant; bei beiden Merkmalen ist Stufe 1 und 2 kaum unterscheidbar.

Für das Itemrating zu den schwierigkeitsbestimmenden Merkmalen (vgl. Kapitel 11.6.1.b) sprechen die obigen Ergebnisse bei den Merkmalen „Lösungsoffenheit“, „Inhaltliche Komplexität“ und „Informationsvorgabe“ für eine Beibehaltung der drei Stufen. Beim Merkmal „Antwortformat“ ist für weitere Analysen eine Zusammenlegung der Stufen naheliegend: (f1) Selected-Response und (f2/f3) Constructed-Response. Ebenfalls sollte bei den Merkmalen „Kognitionsprozess“ und „Wissensart“ für weitere Analysen eine Zusammenlegung der Stufen bedacht werden: (k1/k2) wiedergeben & verstehen/anwenden und (k3) vergleichen/beurteilen/entscheiden bzw. (12) Faktenwissen & Konzeptwissen und (3) Handlungswissen.

11.6.2.c Prädiktion der Itemparameter durch schwierigkeitsbestimmende Merkmale (Regressionsanalyse)

Die Bedeutung der schwierigkeitsbestimmenden Merkmale bzw. Merkmalsausprägungen für die Itemparameter muss regressionsanalytisch untersucht werden. Für die Prädiktion der probabilistischen Item-Schwierigkeiten (Itemparameter) durch die schwierigkeitsbestimmenden Merkmale wird eine multiple Regressionsanalyse gerechnet (z.B. Fromm, 2012, S. 83ff.; Kuckartz et al., 2013, S. 259ff.). Die Merkmale werden in sog. Dummy-Variablen (d) rekodiert, die nur Werte von 0 und 1 annehmen. Bei einem dichotomen Merkmal wird die einfachere Ausprägung mit 0 kodiert, die schwierigere mit 1. Bei einem polytomen Merkmal werden k-1 Dummy-Variablen für die Ausprägungen gebildet, 0 steht für das Nicht-Vorhandensein der Ausprägung, 1 für das Vorhandensein.¹⁰¹

Im linearen Regressionsmodell wird die probabilistische Schwierigkeit des Items (σ_i) als (abhängige) Zielvariable definiert, die Merkmale bzw. die Dummy-kodierten Merkmalsausprägungen des Items (x_{di}) als (unabhängige) Einflussvariable. Mit anderen Worten, die probabilistische Item-Schwierigkeit wird als gewichtete Summe seiner einzelnen Merkmale bzw. Merkmalsausprägungen modelliert (Hartig, 2007, S. 91):

$$\sigma_i = b_0 + b_1 * x_{1i} + b_2 * x_{2i} + \dots + b_d * x_{di} + \varepsilon_i, \quad i = 1, \dots, m,$$

wobei σ den Itemparameter, b das unstandardisierte Regressionsgewicht, x Wert des Merkmals bzw. Merkmalsausprägung (als Dummy-Variable: 0 = nicht vorhanden, 1 = vorhanden) und ε das Residuum bezeichnet.

¹⁰¹ Beispiel: Für ein Merkmal mit drei Ausprägungen werden zwei Dummy-Variablen gebildet. Die Ausprägung 1 bildet die Referenzkategorie, Ausprägung 2 und Ausprägung 3 sind je eine Dummy-Variable mit den Werten 0 und 1.

Die Modellgüte, d.h. inwieweit die Varianz in den Itemparametern durch die Varianz in den Merkmalen erklärt wird, kann am Determinationskoeffizienten (R^2) abgelesen werden: je höher R^2 , desto besser erklären die ausgewählten Merkmale die Itemparameter; bei einem R^2 nahe bei 0 leisten die ausgewählten Merkmale keine oder eine nur unbedeutende Erklärungskraft. Der standardisierte Regressionskoeffizient (β) kann als Effekt des jeweiligen Merkmals auf die Itemparameter interpretiert werden: Merkmale mit einem hohen Regressionskoeffizienten ($\beta \geq 0.5$) sind besonders bedeutsam für den Itemparameter, Merkmale mit einem sehr geringen Regressionskoeffizienten ($\beta \approx 0.0$) sind nicht bedeutsam. Der unstandardisierte Regressionskoeffizient (b) ist wie folgt zu interpretieren: Ein Item mit der Merkmalsausprägung $x_{di} = 1$ ist um b_d -Logits schwieriger als ein Item ohne diese Merkmalsausprägung $x_{di} = 0$ (Hartig, 2007, S. 90).

Für den WBK-T2 wurden vier der sechs schwierigkeitsbestimmenden Merkmale für die Regressionsanalyse verwendet. Um Multikollinearität unter den Prädiktoren zu vermeiden, wurden die zwei Merkmale aus dem Konstruktionsschema nicht berücksichtigt („Antwortformat“ ist hoch korreliert mit „Lösungsoffenheit“, „Kognitionsprozess“ mit „Wissensart“) (vgl. Tabelle 11-18). Aus den vier Merkmalen wurden acht Dummy-Variablen gebildet¹⁰² und simultan als Prädiktoren in die Regressionsanalyse eingeführt.

Das Regressionsmodell zeigt eine Varianzaufklärung von 22 % ($R^2 = .42$, korr. $R^2 = .22$) und ist nicht signifikant ($F_{(8;23)} = 2.10$, $p = .078$) (vgl. Tabelle 11-20). Vom Merkmal „Informationsvorgabe“ hat der Prädiktor info3 einen starken und signifikanten Effekt auf den Itemparameter ($\beta = .70$, $p = .013$) und info2 einen moderaten und knapp signifikanten Effekt ($\beta = .51$, $p = .054$). Beim Merkmal „inhaltliche Komplexität“ haben beide Prädiktoren complex2 und complex3 zwar moderate Effekte, aber sind deutlich nicht signifikant; beim Merkmal „Lösungsoffenheit“ haben die Prädiktoren open2 und open3 geringe Effekte und sind ebenfalls deutlich nicht signifikant. Beim Merkmal „Wissensart“ sind die Prädiktoren know2 und know3 sogar negativ und nicht signifikant.

Das Ergebnis der Regressionsanalyse ist insofern überraschend, als dass die Ausprägungen der drei theoretisch fundierten Merkmale (vgl. Kapitel 11.6.1.b) eine nicht signifikante Erklärungskraft für die Itemparameter im WBK-T2 haben, und dass sie sogar gering („Lösungsoffenheit“) bzw. negativ („Wissensart“) ist. Entscheidend für die Erklärungskraft der Itemparameter im WBK-T2 ist die Informationsvorgabe, d.h. ob und wie viel Information aus dem Gedächtnis abgerufen werden muss. Es drängt sich hier die Frage auf, ob die ausgewählten schwierigkeitsbestimmenden Merkmale für die Items des WBK-T2 tatsächlich geeignet sind oder ob sich mittels einer vertieften Itemanalyse treffendere Merkmale bzw. Ausprägungen finden lassen. Diese Frage sollte bei einer allfälligen Re-Analyse berücksichtigt werden.

¹⁰² Bei vier Merkmalen mit je drei Ausprägungen sind $4 \cdot (3-1) = 8$ Dummy-Variablen erforderlich.

Tabelle 11-20. Regressionsanalyse der Itemparameter auf Ausprägungen der schwierigkeitsbestimmenden Merkmale

Variable	M	Korrelation		Regression			
		r	p	b	SE b	β	p
Konstante	---	---	---	-1.511	0.476	---	.004
open2	32	0.12	.526	0.985	0.837	0.20	.251
open3	32	0.22	.220	0.320	0.532	0.13	.553
know2	32	0.11	.535	-1.250	0.755	-0.51	.111
know3	32	0.18	.314	-0.109	1.156	-0.03	.926
complex2	32	0.10	.571	0.900	0.651	0.37	.180
complex3	32	0.28	.116	1.308	0.890	0.46	.155
info2	32	0.02	.924	1.214	0.599	0.51°	.054
info3	32	0.29	.111	1.716	0.635	0.70*	.013
Varianzaufklärung				$R^2 = .42$, korr. $R^2 = .22$			
Signifikanz				$F_{(8;23)} = 2.073$, $p = .082$			

Hinweis: $m = 32$ Items. Signifikanzniveaus: ** $p < .01$, * $p < .05$, ° $p < .10$.

Parametrische Korrelationsanalyse. Bivariate Korrelationskoeffizienten als Pearsons r .

Multiple simultane Regressionsanalyse. Abhängige Variable: Itemparameter. Unabhängige Dummy-Variablen aus den ursprünglich ordinal-skalierten Variablen: Informationsvorgabe (info), Antwortformat (format), Wissensart (know), inhaltliche Komplexität (complex). Stufe 1 = Referenzkategorie (Konstante), Stufe 2 = z.B. info2, Stufe 3 = z.B. info3. Multikollinearität unproblematisch (alle TOL > 0.25).

Tabelle 11-21. Bedeutsame schwierigkeitsbestimmende Merkmale für die Items des WBK-T2

Merkmal und Merkmalsausprägung	b	SE b	β	p
Regressionskonstante	-1.511	0.476	---	.004
Komplexität: 2 Konzepte verbinden (complex2)	0.900	0.651	.37	.180
Informationsvorgabe: Information teilweise vorgegeben (info2)	1.214	0.599	.51	.054
Komplexität: > 2 Konzepte verbinden (complex3)	1.308	0.890	.46	.155
Informationsvorgabe: Information nicht vorgegeben (info3)	1.716	0.635	.70	.013

Es zeigt sich auch in dieser Arbeit, dass die Taxonomien für kognitive Prozesse, die ursprünglich für die Formulierung von Lernzielen und Lernaufgaben konzipiert wurden (vgl. Kapitel 4.4.1 und 11.6.1.a), sich nicht als kognitionsbezogene schwierigkeitsbestimmende Merkmale von Testitems eignen. Mit anderen Worten, aus dem Merkmal „Kognitionsprozess“ kann nicht darauf geschlossen werden, wie die Testpersonen die Items lösen. Auch in vergleichbaren Studien aus dem Bereich der Wirtschaftspädagogik, die schwierigkeitsbestimmende Merkmale für Leistungstests zur Erfassung beruflich-kaufmännischer Kompetenzen untersuchten, zeigte sich ein ähnliches Bild: Schwierigkeitsbestimmende Merkmale zum kognitiven Prozess liessen sich nicht für die

verstehensbasierte Subskalen modellieren¹⁰³, eher für die handlungsbasierten Subskalen (vgl. Guggemos & Schönlein, 2015; Winther & Achtenhagen, 2009b).

11.6.2.d Bestimmung der Schwellen auf der Logit-Skala

Die *erwartete Schwierigkeit* eines Items ($\hat{\delta}_i$) kann anhand seiner spezifischen Merkmalskombinationen bestimmt bzw. aus der Summe seiner bedeutsamen unstandardisierten Regressionskoeffizienten (b_m) berechnet werden (Hartig, 2007, S. 91f.). Die unstandardisierten Regressionskoeffizienten können als Rangwerte interpretiert werden: ein Item, bei dem das Merkmal (x_m) vorhanden ist, ist um den Betrag des Regressionskoeffizienten (b_m) Logits schwieriger als ein Item, bei dem dieses Merkmal nicht vorhanden ist (Winther, 2010, S. 223).

Für die quantitative Bestimmung der Schwellen der Kompetenzniveaus auf der Logit-Skala werden die unstandardisierten Regressionskoeffizienten der *bedeutsamen* Merkmale bzw. Merkmalsausprägungen (b_m) verwendet, wobei bedeutsam die Effektstärke (z.B. $\beta \geq .30$), nicht die Signifikanz meint (Hartig, 2007, S. 94f.).¹⁰⁴ Der Logit-Wert, wo ein oder mehrere bedeutsame Merkmale erstmalig systematisch auftreten, kann als untere Schwelle eines Kompetenzniveaus interpretiert werden. Die untere Schwelle eines Kompetenzniveaus wird durch die Merkmalskombinationen der dort lokalisierten Items charakterisiert und von deren erwarteten Schwierigkeiten bestimmt. Aus den vielzähligen Merkmalskombinationen der Testitems werden diejenigen ausgewählt, die den Übergang von einem Kompetenzniveau zum nächsten charakterisieren (Hartig, 2007, S. 92f.).

Die unstandardisierten Regressionskoeffizienten werden in aufsteigender Rangfolge geordnet und aus ihren Teilsummen die jeweils untere Schwelle eines Kompetenzniveaus auf der Logit-Skala berechnet: Die Regressionskonstante b_0 entspricht der Schwierigkeit eines Items „ohne ausgeprägte Schwierigkeit“, d.h. alle schwierigkeitsbestimmenden Merkmale sind auf der jeweils untersten Stufe kodiert. Die Summe aus der Regressionskonstante b_0 und dem Regressionskoeffizienten b_1 entspricht der Schwierigkeit eines Items, bei der nur das Merkmal x_1 vorliegt. Die Summe aus b_0 , b_1 und b_2 entspricht der Schwierigkeit eines Items, bei dem x_1 und x_2 vorliegen etc. (Hartig, 2007, S. 92f.).

Aus der Regressionsanalyse des WBK-T2 (vgl. Tabelle 11-20) ergeben sich vier bedeutsame Merkmale ($\beta \geq .30$) (vgl. Tabelle 11-21). Die b-Koeffizienten dieser Merkmale werden nun in ihrer

¹⁰³ Winther und Achtenhagen (2009) wählten für das Merkmal „Kognition“ vier Stufen: (1) Abfragen/Wiedergeben, (2) Elaborieren und Verstehen, (3) Analysieren und Validieren, (4) Nutzen und Anwenden – und konnten bloss für die Stufe 3 einen signifikanten Einfluss zeigen. Guggemos und Schönlein (2015) wählten für das vergleichbare Merkmal „Informationsverarbeitung“ drei Stufen: (1) Verstehen, (2) Anwenden, (3) Analyse – und konnten für keine Stufe einen signifikanten Einfluss zeigen.

¹⁰⁴ Bei der Etablierung eines Kompetenzniveaumodells geht es nicht um die Generalisierbarkeit eines regressionsanalytischen Zusammenhangs zwischen Item-Schwierigkeit und Merkmalen, sondern um die Beschreibung von spezifischen Anforderungssituationen auf einem bestimmten Niveau, daher ist hier die Effektstärke eines Merkmals relevant, nicht seine Signifikanz (Hartig, 2007, S. 94f.).

aufsteigenden Rangfolge angeordnet und summiert, woraus sich für den WBK-T2 fünf untere Schwellen ergeben: die erste Schwelle liegt bei -1.511 Logits ($b_{\text{Konst.}}$), die zweite Schwelle bei -0.611 Logits ($b_{\text{Konst.}} + b_{\text{complex2}} = -1.511 + 0.900$), die dritte Schwelle bei +0.603 Logits ($b_{\text{Konst.}} + b_{\text{complex2}} + b_{\text{info2}} = -1.511 + 0.900 + 1.214$), die vierte bei +1.911 Logits und die fünfte bei +3.627 Logits.

11.6.3 Qualitative Beschreibung der Kompetenzniveaus

Ein Kompetenzniveau wird zwischen zwei Schwellen bestimmt, d.h. zwischen der unteren und der oberen Schwelle auf der Logit-Skala. Die in unmittelbarer Nähe der Schwelle lokalisierten Item-Schwierigkeiten sind charakteristisch für die Personenfähigkeiten auf diesem Niveau und auf dem nächsthöheren Niveau. Diese Items werden nun gebündelt herangezogen für eine qualitative Beschreibung der fachbezogenen Anforderungen auf diesem Kompetenzniveau. Die anforderungsbezogenen Beschreibungen der Kompetenzniveaus erlauben eine kriteriumsorientierte Interpretation der Testergebnisse, d.h. eine Aussage darüber, über welche spezifischen Kompetenzen Testpersonen auf einem bestimmten Niveau verfügen, welche Anforderungen sie gerade noch (unterhalb der Schwelle) und welche sie nicht mehr (oberhalb der Schwelle) bewältigen können (Hartig, 2007, S. 86).

Der entscheidende Vorteil von Kompetenzniveau modellen ist die kriteriumsorientierte Interpretation der Personenfähigkeiten auf einem Kompetenzniveau. Dabei darf aber nicht ausser Acht gelassen werden, dass die Kompetenzen von Personen nahe unterhalb und nahe oberhalb einer Schwelle ähnlicher sind als die Kompetenzen von Personen im unteren und oberen Bereich des Niveaus. Ein allgemeiner Nachteil von Kompetenzniveau modellen wird im Informationsverlust gesehen, der von der Unterteilung der kontinuierlichen Skala in ordinale Abschnitte herrührt, so dass Unterschiede in Personenfähigkeiten innerhalb eines Niveaus nicht mehr berücksichtigt/interpretiert werden können (Hartig, 2007, S. 85). Die Grenzen der Kompetenzniveau modelle liegen darin, dass die in ihnen definierten Schwellen und beschriebenen Kompetenzniveaus an das verwendete Testmaterial gebunden sind und sich nicht ohne weiteres auf neue Items, z.B. modifizierte und neukonstruierte Items, die bei der Testrevision entstehen, anwenden lassen (Hartig, 2007, S. 96).

Für den WBK-T2 ergeben sich aufgrund der regressionsanalytisch bestimmten b-Koeffizienten und berechneten Schwellen fünf Kompetenzniveaus (vgl. Abbildung 11-22). Die Anforderungskriterien für diese Kompetenzniveaus lassen sich wie folgt beschreiben:

- Kompetenzniveau A: Testpersonen bearbeiten Anforderungssituationen ohne ausgeprägte Schwierigkeiten (Konstante: open1, know1, complex1, info1). Alle benötigten Informationen sind in den Materialien vorgegeben; die inhaltliche Komplexität ist gering, d.h. es müssen keine Zusammenhänge zwischen fachlichen Konzepten hergestellt werden.

- Kompetenzniveau B: Testpersonen bearbeiten Anforderungssituationen mit mittlerer Komplexität (complex2). Alle benötigten Informationen sind in den Materialien vorgegeben, aber es muss ein Zusammenhang zwischen zwei Konzepten hergestellt werden.
- Kompetenzniveau C: Testpersonen bearbeiten Anforderungssituationen mit mittlerer Komplexität und teilweiser Informationsvorgabe (complex2, info2). Die benötigten Informationen sind nur ansatzweise in den Materialien vorgegeben und es muss ein Zusammenhang zwischen zwei Konzepten hergestellt werden.
- Kompetenzniveau D: Testpersonen bearbeiten Anforderungssituationen mit hoher Komplexität und teilweiser Informationsvorgabe (complex3, info2). Die benötigten Informationen sind nur ansatzweise in den Materialien vorgegeben und es muss ein Zusammenhang zwischen mehr als zwei Konzepten hergestellt werden.
- Kompetenzniveau E: Testpersonen bearbeiten Anforderungssituationen mit hoher Komplexität und keiner Informationsvorgabe (complex3, info3). Die benötigten Informationen sind nicht in den Materialien vorgegeben und es muss ein Zusammenhang zwischen mehr als zwei Konzepten hergestellt werden.

Für das vorgeschlagene Kompetenzniveaumodell müssen jedoch folgende Einschränkungen angebracht werden: Die Kompetenzniveaus B (Range = 1.214) und C (Range = 1.308) sind auf einem sehr breiten Logit-Intervall definiert, so dass die formulierten Anforderungskriterien für die Items und Personen auf diesem Niveau als zu vage erscheinen. Eine überwiegende Mehrheit der Testpersonen befindet sich auf den Kompetenzniveaus B (53 %) und C (42 %), so dass eine differenzierte Interpretation der Personenfähigkeiten der Stichprobe WBKgym nur an zwei Schwellen ([-0.611 und 0.603) möglich und sinnvoll ist.

Für die Stichprobe WBKgym und den WBK-T2 zeigt sich folgende Verteilung der Personenparameter auf den Kompetenzniveaus (vgl. Abbildung 11-22): 2.6 % der Testpersonen (N = 10) befinden sich auf Kompetenzniveau A, 54.4 % (N = 200) auf Niveau B, 42 % (N = 157) auf Niveau C und 1.9 % (N = 7) auf Niveau D. Es befindet sich keine Testperson auf dem Kompetenzniveau E, allerdings ist eine Person in einem Bereich unter dem Kompetenzniveau A. Eine überwiegende Mehrheit der Testpersonen (95 %) befindet sich also im Kompetenzniveau C und darüber.

Ob dieses Ergebnis, d.h. die Kumulierung der Personenfähigkeiten auf den Kompetenzniveaus B und C speziell für die Stichprobe WBKgym gilt oder ob es sich auch für andere Stichproben aus der Zielgruppe des Konstrukts reproduzieren lässt, bedarf einer weiteren Validierung mit anderen Stichproben aus der Zielgruppe.

11.7 Zwischenfazit

In diesem Kapitel wurde Evidenz für die interne Struktur des WBK-T2 bzw. seine psychometrische Qualität gesammelt. Für die Validierung der internen Struktur des WBK-T2 wurden psychometrische Analysen mit einem probabilistischen Testmodell, dem Partial-Credit-Rasch-Modell gerechnet. Die Ergebnisse dieses Kapitels sollen Evidenz für die postulierte interne Struktur des WBK-T2 liefern, insbesondere für die psychometrische Qualität des WBK-T2 (Dimensionalität, Skalierbarkeit, Reliabilität) und der einzelnen Items (Homogenität, Trennschärfe, Schwierigkeit).

Für die *Dimensionalität* des WBK-T2 wurde ein eindimensionales Modell gegen drei mehrdimensionale Modelle getestet. Das eindimensionale Modell mit einem latenten Generalfaktor (M1) überwiegt die zweidimensionalen Modelle „Kognitionsprozess“ (M3) und „Antwortformat“ (M4): Die Modellfit-Indizes von M3 und M4 sind zwar leicht besser als derjenige von M1, jedoch sind die latenten Korrelationen der Faktoren im M3 und M4 sehr hoch. Ebenso ist das eindimensionale Modell M1 dem vierdimensionalen Modell „Testinhalt“ (M2) vorzuziehen: M4 hat zwar einen signifikant besseren Modellfit als M1, jedoch sind die latenten Korrelationen zwischen den Faktoren hoch und die Reliabilitäten der einzelnen Faktoren unbefriedigend. Die Daten liefern also nicht genügend empirische Evidenz für ein multidimensionales Konstrukt WBK-soek bzw. eine multifaktorielle Struktur des WBK-T2.

Die *Schätzung* und *Skalierung* der Parameter erfolgten nach der MML-Methode. Die Itemparameter streuen ausreichend breit auf der Logit-Skala und streuen insbesondere breiter als die Personenparameter. Die *Person-Item-Map* illustriert, dass der WBK-T2 für die vorliegende Stichprobe WBKgym – Gymnasiast*innen im Kanton St.Gallen – etwas zu einfach konstruiert ist, da sich der mittlere Personenparameter im positiven Logit-Bereich befindet. Die *Schätzungsgenauigkeit* der Personenparameter kann anhand der *Test Information Function* (TIC) und der Standardfehler als ausreichend genau beurteilt werden. Allerdings ist die Parameterschätzung in den Randbereichen der Logit-Skala am ungenauesten. Zudem können Personen mit etwas hoher Fähigkeit weniger genau geschätzt werden als Personen mit etwas geringer. Die probabilistischen und klassischen Reliabilitätskoeffizienten für die Personenparameter bzw. Personen-Testwerte sind als befriedigend zu beurteilen.

Die *Modellgültigkeit* des Partial-Credit-Rasch-Modells bzw. die *Item-Homogenität* wurde anhand der Item-Infits (wMNSQ), der Schwellenparameter und des *Differential Item Functioning* (DIF) untersucht. Die Item-Infits befinden sich bei allen Items im tolerierbaren Intervall und sprechen für annähernd gleiche Item-Trennschärfe. Die Schwellenparameter der polytomen Items sind – mit wenigen Ausnahmen – ansteigend angeordnet. DIF bezüglich des Schwerpunktfachs (Gruppe WuR, Gruppe Nicht-WuR) wurde auf globaler Ebene bestätigt und konnte bei vier Items festgestellt werden. Auch das DIF bezüglich des Geschlechts (Jungen, Mädchen) wurde auf globaler Ebene

bestätigt, auf lokaler Ebene konnte es bei neun Items nachgewiesen werden. Insgesamt liefern die Daten jedoch genügend empirische Evidenz für die Gültigkeit des Partial Credit Rasch-Modells bzw. für die Item-Homogenität.

Für die *Verteilung der Testwerte* im WBK-T2 konnte eine annähernde Normalverteilung konstatiert werden. Die Qualität der Items im WBK-t2 wurde anhand der klassischen Schwierigkeiten (adjustierte Item-Mittelwerte) und der klassischen Item-Trennschärfen untersucht. Die klassischen Item-Schwierigkeiten streuen auf dem gesamten Wertebereich, jedoch sind sie für fünf Items als kritisch zu beurteilen (d.h. Werte im oberen oder unteren 10 %-Bereich). Die klassischen Item-Trennschärfen sind für die Mehrheit der Items als befriedigend bis gut zu beurteilen; bei vier Items sind sie kritisch. Die Distraktoren-Trennschärfen der Selected-Response-Items sind mehrheitlich als befriedigend zu beurteilen, nur bei zwei Items sind sie kritisch.

Bei der *Graduierung* des WBK-T2 durch die vier schwierigkeitsbestimmenden Merkmale erwiesen sich die Merkmale „inhaltliche Komplexität“ und „Informationsvorlage“ als besonders bedeutsam für die Prädiktion der probabilistischen Item-Schwierigkeiten. Es konnten fünf Kompetenzniveaus beschrieben werden, die eine kriteriumsorientierte Interpretation der Personenfähigkeiten erlauben, aber die Stichprobe WBKgym zu wenig differenzieren.

Aufgrund der psychometrischen und klassischen Itemanalysen in diesem Kapitel sollten folgende 4 bzw. 7 Items – unter Berücksichtigung der Konstrukt-Repräsentativität – modifiziert bzw. ausgeschlossen werden:

- STA_6m und MAN_9n: Antwortoptionen modifizieren aufgrund fachlicher Klarheit/Korrektheit und psychometrischer Indizes ($r_{it-total} \leq .20$, Distraktoren $r_{ptbis} > 0$).
- AHV_13n: ausschliessen aufgrund fachdidaktischer Passung und psychometrischer Indizes ($r_{it-total} \leq .20$, $wMNSQ > 1.15$).
- AHV_2: ausschliessen aufgrund inhaltlicher Relevanz, fachdidaktischer Passung und psychometrischer Indizes ($r_{it-total} \leq .20$).
- AHV_5, ENE_4m, ENE_9m: inhaltlich evaluieren aufgrund psychometrischer Indizes (DIF), ggf. belassen oder ausschliessen.

12 Validierung der Beziehung des WBK-T2 zu externen Merkmalen

In diesem Kapitel wird der Leitfrage (9) nachgegangen (vgl. Kapitel 1.2):

(9) Wie unterscheiden sich verschiedene Personengruppen (Zielgruppe und Kontrastgruppe, Zielsubgruppen) in der sozioökonomischen Facette der WBK? Wie hängen externe Leistungs- und Interessensmerkmale der Testpersonen mit der sozioökonomischen Facette der WBK zusammen?

Mit dieser Leitfrage wird Evidenz für die Beziehung des WBK-T2 mit externen Merkmalen gesammelt (AERA et al., 2014, S. 16ff.). Sie geht einher mit Annahme X bis Annahme XVII zur Beziehung des WBK-T2 zu externen Merkmalen (vgl. Kapitel 8.1.4.a), die nachfolgend überprüft werden. Zur Beantwortung dieser Leitfrage wurden deskriptiv- und inferenzstatistische Analysen mit dem Programm SPSS Statistics Version 25 gerechnet (IBM, 2017). Eingangs und einschränkend soll jedoch betont werden, dass sowohl die Stichprobe WBKgym als auch die aus anderen Projekten verwendeten Vergleichsstichproben nicht global-repräsentativ sind und insofern die nachfolgend präsentierten Ergebnisse nur beschränkt generalisierbar sind, insbesondere was die Aussagen zu Effektstärken und Signifikanz betrifft (vgl. Bortz & Schuster, 2010, S. 89, S. 169f.; Döring & Bortz, 2016, S. 297f.).

In diesem Kapitel werden diverse Ergebnisse zur WBK-soek und externen Merkmalen dargestellt und diskutiert: die kriteriale Validität von Zielgruppe und Kontrastgruppe (Kapitel 12.1), die kriteriale Validität von verschiedenen Zielsubgruppen (Kapitel 12.2). Schliesslich wird die divergente Validität der WBK-soek zu Leistungs- und Interessensmerkmalen untersucht (Kapitel 12.3).

12.1 Kriteriale Validität: WBK-soek von Zielgruppe und Kontrastgruppe

Wie unterscheiden sich Zielgruppe und Kontrastgruppe in der sozioökonomischen Facette der WBK (WBK-soek)? Mit anderen Worten: Welcher Unterschied in der WBK-soek zeigt sich bezüglich des curricularen Vorwissens zwischen der Zielgruppe (WuR) und der Kontrastgruppe (Nicht-WuR)? Aufgrund des Lehrplans und der Stundendotation für das wählbare Schwerpunktfach und das obligatorische Einführungsfach WuR ist davon auszugehen, dass die Zielgruppe WuR mehr curriculares Vorwissen für die WBK-soek mitbringt als die Kontrastgruppe Nicht-WuR (vgl. Annahme X).

Dieses Kapitel ist wie folgt gegliedert: Zunächst werden die Curricula von Zielgruppe und Kontrastgruppe hinsichtlich sozioökonomischer Problemsituationen qualitativ untersucht (Kapitel 12.1.1), um die Testergebnisse der beiden Gruppen auf Ebene der Problemsituationen zu

interpretieren. Danach wird die WBK-soek der Zielgruppe WuR mit der Kontrastgruppe Nicht-WuR deskriptivstatistisch und inferenzstatistisch verglichen (Kapitel 12.1.2).

12.1.1 Qualitative Analysen zu den Lerninhalten in den Curricula

Der WBK-T2 wurde für Jugendliche am Ende der Sekundarstufe II konzipiert (vgl. Kapitel 8.1.3). Für den Testzweck wurde eine Zielgruppe (kaufmännische Berufsfachschule, Berufsmaturitätsschule im Schwerpunkt WuR, gymnasiale Maturitätsschule im Schwerpunkt WuR) und eine Kontrastgruppe (nicht-kaufmännische Berufsfachschulen, Berufsmaturitätsschulen und gymnasiale Maturitätsschule ohne Schwerpunkt WuR) bestimmt. Die Zielgruppe und Kontrastgruppe unterscheiden sich hinsichtlich ihres curricularen fachlichen Vorwissens, insbesondere Fakten- und Konzeptwissen, das für die (abstrakte) Lösung sozioökonomischer Problemsituationen notwendig ist.

Im Projekt WBKgym wurde die Stichprobe entsprechend der im Testzweck des WBK-T2 genannten Zielgruppen und Kontrastgruppen zusammengesetzt (vgl. Kapitel 8.1.3 und 10.1.3.b) und für die Analysen berücksichtigt: Als Zielgruppe fungieren Schüler*innen mit „Wirtschaft & Recht“ im Schwerpunktfach (Gruppe WuR), als Kontrastgruppe jene mit einem anderen Schwerpunktfach bzw. mit „Wirtschaft & Recht“ im Einführungsfach (Gruppe Nicht-WuR). Für die Beurteilung des *potenziellen* curricularen Vorwissens der Gruppen bezüglich sozioökonomischer Problemsituationen sowie für die gruppenspezifische Interpretation der Testergebnisse des WBK-T2 wurden die Lehrpläne der Sekundarstufe II analysiert.¹⁰⁵ Die Inhaltsanalyse bezog sich darauf, inwiefern die Lernziele/Lerninhalte in den Lehrplänen direkt oder indirekt auf sozioökonomische Problemsituationen hinweisen (vgl. Anhang Tabelle A- 2).

Die Beurteilung der Formulierungen erfolgte auf einer 4-stufigen Skala von 0 bis 3 (vgl. Förster, Brückner, et al., 2017) (vgl. Tabelle 12-1). Code 3 („vollumfänglich vorhanden“) bedeutet, dass die Problemsituation der Formulierung des Lernziels explizit genannt ist; z.B. das Lernziel „Arbeitslosigkeit“ in allen drei Lehrplänen oder das Lernziel „Staatsverschuldung“ in zwei Lehrplänen. Mit Code 2 („grösstenteils vorhanden“) ist gemeint, dass die Problemsituation zu einem Lernziel subsumiert, d.h. im Sinne eines Exkurses im Unterricht behandelt werden kann; z.B. Agrarhandel zum Lernziel „internationale Arbeitsteilung/Globalisierung/Freihandel“ oder Umwelt-/Klimaschutz zu den Lernzielen „Marktwirtschaft“ und „Wirtschaftspolitik“. Bei Code 1 („ansatzweise vorhanden“) sind aus dem Lernziel Bezüge zur Problemsituation möglich, aber nicht offensichtlich, wie z.B.

¹⁰⁵ Für die gymnasialen Maturitätsschulen: Rahmenlehrplan für Maturitätsschulen (EDK, 1994) und exemplarisch Lehrpläne für das Gymnasium der Kantone Bern und St. Gallen (Bern, 2005; St.Gallen, 2006). Für die Berufsmaturitätsschulen: Rahmenlehrplan für die Berufsmaturität (SBFI, 2012). Für die kaufmännischen Berufsfachschulen: Bildungsplan und Lernzielkatalog Kauffrau/Kaufmann EFZ (SKKAB, 2011a, 2011b). Für alle anderen Berufsfachschulen: Rahmenlehrplan für den allgemeinbildenden Unterricht (SBFI, 2006).

Flüchtlingsströme beim Lernziel „Sozialpolitik“ oder „Globalisierung“. Bei Code 0 („nicht vorhanden“) finden sich weder explizite noch implizite Hinweise und Bezüge zur Problemsituation.

Insgesamt zeigt sich eine hohe curriculare Passung (mittleres Rating: > 2.0) bei den sozioökonomischen Problemsituationen Arbeitslosigkeit, Energieversorgung, Umwelt-/Klimaschutz, Staatsverschuldung. Eine moderate curriculare Passung (mittleres Rating: $1.5 \leq x \leq 2.0$) kann folgenden acht Problemsituationen attestiert werden: Agrarhandel, Altersvorsorge, bilaterale Abkommen (Rahmenabkommen), Gehaltsgerechtigkeit (Managervergütungen), Krankenversicherung, Steuerwettbewerb/Finanzausgleich, Wohnungsmarkt/Mietpreise, Strassenverkehr.

Zwischen den drei Zielsubgruppen kaufmännische Berufsfachschule (total Rating 30), Berufsmaturitätsschule im Schwerpunkt WuR (total Rating 28), gymnasiale Maturitätsschule im Schwerpunkt WuR (total Rating 30) zeigen sich über alle Problemsituationen hinweg keine substanziellen Unterschiede, wohl aber bei den einzelnen Problemsituationen.

Hingegen ergeben sich substanzielle Unterschiede zwischen den zwei Kontrastsubgruppen nicht-kaufmännische Berufsfachschule (total Rating 16) und gymnasiale Maturitätsschule im nicht-wirtschaftlichem Schwerpunkt (total Rating 7). Bei den nicht-kaufmännischen Berufsfachschulen sind die entsprechenden Lernziele im Lernbereich „Gesellschaft“ von ABU zu finden, der sich in acht Aspekte gliedert, dazu gehören u.a. Ethik, Ökologie, Politik, Recht, Wirtschaft. Allerdings sind die Lernziele des Aspekts Wirtschaft sehr allgemein gehalten, so dass eine valide Zuordnung zu den sozioökonomischen Problemsituationen schwierig ist. Bei der gymnasialen Maturität im nicht-wirtschaftlichem Schwerpunkt umfasst das obligatorische Fach „Einführung WuR“ nur wenig passende Lernziele, da hier Grundlagen aus allen drei Teilbereichen (VWL, BWL, Recht) vermittelt werden und die Stundentafel vergleichsweise gering ist (2 bis 4 Jahreswochenstunden). Bei nicht-kaufmännischen Berufsmaturitätsschulen wird kein obligatorisches Fach „Einführung WuR“ angeboten.

Aufgrund dieser Beurteilung bestätigt sich die Einteilung in Zielgruppe und Kontrastgruppe im Testzweck des WBK-T2 (vgl. Kapitel 8.1.3). Des Weiteren ist diese Beurteilung auch für zukünftige Forschungsprojekte entscheidend, wenn es um die Auswahl/Ziehung einer merkmalsrepräsentativen Stichprobe (Ziel- und Kontrastgruppe) geht. Die Beurteilung ist bei der Interpretation der Ergebnisse des WBK-T2 zu den Gruppen (vgl. Kapitel 12.1 und 12.2.3) als *Erklärungskriterium* zu berücksichtigen, und zwar für den gesamten WBK-T2 (welche Problemsituationen sind enthalten?) als auch für die einzelnen Problemsituationen (inwiefern ist diese Problemsituation curricular bekannt?).

Tabelle 12-1. Beurteilung des potenziellen curricularen Vorwissens für sozioökonomische Problemsituationen der Zielgruppe und Kontrastgruppe auf der Sekundarstufe II

Politikfeld	Sozioökonomische Problemsituation	Zielgruppen				Kontrastgruppen		
		gymMS: WuR	BMS: WuR	kfmBS: WuG	M	gymMS: Nicht-WuR	BS: ABU	M
Arbeit	Gehaltsgerechtigkeit (Managervergütungen)	2	2	2	2.0	1	1	1.0
Energie & Umwelt	Energieversorgung	2	2	3	2.3	0	2	1.0
	Umwelt- und Klimaschutz	2	2	3	2.3	1	3	2.0
Finanzen & Steuern	Staatsverschuldung	3	3	2	2.7	1	1	1.0
	Steuerwettbewerb/Finanzausgleich	2	2	2	2.0	0	1	0.5
Gesundheit	Krankenversicherung	2	1	3	2.0	0	1	0.5
Internationale Beziehungen & wirtschaftliche Integration	Bilaterale Abkommen Schweiz/EU (Rahmenabkommen)	2	3	1	2.0	1	2	1.5
Landwirtschaft	Agrarhandel	2	2	2	2.0	1	2	1.5
Migration & Integration	Flüchtlingsströme	1	1	1	1.0	0	1	0.5
Raumplanung	Wohnungsmarkt/Mietpreise	2	2	2	2.0	0	0	0.0
Soziale Sicherheit & Vorsorge	Altersvorsorge	2	1	3	2.0	1	1	1.0
	Arbeitslosigkeit	3	3	3	3.0	1	1	1.0
Verkehr	Strassenverkehr	2	2	1	1.7	0	0	0.0
Wettbewerb	Service Public	2	1	1	1.3	0	0	0.0
	Strukturpolitik	1	1	1	1.0	0	0	0.0
Total	----	30	28	30	2.0	7	16	0.8

Hinweis: Ratingskala: 0 = nicht vorhanden, 1 = ansatzweise vorhanden, 2 = grösstenteils vorhanden, 3 = vollumfänglich vorhanden. M = Mittelwert. kfmBS = kaufmännische Berufsfachschule, BMS = Berufsmaturitätsschule, gymMS = gymnasiale Maturitätsschule, BS = nicht-kaufmännische Berufsfachschulen. WuG = Wirtschaft & Gesellschaft, WuR = Wirtschaft & Recht als Schwerpunktfach, Nicht-WuR = Wirtschaft & Recht als Einführungsfach, ABU = Allgemeinbildender Unterricht.

Quelle: BSF (2017), SBFI (2006, 2012), SBFI & SKKAB (2011a, 2011b), Erziehungsdirektion des Kantons Bern (2005), Erziehungsdirektion des Kantons St.Gallen (2006); eigene Entwicklung und Darstellung.

12.1.2 Quantitative Analysen zum curricularen Vorwissen (Fach WuR)

Für die deskriptivstatistischen Analysen wurden die adjustierten Item-Mittelwerte verwendet, um die adjustierten Testwerte für den gesamten WBK-T2 und für die einzelnen Problemsituationen zu vergleichen. Der mittlere adjustierte Testwert der gesamten Stichprobe ($M_{\text{adj}} = 0.63$) verweist auf eine moderat ausgeprägte WBK-soek, ebenfalls die mittleren adjustierten Testwerte der Gruppe Nicht-WuR ($M_{\text{adj}} = 0.59$) und der Gruppe WuR ($M_{\text{adj}} = 0.67$) (vgl. Abbildung 12-1, vgl. Kapitel 11.5.1.c).

Der adjustierte Mittelwert des gesamten WBK-T2 fällt für die Gruppe WuR ($M = 0.67$) um 8 % höher aus als für die Gruppe Nicht-WuR ($M = 0.59$, $\Delta M = 0.08$), d.h. dass der Test von der Gruppe WuR besser gelöst wurden als von der Gruppe Nicht-WuR (vgl. Abbildung 12-1). Dieses Ergebnis lässt sich nicht nur für den gesamten WBK-T2, sondern auch für jede einzelne Problemsituation festhalten. Bemerkenswert ist hingegen, dass die beiden Gruppen bei der Problemsituation AHV nur um 4 % ($\Delta M = 0.04$) auseinanderliegen, bei den anderen drei Problemsituationen sind es 9 bzw. 10 % ($\Delta M = 0.09$ bzw. 0.10). Diese deskriptiven Ergebnisse sind mit den theoretischen Vorüberlegungen konform: Die Gruppe WuR bringt aufgrund des gewählten Schwerpunktfaches „Wirtschaft & Recht“ mehr potenzielles curriculares Vorwissen bezüglich sozialökonomischer Problemsituationen mit (vgl. Kapitel 10.1.3).

Für die inferenzstatistischen Analysen wurden Mittelwertvergleiche (t-Tests für unabhängige Stichproben) gerechnet (vgl. Fromm, 2012, S. 12ff.; Kuckartz et al., 2013, S. 159ff.). Als metrische Zielvariable für die WBK-soek wurde der summierte Testwert des WBK-T2 mit 32 Items verwendet, als dichotome Gruppierungsvariable das Schwerpunktfach (WuR = 1, Nicht-WuR = 0) (vgl. Tabelle 12-2). Der Mittelwert in der WBK-soek der Gruppe WuR ist höchst signifikant höher als jener der Gruppe Nicht-WuR ($\Delta M = -4.9$, $t_{(373)} = -7.835$, $p < .001$). Die Gruppenzugehörigkeit „Schwerpunktfach“ hat einen grossen Effekt (Cohens $|d| = 0.80$, $\eta^2 = 0.13$)¹⁰⁶. Die Annahme X zum curricularen Vorwissen kann also bestätigt werden.

Somit können die Befunde aus früheren Studien zur Bedeutung des wirtschaftlichen Schwerpunktfachs für die ökonomische Kompetenz von Gymnasiast*innen repliziert werden (vgl. K. Beck, 1993; Schumann & Eberle, 2014a; Schumann et al., 2013). Allerdings muss bei der Interpretation dieser Ergebnisse berücksichtigt werden, dass der Inhalt des WBK-T2 nicht curricular ausgerichtet ist (vgl. Kapitel 8.2), und dass neben curricularem Vorwissen (formelles/schulisches Lernen) auch nicht-curriculares Vorwissen (informelles/ausserschulisches Lernen) zur WBK-soek beitragen kann. Eine umfassendere Studie zu schulischen und ausserschulischen Kontextfaktoren als Prädiktoren der WBK-soek ist derzeit noch in Arbeit (Ackermann & Peng, 2018, in progress).

¹⁰⁶ Effektstärken werden nach Cohen wie folgt kategorisiert (Cohen, 1988, S. 24-27): $|d| \geq 0.2$ als „small“, $|d| \geq 0.5$ als „medium“ und $|d| \geq 0.8$ als „large“.

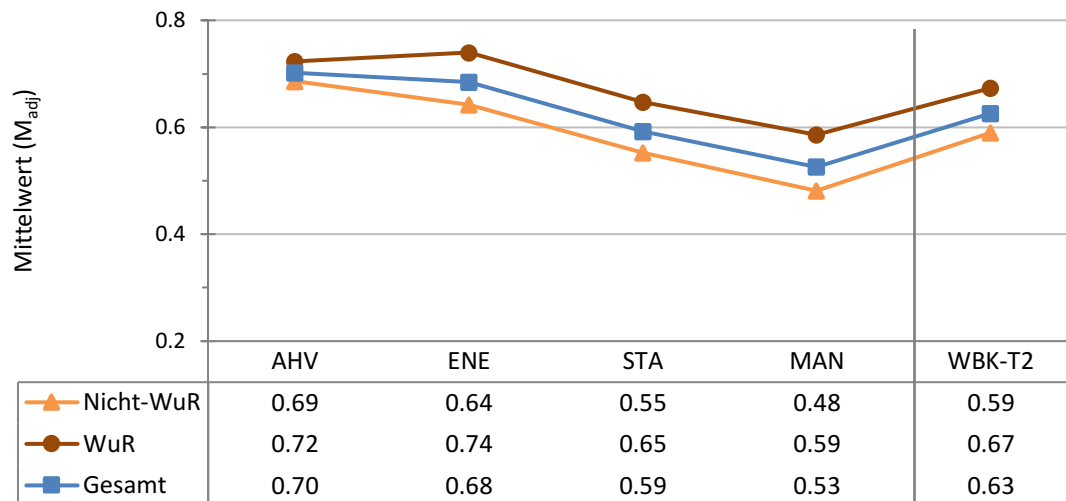


Abbildung 12-1. Adjustierte Testwerte nach Problemsituationen und Schwerpunktfach

Hinweis: $N = 375$, $m = 32$. Adjustierte Testwerte berechnet aus adjustierten Mittelwerten für dichotome und polytome Items, Wertebereich $[0,1]$. Nicht-WuR = anderes Schwerpunktfach, WuR = Schwerpunktfach „Wirtschaft & Recht“. AHV = Altersvorsorge, ENE = Energieversorgung, STA = Staatsverschuldung, MAN = Managervergütungen.

Tabelle 12-2. Mittelwertunterschiede in WBK-soek nach Schwerpunktfach

Gruppe	N	M	SD	ΔM	t	df	p (2-seitig)	d	η^2
Nicht-WuR	216	33.0	6.67	-4.9	-7.835	273	< .001	0.80	0.13
WuR	159	37.9	5.36						
Total	375	35.0	6.60	---	---	---	---	---	---

Hinweis: $N = 375$ Cases, $m = 32$ Items. Zielvariable: WBK-soek (WBK-T2), Summenwert mit Wertebereich $[0,55]$. Gruppierungsvariable: Schwerpunktfach (WuR = 1, Nicht-WuR = 0). Nicht-WuR = anderes Schwerpunktfach, WuR = Schwerpunktfach „Wirtschaft & Recht“. Levene-Test auf Varianzhomogenität nicht angenommen: $F_{(1,373)} = 4.831$, $p = .029$. t-Test für unabhängige Stichproben.

Zudem sind diese Ergebnisse nur mit Vorbehalt für die Grundgesamtheit der Gruppen WuR und Nicht-WuR im Kanton St.Gallen verallgemeinerbar, da es sich beim Projekt WBKgym nicht um eine global-repräsentative Stichprobe, sondern um eine merkmalspezifisch-repräsentative Stichprobe gemäss Quotenplan handelt (vgl. Kapitel 10.1.3). Für eine umfassendere Validierung der WBK-soek bezüglich des curricularen Vorwissens müssten Gymnasiast*innen aus weiteren Deutschschweizer Kantonen untersucht werden. Eine repräsentative Untersuchung für die gesamte Deutschschweiz wurde im Projekt OEKOMA erreicht (Schumann & Eberle, 2014a), allerdings ist das dort untersuchte Konstrukt ein anderes, nämlich die sog. Kerndimension ökonomischer Kompetenzen, die sich auf fachwissenschaftliche Konzepte bezieht (vgl. Kapitel 3.4.2). Die Ergebnisse zum curricularen Vorwissen sind jedoch vergleichbar mit denjenigen im Projekt WBKgym.

12.2 Kriteriale Validität: WBK-soek von Zielsubgruppen

Wie unterscheiden sich verschiedene Zielsubgruppen in der sozioökonomischen Facette der WBK (WBK-soek)?

Dieses Kapitel ist wie folgt gegliedert: Zum einen werden Zielsubgruppen nach biologischem Geschlecht (Kapitel 12.2.1) und soziokulturellem Hintergrund (Kapitel 12.2.2) anhand der Daten aus dem Projekt WBKgym verglichen. Zum anderen werden Zielsubgruppen nach Schulform/Bildungsgang (Kapitel 12.2.3) und Land/Politiksystem (Kapitel 12.2.4) mithilfe von Daten aus verwandten Forschungsprojekten verglichen.

12.2.1 Biologisches Geschlecht

Welcher Unterschied in der WBK-soek zeigt sich bezüglich des biologischen Geschlechts, d.h. bei Mädchen und Jungen, bei den untersuchten Gymnasiast*innen? Bei der Modellierung der WBK-soek und bei der Spezifizierung des WBK-T2 wurden keine Gender-Spezifika intendiert, so dass keine Kompetenzunterschiede zwischen Mädchen und Jungen angenommen werden (vgl. Annahme XI).

Für die inferenzstatistischen Analysen wurden Mittelwertvergleiche (t-Tests für unabhängige Stichproben) gerechnet (vgl. Fromm, 2012, S. 12ff.; Kuckartz et al., 2013, S. 159ff.). Die WBK-soek wurde als metrische Zielvariable (Summenwert des WBK-T2 mit 32 Items) und das Geschlecht als dichotome Gruppierungsvariable (männlich = 1, weiblich = 0) verwendet (vgl. Tabelle 12-3). Jungen haben einen höchst signifikant höheren Mittelwert in der WBK-soek als Mädchen ($\Delta M = 3.1$, $t_{(373)} = -4.579$, $p < .001$). Die Gruppierungsvariable „Geschlecht“ hat jedoch einen (knapp) geringen Effekt auf die WBK-soek (Cohens $|d| = 0.49$, $\eta^2 = 0.05$).

Dieses (vorläufige) Ergebnis würde die bisherigen Befunde zum Geschlechtereffekt bestätigen (vgl. K. Beck, 1993; Brückner et al., 2015; Förster & Happ, under review; Rudeloff et al., under review; Schumann & Eberle, 2014a; Schumann et al., 2013; Walstad et al., 2013a) – zumindest wenn man unterstellte, dass die Stichprobe bezüglich kontextueller Merkmale, insbesondere der schulischen Lerngelegenheiten und des potenziellen curricularen Vorwissens, homogen sei. Eine solche Unterscheidung wurde bei den oben erwähnten Studien für die Analyse des Geschlechtereffekts nicht vorgenommen. Untersucht man jedoch den Mittelwertunterschied in der WBK-soek für das Geschlecht getrennt nach Schwerpunktfach, ergibt sich ein differenziertes Bild (vgl. Tabelle 12-4).

Tabelle 12-3. Mittelwertunterschiede in WBK-soek nach Geschlecht

Gruppe	N	M	SD	ΔM	t	df	p (2-seitig)	d	η^2
weiblich	234	33.9	6.60	-3.1	-4.579	373	< .001	0.49	0.05
männlich	141	37.0	6.15						
Total	375	35.0	6.60	---		---	---	---	---

Hinweis: N = 375 Cases, m = 32 Items. Zielvariable: WBK-soek (WBK-T2), Summenwert mit Wertebereich [0,55]. Gruppierungsvariable: Geschlecht (männlich = 1, weiblich = 0). Levene-Test auf Varianzhomogenität angenommen: $F_{(1,373)} = 1.247$, $p = 0.265$. t-Test für unabhängige Stichproben.

Tabelle 12-4. Mittelwertunterschiede in WBK-soek nach Geschlecht innerhalb Schwerpunktfach

Schwerpunkt	Geschlecht	N	M	SD	ΔM	t	df	p (2-seitig)	d	η^2
Nicht-WuR	weiblich	160	31.9	6.22	-4.3	-4.268	214	< .001	0.65	0.08
	männlich	56	36.1	6.69						
WuR	weiblich	74	38.2	5.20	0.6	0.599	157	.484	0.11	0.00
	männlich	85	37.6	5.51						

Hinweis: N = 375 Cases, m = 32 Items. Zielvariable: WBK-soek (WBK-T2), Summenwert mit Wertebereich [0,55]. Gruppierungsvariable: Geschlecht (männlich = 1, weiblich = 0). Levene-Test auf Varianzhomogenität für Gruppe Nicht-WuR angenommen: $F_{(1,213)} = 1.118$, $p = 0.291$; für Gruppe WuR angenommen: $F_{(1,158)} = 0.105$, $p = 0.747$. t-Test für unabhängige Stichproben.

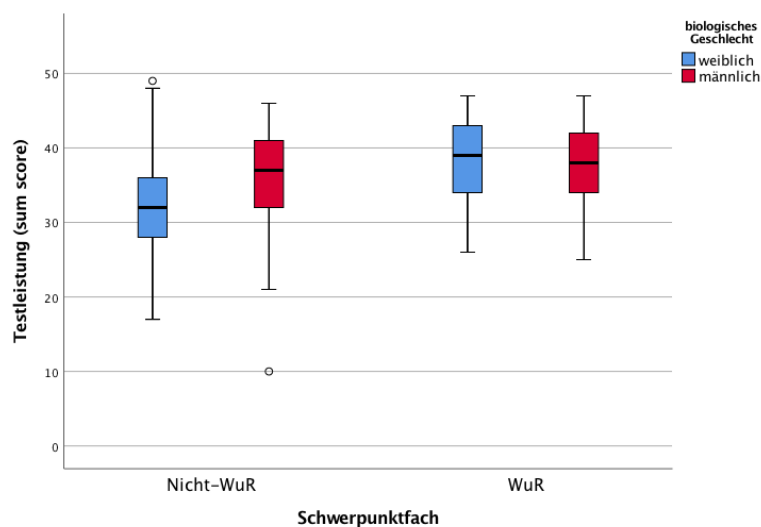


Abbildung 12-2. Boxplots für Testwerte nach Schwerpunktfach und Geschlecht

Innerhalb der Gruppe Nicht-WuR zeigen Jungen eine signifikant höhere WBK-soek als Mädchen ($\Delta M = -4.3$, $t_{(214)} = -4.268$, $p < .001$); die Effektstärke des Mittelwertunterschieds ist jedoch moderat (Cohens $|d| = 0.65$, $\eta^2 = 0.08$). Umgekehrt haben innerhalb der Gruppe WuR Mädchen eine leicht höhere WBK-soek als Jungen, jedoch ist der Unterschied nicht signifikant ($\Delta M = 0.6$, $t_{(157)} = 0.599$, $p = .484$) und die Effektstärke ist gering (Cohens $|d| = 0.11$, $\eta^2 = 0.00$). Der Mittelwertunterschied zugunsten des männlichen Geschlechts ist also nur in der Gruppe Nicht-WuR vorhanden, in der

Gruppe WuR ist er nicht vorhanden. Durch die Wahl und den Besuch des Schwerpunktfaches WuR scheint bei Mädchen und Jungen ein hohes Interesse an ökonomischen Themen zu bestehen bzw. fortzubestehen, was den *gender gap* in der WBK-soek zu nivellieren vermag. Annahme XI zum biologischen Geschlecht kann also nur für die Zielgruppe WuR, nicht aber für die Kontrastgruppe Nicht-WuR bestätigt werden.

Anhand der obigen Ergebnisse lässt sich zudem ablesen, dass Jungen der Gruppe WuR ($M_{\text{WuR(m)}} = 37.6$) bloss eine leicht höhere WBK-soek haben als jene der Gruppe Nicht-WuR ($M_{\text{Nicht-WuR(m)}} = 36.1$, $\Delta M = -1.4$) (vgl. Abbildung 12-2). Demgegenüber haben Mädchen der Gruppe WuR ($M_{\text{WuR(f)}} = 38.2$) eine deutlich höhere WBK-soek als jene der Gruppe Nicht-WuR ($M_{\text{Nicht-WuR(f)}} = 31.9$, $\Delta M = -6.3$). Interessanterweise wird der höchste (49 Punkte) und geringe Testwert (10 Punkte) von je einer Schülerin der Gruppe Nicht-WuR erreicht.

Zusammenfassend können also die Befunde bisheriger Studien zum Geschlechtereffekt bei ökonomischen Kompetenzen von Gymnasiast*innen nur bedingt repliziert werden bzw. müssen dahingehend relativiert werden, als dass sich der Geschlechtereffekt abhängig von einem wirtschaftlichen Schwerpunktfach zeigt. Eine vertiefte Studie der geschlechtsspezifischen Unterschiede im WBK-T2, insbesondere in Abhängigkeit vom Itemformat erscheint demnächst (Ackermann & Siegfried, under review).

12.2.2 Soziokultureller Hintergrund (Erstsprache, Staatsbürgerschaft)

Welcher Unterschied in der WBK-soek zeigt sich bezüglich des soziokulturellen Hintergrundes, d.h. Erstsprache und Staatsbürgerschaft, bei den untersuchten Gymnasiast*innen? Gemäss den Spezifikationen des WBK-T2 sollte die Erstsprache Deutsch das sprachliche Verständnis von sozioökonomischen Problemstellungen begünstigen (vgl. Annahme XII) und die Schweizer Staatsbürgerschaft persönliche Betroffenheit und Interessiertheit für sozioökonomische Problemstellungen erzeugen (vgl. Annahme XIII).

Für die inferenzstatistischen Analysen wurden Mittelwertvergleiche (t-Tests für unabhängige Stichproben) gerechnet (vgl. Fromm, 2012, S. 12ff.; Kuckartz et al., 2013, S. 159ff.). Als erstes wurde der sprachliche Hintergrund als dichotome Gruppierungsvariable (Deutsch = 1, Nicht-Deutsch = 0) verwendet, um den Unterschied in der WBK-soek zu untersuchen (vgl. Tabelle 12-5). Der Mittelwertunterschied in der WBK-soek zwischen der Gruppe DTS und der Gruppe Nicht-DTS ist hoch signifikant zugunsten der ersten Gruppe ($\Delta M = -4.8$, $t_{(373)} = -3.029$, $p = .003$). Die Gruppierungsvariable Erstsprache hat einen mittleren Effekt auf die WBK-soek (Cohens $|d| = 0.74$, $\eta^2 = 0.02$). Annahme XII zur Erstsprache kann also bestätigt werden.

Tabelle 12-5. Mittelwertunterschiede in WBK-soek nach Erstsprache

Gruppe	N	M	SD	ΔM	t	df	p (2-seitig)	d	η^2
Nicht-Deutsch	18	30.5	6.04	-4.8	-3.029	373	.003	0.74	0.02
Deutsch	357	35.3	6.55						
Total	375	35.0	6.60	---	---	---	---	---	---

Hinweis: N = 375 Cases, m = 32 Items. Zielvariable: WBK-soek (WBK-T2), Summenwert mit Wertebereich [0,55]. Gruppierungsvariable: sprachlicher Hintergrund (Deutsch = 1, Nicht-Deutsch = 0). Levene-Test auf Varianzhomogenität nicht angenommen: $F_{(1;373)} = 0.167$, $p = .683$. t-Test für unabhängige Stichproben.

Tabelle 12-6. Mittelwertunterschiede in WBK-soek nach Staatsbürgerschaft

Gruppe	N	M	SD	ΔM	t	df	p (2-seitig)	d	η^2
Nicht-Schweiz	37	32.9	6.99	-2.4	-2.103	373	.036	0.37	0.01
Schweiz	338	35.3	6.52						
Total	375	35.0	6.60	---	---	---	---	---	---

Hinweis: N = 375 Cases, m = 32 Items. Zielvariable: WBK-soek (WBK-T2), Summenwert mit Wertebereich [0,55]. Gruppierungsvariable: staatsbürgerlicher Hintergrund (Schweiz = 1, Nicht-Schweiz = 0). Levene-Test auf Varianzhomogenität nicht angenommen: $F_{(1;373)} = 0.143$, $p = .705$. t-Test für unabhängige Stichproben.

Als zweites wurde der staatsbürgerliche Hintergrund als dichotome Gruppierungsvariable (Schweiz = 1, Nicht-Schweiz = 0) verwendet, um die Unterschiede in der WBK-soek zu untersuchen (vgl. Tabelle 12-6). Der Mittelwertunterschied in der WBK-soek zwischen Schüler*innen mit Staatsbürgerschaft Schweiz und jenen mit anderer Staatsbürgerschaft (Nicht-Schweiz) ist signifikant zugunsten der ersten Gruppe ($\Delta M = -2.4$, $t_{(373)} = -2.103$, $p = .036$). Die Gruppierungsvariable hat einen kleinen Effekt auf die WBK-soek (Cohens $|d| = 0.37$, $\eta^2 = 0.01$). Annahme XIII zur Staatsbürgerschaft kann also bestätigt werden.

Die Befunde aus früheren Studien zur Bedeutung des soziokulturellen Hintergrunds für die ökonomische Kompetenz von Gymnasiast*innen können somit repliziert werden (vgl. Jüttler & Schumann, 2016). Einschränkend muss zu den obigen Ergebnissen bemerkt werden, dass die Stichprobe bezüglich der Erstsprache (95 % Schweizerdeutsch/Deutsch als Erstsprache, 5% Schweizerdeutsch/Deutsch als Zweitsprache) und bezüglich der Staatsbürgerschaft (90 % Schweizer Staatsbürgerschaft, 10 % andere Staatsbürgerschaft) ungleich verteilt ist (vgl. Tabelle 10-4). Diese Verteilungen scheinen aber für die Grundgesamtheit der Gymnasiast*innen im Kanton St.Gallen repräsentativ zu sein. Eine umfassendere Studie zu schulischen und ausserschulischen Kontextfaktoren als Prädiktoren der WBK-soek ist derzeit noch in Arbeit (Ackermann & Peng, in progress; Peng, 2018).

12.2.3 Schulform/Bildungsgang

Welcher Unterschied in der WBK-soek zeigt sich bezüglich der Schulform bzw. des Bildungsgangs, d.h. bei Schüler*innen der gymnasialen Maturitätsschule und der kaufmännischen Berufsfachschule? Aufgrund der Selektionsmechanismen im Bildungssystem, insbesondere beim Übergang von der Sekundarstufe I zur Sekundarstufe II, ist davon auszugehen, dass Gymnasiast*innen höhere allgemeine kognitive Fähigkeiten haben als (kaufmännische) Berufslernende und daher eine höhere WBK-soek verfügen (vgl. Annahme XIV).

Für den Vergleich der WBK-soek nach Schulform wurden Daten aus dem vorangegangenen Projekt CoBALIT beigezogen (Eberle et al., 2016; Schumann et al., 2017). Die Gruppe kaufmännische Berufslernende (kfmBS) entspricht der Schweizer Stichprobe des Projekts CoBALIT. Die Lernenden befanden sich zum Erhebungszeitpunkt im dritten und letzten Ausbildungsjahr der kaufmännischen Branchen SWISSMEM oder SPEDLOGSWISS. Die Gruppe Gymnasiast*innen (gymMS) stammt aus der Stichprobe des Projekts WBKgym (vgl. Kapitel 10.1.3) und befand sich zum Erhebungszeitpunkt in der 11. bzw. 12. Klassenstufe.

Für eine solche vergleichende Analyse ergeben sich zwei Einschränkungen: Erstens kommen aufgrund der unterschiedlichen Zusammensetzung der Testhefte nur die Items der Problemsituationen Energieversorgung (ENE) und Staatsverschuldung (STA) in Frage (vgl. Kapitel 7.1.3). Zweitens können nur diejenigen Items verwendet werden, die durch die Testrevision unverändert geblieben sind oder nur leicht modifiziert worden waren, z.B. sprachlich (vgl. Kapitel 7.2.2). So stehen als Schnittmenge des WBK-T1 und WBK-T2 insgesamt 9 Items aus ENE und STA zur Verfügung, die sich zu einem Testwert von 12 Punkten summieren.

Für die deskriptivstatistischen Analysen wurden die adjustierten Item-Mittelwerte verwendet. Die mittleren adjustierten Testwerte verweisen für beide Gruppen gymMS und kfmBS auf eine moderat ausgeprägte WBK-soek, allerdings befindet sich die Gruppe gymMS ($M_{\text{gymMS}} = 0.64$) im oberen Bereich, die Gruppe kfmBS ($M_{\text{kfmBS}} = 0.54$) im unteren (vgl. Abbildung 12-3, vgl. Kapitel 11.5.1.c). Der Mittelwert des reduzierten WBK-T1/WBK-T2 ist bei der Gruppe gymMS ($M_{\text{gymMS}} = 0.64$) um 10 % höher als von der Gruppe kfmBS ($M_{\text{kfmBS}} = 0.54$, $\Delta M = 0.10$) (vgl. Abbildung 12-3). Die Problemsituation ENE wurde von beiden Gruppen etwa gleich gut gelöst ($\Delta M = -0.01$). Hingegen wurde die Problemsituation STA von der Gruppe gymMS deutlich besser gelöst als von der Gruppe kfmBS ($\Delta M = 0.16$).

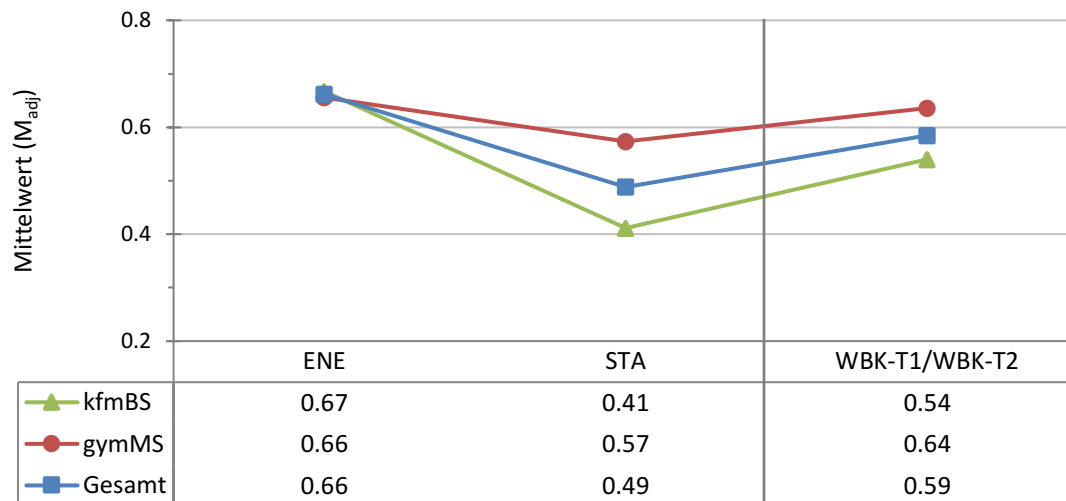


Abbildung 12-3. Adjustierte Testwerte nach Problemsituationen und Schulform

Hinweis: $N = 792$ Cases, $m = 9$ Items. Adjustierte Testwerte berechnet aus adjustierten Mittelwerten für dichotome und polytome Items, Wertebereich $[0,1]$. kfmBS = kaufmännische Berufsfachschule, gymMS = gymnasiale Maturitätsschule. ENE = Energieversorgung, STA = Staatsverschuldung.

Tabelle 12-7. Mittelwertunterschiede in WBK-soek nach Schulform

Gruppe	N	M	SD	ΔM	t	df	p (2-seitig)	d	η^2
kfmBS	417	5.8	2.02	-0.6	-4.082	790	< .001	0.29	0.02
gymMS	375	6.4	2.16						
Total	792	6.1	2.11	---	---	---	---	---	---

Hinweis: $N = 792$ Cases, $m = 9$ Items. Zielvariable: WBK-soek (WBK-T1/WBK-T2 reduziert, ENE und STA), Summenwert mit Wertebereich $[0,12]$. Gruppierungsvariable: Schulform (gymMS = 1, kfmBS = 0). kfmBS = kaufmännische Berufsfachschule, gymMS = gymnasiale Maturitätsschule. Levene-Test auf Varianzhomogenität nicht angenommen: $F_{(1,373)} = 0.167$, $p = .683$. t-Test für unabhängige Stichproben. Stichprobe kfmBS aus Projekt CoBALIT, Branchen SWISSMEM und SPEDLOGSWISS (Schumann et al., 2017).

Diese deskriptiven Ergebnisse sind konsistent mit der Beurteilung des potenziellen curricularen Vorwissens (vgl. Kapitel 12.1.1): „Energie & Ökologie“ ist ein explizites Lernziel im kaufmännischen Bildungsplan für den Lernbereich WuG, was erklären kann, warum die Gruppe kfmBS bei der Problemsituation Energieversorgung gleich gut abschneidet wie die Gruppe gymMS. „Staatsverschuldung“ hingegen ist nur ein implizites Lernziel im kaufmännischen Bildungsplan, was eine Erklärung dafür sein kann, warum die Gruppe kfmBS bei der Problemsituation Staatsverschuldung schlechter abschneidet als die Gruppe gymMS.

Für die inferenzstatistischen Analysen wurden Mittelwertvergleiche (t-Tests für unabhängige Stichproben) gerechnet (vgl. Fromm, 2012, S. 12ff.; Kuckartz et al., 2013, S. 159ff.). Als metrische Zielvariable für die WBK-soek wurde der summierte Testwert des WBK-T1/WBK-T2 mit 9 Items und als dichotome Gruppierungsvariable die Schulform (gymMS = 1, kfmBS = 0) verwendet (vgl. Tabelle 12-7). Der Mittelwert in der WBK-soek der Gruppe gymMS ist höchst signifikant höher jener der

Gruppe kfmBS ($\Delta M = -0.6$, $t_{(790)} = -4.082$, $p < .001$). Die Gruppierungsvariable Schulform hat einen kleinen Effekt auf die WBK-soek (Cohens $|d| = 0.29$, $\eta^2 = 0.02$). Annahme XIV zur Schulform bzw. zum Bildungsgang kann somit bestätigt werden.

Die Interpretation dieser Ergebnisse muss in zweifacher Hinsicht eingeschränkt werden: Erstens ist der für den Vergleich verfügbare Itempool sehr klein ($m = 9$) und die Items stammen aus nur zwei Problemsituationen. Zweitens umfasst die Stichprobe der kaufmännischen Lernenden nur 2 von 21 Branchen in der Schweiz, bei diesen beiden Branchen entspricht sie allerdings beinahe der Grundgesamtheit.

Der Unterschied zwischen den beiden Gruppen dürfte noch grösser ausfallen, wenn in der Gruppe gymMS nur die Schüler*innen im Schwerpunktfach WuR berücksichtigt werden (vgl. Kapitel 12.1). Umgekehrt dürften sich die Gruppen weniger stark unterscheiden, wenn für die Gruppe kfmBS Lernende aus weiteren kaufmännischen Branchen berücksichtigt werden, z.B. Banken, Privatversicherungen, öffentliche Verwaltung. Eine umfassendere Validierung der WBK-soek bezüglich der Schulform kann im Rahmen dieser Dissertation jedoch nicht geleistet werden, wäre aber für eine *Follow-up*-Studie denkbar.

12.2.4 Politiksystem/Land

Welcher Unterschied in der WBK-soek zeigt sich bezüglich des Politiksystems bzw. Lands, d.h. zwischen Schweizer Gymnasiast*innen und deutschen Gymnasiast*innen sowie zwischen Schweizer kaufmännischen Lernenden und deutschen kaufmännischen Lernenden? Aufgrund des politischen Systems der Schweiz, das Staatsbürger*innen mit direkt-demokratischen Rechten ausstattet, werden Schweizer Schüler*innen eher in politische Debatten und Entscheidungen zu sozioökonomischen Problemstellungen eingebunden als deutsche Schüler*innen (vgl. Annahme XV).

Für die inferenzstatistischen Analysen wurden Mittelwertvergleiche (t-Tests für unabhängige Stichproben) gerechnet (vgl. Fromm, 2012, S. 12ff.; Kuckartz et al., 2013, S. 159ff.).

12.2.4.a Gymnasiast*innen in Schweiz und Deutschland

Für den Vergleich der WBK-soek nach Land bei Gymnasiast*innen wurden die Daten der nachgelagerten Studie WBKgymCHDE beigezogen (Siegfried & Ackermann, under review). Die Gruppe deutsche Gymnasiast*innen (gymDE) stammt aus beruflichen Gymnasien im Bundesland Hessen im Leistungskurs „Politik & Wirtschaft“; die Gruppe Schweizer Gymnasiast*innen (gymCH) aus gymnasialen Maturitätsschulen im Kanton St.Gallen im Schwerpunktfach „Wirtschaft & Recht“ (vgl. Kapitel 10.1.3). Für die vergleichende Analyse wurde der reduzierte WBK-T2 mit den zwei Problemsituationen AHV und STA verwendet; die insgesamt 18 Items summieren sich zu einem Testwert von 30 Punkten.

Die Mittelwerte in der WBK-soek der Gruppe gymCH und der Gruppe gymDE unterscheiden sich höchst signifikant zugunsten der Schweizer Gymnasiast*innen ($\Delta M = -4.50$, $t_{(503)} = -11.024$, $p < .001$) (vgl. Tabelle 12-8). Der Effekt der Gruppenzugehörigkeit „Land“ ist gross (Cohens $|d| = 1.20$, $\eta^2 = 0.22$). Annahme XV zum Politiksystem bzw. Land kann für Gymnasiast*innen bestätigt werden.

Allerdings muss dieses Ergebnis in zweifacher Hinsicht relativiert werden. Erstens sind die beiden Stichproben nicht global-repräsentativ für ihre jeweilige Grundgesamtheit und die beiden Gruppen ungleich gross ($N_{\text{gymCH}} = 375$, $N_{\text{gymDE}} = 130$), was zu einer Verzerrung des geschätzten Effekts führen kann (vgl. Bortz & Schuster, 2010, S. 89, S. 169f.; Döring & Bortz, 2016, S. 297f.). Zweitens sind die beiden Gruppen nur bedingt vergleichbar, da das berufliche Gymnasium in Deutschland eine andere „Klientel“ bedient als das allgemeine Gymnasium und somit kein echtes Pendant zum Schweizer Gymnasium darstellt. Eine detaillierte Diskussion dazu findet sich in der eingangs erwähnten Studie (Siegfried & Ackermann, under review).

12.2.4.b Kaufmännische Lernende in Schweiz und Deutschland

Für den Vergleich der WBK-soek nach Land bei kaufmännischen Berufslernenden wurden wiederum die Daten der vorangegangenen Studie aus dem Projekt CoBALIT beigezogen (Eberle et al., 2016; Schumann et al., 2017). Die Gruppe Schweizer kaufmännische Berufsschüler*innen (kfmCH) entspricht der Schweizer Stichprobe, hingegen ist die Gruppe deutsche kaufmännische Lernende (kfmDE) eine Teilstichprobe der deutschen Stichprobe. Für die vergleichende Analyse wurde der reduzierte WBK-T1 verwendet, d.h. das Testheft 6 des rotierten *booklet designs*, das bei beiden Gruppen eingesetzt worden war. Es umfasst die zwei Problemsituationen ENE und STA mit 13 Items, die einen Testwert von 17 Punkten bilden.

Die Mittelwerte der Gruppe kfmCH und der Gruppe kfmDE unterscheiden sich nicht signifikant ($\Delta M = 0.26$, $p = .415$) (vgl. Tabelle 12-9). Der Effekt der Gruppenzugehörigkeit „Land“ ist unbedeutend (Cohens $|d| = 0.09$, $\eta^2 = 0.00$). Annahme XV zum Politiksystem bzw. Land kann also für kaufmännische Berufsschüler*innen nicht bestätigt werden.

Allerdings muss dieses Ergebnis dahingehend relativiert werden, dass erstens die Stichprobengrösse der beiden Gruppen ungleich gross ist und dass zweitens sich die beiden Gruppen hinsichtlich ihrer Vorbildung unterscheiden: Ca. 60 % der deutschen kaufmännischen Lernenden haben ein Abitur, aber bloss ca. 3 % der Schweizer kaufmännischen Lernenden eine Matura (Eberle et al., 2016; Schumann et al., 2017).

Tabelle 12-8. Mittelwertunterschiede in WBK-soek nach Land für Gymnasiast*innen

Gruppe	N	M	SD	ΔM	t	df	p (2-seitig)	d	η^2
gymDE	130	15.5	4.15	-4.5	-11.024	503	< .001	1.20	0.22
gymCH	375	20.0	3.59						
Total	505	18.9	4.22	---	---	---	---	---	---

Hinweis: $N = 505$ Cases, $m = 18$ Items. Zielvariable: WBK-soek (WBK-T2 reduziert, AHV und STA), Summenwert mit Wertebereich $[0,30]$. Gruppierungsvariable: Land (gymCH = 1, gymDE = 0). gymCH = Gymnasium Schweiz, gymDE = Gymnasium Deutschland. Levene-Test auf Varianzhomogenität nicht angenommen: $F_{(1,503)} = 4.763$, $p = .030$. t-Test für unabhängige Stichproben.

Stichprobe gymCH aus Projekt WBKgym, Stichprobe gymDE aus Projekt WBKgymCHDE (Siegfried & Ackermann, under review).

Tabelle 12-9. Mittelwertunterschiede in WBK-soek nach Land für kaufmännische Lernende

Gruppe	N	M	SD	ΔM	t	df	p (2-seitig)	d	η^2
kfmDE	138	8.1	3.44	0.26	0.816	553	.415	0.09	0.00
kfmCH	317	7.8	2.68						
Total	555	7.9	2.88	---	---	---	---	---	---

Hinweis: $N = 555$ Cases, $m = 13$ Items. Zielvariable: WBK-soek (WBK-T1 reduziert, ENE und STA), Summenwert mit Wertebereich $[0,17]$. Gruppierungsvariable: Land (CH = 1, DE = 0). kfmCH = kaufmännische Berufsfachschule Schweiz, kfmDE = kaufmännische Berufsschule Deutschland. Levene-Test auf Varianzhomogenität nicht angenommen: $F_{(1,553)} = 10.523$, $p = .001$. t-Test für unabhängige Stichproben.

Stichproben kfmCH und kfmDE aus Projekt CoBALIT, Branchen MEM und SpedLog in CH, Berufe Industriekaufleute und Speditionskaufleute in DE (Schumann et al., 2017).

12.3 Divergente Validität: WBK-soek, Leistungs- und Interessensmerkmale

Wie hängen externe Leistungs- und Interessensmerkmale der Testpersonen mit der sozioökonomischen Facette WBK (WBK-soek) zusammen? Zwischen den Testwerten des WBK-T2 und den Schulleistungen in den Fächern Deutsch und WuG wird ein positiver, aber geringer Zusammenhang angenommen (vgl. Annahme XVI). Zwischen den Testwerten des WBK-T2 und dem Interesse an sozioökonomischen Problemsituationen wird ein positiver, aber geringer Zusammenhang vermutet (vgl. Annahme XVII).

Für die Beziehung zwischen der WBK-soek und den manifesten Merkmalen wurden Korrelationsanalysen (z.B. Kuckartz et al., 2013, S. 207ff.) mit den Variablen WBK-soek, Schulnote WuR, Schulnote Deutsch, Schulnote Mathematik und Interesse SOEP gerechnet. Die Variable WBK-soek entspricht dem Summenwert des WBK-T2 (Wertebereich [0,55]). Die drei Variablen zu den Schulnoten sind jeweils eine Selbstdeklaration der Testpersonen und wurden mit dem Fragebogen WBK-Q1 erhoben.¹⁰⁷ Die Variable Interesse SOEP (Interesse an sozioökonomischen Problemsituationen) ist ein Skalen-Mittelwert aus drei Items, die mit dem Fragebogen WBK-Q1 erhoben wurden.¹⁰⁸

WBK-soek korreliert schwach positiv mit den Schulnoten des Faches WuR ($r = .21$, $p < .001$), Deutsch ($r = .14$, $p < .01$) und Mathematik ($r = .15$, $p < .01$) (vgl. Tabelle 12-10). Annahme XVI zu den Schulnoten WuR und Deutsch kann also bestätigt werden. Jedoch fällt in dieser Analyse die Korrelation zwischen den Testwerten (WBK-T2) und der Deutschnote geringer aus als in anderen Studien die Korrelation zwischen den Testwerten (OEKOMA-Test) und der Deutschleistung (vgl. Schumann & Eberle, 2014a, 2014b). Dies mag dadurch begründet sein, dass das Schulfach Deutsch inhaltlich breiter ist als ein psychologischer Test zur Erfassung der Deutschkompetenz (z.B. Textverständnis, Textproduktion).

Prima vista scheint der Zusammenhang zwischen der WBK-soek und der Schulnote WuR am grössten zu sein, jedoch ist die Korrelation von WBK-soek und WuR ($r = .21$) nicht signifikant stärker als die Korrelation von WBK-soek und dem Mittelwert aus Deutsch und Mathematik ($r = .19$, $p < .001$; $z = 0.411$, $p = .341$) (vgl. Eid & Schmidt, 2014, S. 548f.). Die Schulnote des Faches WuR korreliert moderat mit den Schulfachnoten Deutsch ($r = .33$, $p < .001$) und Mathematik ($r = .35$, $p < .001$).

¹⁰⁷ Die Werte der drei Items zu den Schulnoten wurden als halbe Noten erfasst, z.B. 4.5. Im Schweizer Bildungssystem haben die Schulnoten einen Wertebereich von [1,6], wobei 1 die schlechteste Note und 6 die beste Note ist (vgl. Anhang Tabelle A- 15).

¹⁰⁸ Die drei Items zum „Interesse SOEP“ wurden auf einer 5-stufigen Antwortskala (1 = trifft nicht zu, 2 = trifft eher nicht zu, 3 = trifft teils-teils zu, 4 = trifft eher zu, 5 = trifft zu) erfasst (vgl. Anhang Tabelle A- 15). Die explorative Faktoren- und Itemanalyse spricht für eine Bündelung der Items zu einer Skala (Cronbachs $\alpha = .71$). Die Skala wurde aus dem Mittelwert der drei Items zum Interesse SOEP gebildet und hat einen Wertebereich von [1,5]. Sie hat eine linksschiefe Verteilung ($M = 3.88$, $SD = 0.77$) (vgl. Anhang Tabelle A- 16).

Tabelle 12-10. Korrelation zwischen WBK-soek und externen Merkmalen

	WBK-soek	Schulnote WuR	Schulnote Deutsch	Schulnote Mathematik	Interesse SOEP
WBK-soek (WBK-T2 Summenwert)	1	0.21*** (< .001)	0.14** (.007)	0.15** (.003)	.21*** (< .001)
Schulnote WuR bzw. VWL (Selbstdeklaration)		1	0.33*** (< .001)	0.35*** (< .001)	.13* (.014)
Schulnote Deutsch (Selbstdeklaration)			1	0.16** (.002)	.06 (.282)
Schulnote Mathematik (Selbstdeklaration)				1	.01 (.921)
Interesse SOEP (Skalen-Mittelwert)					1

Hinweis: $N = 375$ Cases, missing pairwise. Korrelationskoeffizienten als Pearsons r .

Signifikanzniveaus (2-seitig): *** $p < .001$, ** $p < .01$, * $p < .05$.

WuR = Wirtschaft & Recht. VWL = Volkswirtschaftslehre. SOEP = sozioökonomischer Problemsituationen.

WBK-soek korreliert schwach positiv mit dem Interesse an sozioökonomischen Problemsituationen (Interesse SOEP) ($r = .21$, $p < .001$) (vgl. Tabelle 12-10). Annahme XVII zum Sachinteresse kann also bestätigt werden. Dieser Zusammenhang zwischen WBK-soek und Sachinteresse kann als *relativ* hoch beurteilt werden, v.a. angesichts aller anderen Faktoren, welche die Kompetenz beeinflussen. Die Richtung und die Höhe der Korrelation zwischen WBK-soek und Interesse ist konform mit den Befunden anderer Studien (vgl. Schumann & Eberle, 2014a, S. 115f.).

12.4 Zwischenfazit

In diesem Kapitel wurde Evidenz für die Beziehung des WBK-T2 zu externen Merkmalen gesammelt. Für die Validierung dieser Beziehung der WBK-soek zu externen Merkmalen wurden deskriptiv- und inferenzstatistische Analysen (Mittelwert- und Korrelationsanalysen) gerechnet. Die Ergebnisse dieses Kapitels sollen Evidenz für die postulierten Beziehungen des WBK-T2 liefern, insbesondere zu manifesten Aussenkriterien.

Für die kriteriale Validierung der WBK-soek wurde zum einen die Zielgruppe und die Kontrastgruppe untersucht, zum anderen verschiedene Zielsubgruppen. Als Zielgruppe für den WBK-T2 wurden in der Stichprobe WBKgym Gymnasiast*innen im Schwerpunktfach „Wirtschaft & Recht“ (Gruppe WuR) bestimmt, als Kontrastgruppe Gymnasiast*innen in einem anderen Schwerpunktfach (Gruppe Nicht-WuR). Schüler*innen der Gruppe WuR haben eine signifikant höhere WBK-soek als jene der Gruppe Nicht-WuR; der Effekt der Gruppenzugehörigkeit „Schwerpunktfach“ ist gross und höchst signifikant. Die Unterschiede in der WBK-soek waren bei der Problemsituation Altersvorsorge am geringsten.

Für die Zielsubgruppen wurden als Kriterien das Geschlecht, der soziokulturelle Hintergrund (Erstsprache, Staatsbürgerschaft), die Schulform/Bildungsgang sowie das Land gewählt. Jungen haben eine signifikant höhere WBK-soek als Mädchen; dieser Effekt zeigt sich ausgeprägt in der Gruppe Nicht-WuR, jedoch nicht in der Gruppe WuR. Testpersonen der mit Erstsprache Deutsch haben eine signifikant höhere WBK-soek als Testpersonen mit anderer Erstsprache; der Effekt der Gruppenzugehörigkeit „Erstsprache Deutsch“ ist moderat und hoch signifikant. Testpersonen mit Schweizer Staatsbürgerschaft haben eine signifikant höhere WBK-soek als Testpersonen mit anderer Staatsbürgerschaft; der Effekt der Gruppenzugehörigkeit „Staatsbürgerschaft Schweiz“ ist schwach und signifikant. Die WBK-soek von Schüler*innen an gymnasialen Maturitätsschulen ist signifikant höher als jene Schüler*innen an kaufmännischen Berufsfachschulen; es handelt sich jedoch um einen geringen Effekt. Die WBK-soek unterscheidet sich signifikant zugunsten der Schweizer Gymnasiast*innen gegenüber den Deutschen Gymnasiast*innen, jedoch nicht signifikant zwischen den Schweizer kaufmännischen Lernenden und den deutschen kaufmännischen Lernenden.

Für die divergente Validierung wurde die Beziehung der WBK-soek zu externen Leistungs- und Interessensmerkmalen untersucht. Hierbei zeigt sich, dass die WBK-soek schwach positiv mit der Schulfachnote WuR und der Schulfachnote Deutsch korreliert, und ebenfalls schwach positiv mit dem Interesse an sozioökonomischen Problemsituationen.

Abschliessend und einschränkend muss jedoch zu allen Ergebnissen dieses Kapitels noch mal betont werden, dass sowohl die Stichprobe WBKgym als auch die Vergleichsstichproben aus anderen Projekten nicht global-repräsentativ sind und die Aussagen zu Effektstärken und Signifikanz aus den inferenzstatistischen Analysen nur beschränkt generalisierbar sind (vgl. Bortz & Schuster, 2010, S. 89, S. 169f.; Döring & Bortz, 2016, S. 297f.).

TEIL D: SCHLUSSBETRACHTUNG

13 Zentrale Ergebnisse und Erkenntnisse

In diesem Kapitel werden die zentralen Ergebnisse und Erkenntnisse zusammengefasst und diskutiert. Es folgen zuerst die Ergebnisse entlang der Leitfragen (Kapitel 13.1), danach Empfehlungen für eine kleine Revision des WBK-T2 (Kapitel 13.2).

13.1 Ergebnisse entlang der Leitfragen

Die vorliegende Dissertation hat eine dreifache Zielsetzung (vgl. Kapitel 1.2): (A) die theorie- und empiriegeleitete Modellierung der wirtschaftsbürgerlichen Kompetenz (WBK-Modell); (B) die modellbasierte und *Standards*-orientierte Testentwicklung für die sozioökonomische Facette der WBK (WBK-Test); (C) die evidenzbasierte Validierung der Testergebnisse hinsichtlich ihrer Interpretation als sozioökonomische Facette der WBK. Anlehnt an diese dreifache Zielsetzung wurden zehn Leitfragen formuliert, die im Rahmen des Dissertationsvorhabens bearbeitet und in der vorliegenden Dissertationsschrift dargelegt wurden.

13.1.1 Zur Kompetenzmodellierung: WBK-Modell

Nachfolgend werden die Ergebnisse zu den Leitfragen (1) bis (4) zusammengefasst.

Ökonomische Bildung bzw. ökonomische Allgemeinbildung zielt darauf ab, junge Menschen zur Bewältigung *ökonomisch geprägter Lebenssituationen* zu befähigen. Fachdidaktische Konzeptionen der ökonomischen Bildung legen normative Leitbilder und Bildungsziele fest, die insbesondere bei der Auswahl und Gewichtung von Lernzielen/Lerninhalten sowie bei der Formulierung von Bildungsstandards und Kompetenzen zum Tragen kommen. Aus den fachdidaktischen Konzeptionen der ökonomischen Bildung und der wirtschaftsbürgerlichen Bildung lässt sich für «wirtschaftsbürgerliche» Anforderungssituationen folgendes ableiten:

Die *Domäne* der «wirtschaftsbürgerlichen» Anforderungssituationen ergibt sich aus der „ökonomischen Perspektive“. Diese Anforderungssituationen stellen sich im *persönlichen, beruflichen* und *gesellschaftlichen* Lebensbereich (Albers, 1988, 1995), in denen der Mensch unterschiedliche Rollen – Verbraucher*in, Erwerbstätige*r und Wirtschaftsbürger*in – innehat (Kaminski et al., 2017; G. Seeber et al., 2012; B. Weber, 2005, 2014). Für «wirtschaftsbürgerliche» Anforderungssituationen sind verschiedene *Fähigkeiten* erforderlich: das *Verstehen* und *Reflektieren* von wirtschaftlichen und gesellschaftlichen Zusammenhängen, das *Beurteilen* von entsprechenden Problemstellungen bzw. Lösungsansätzen sowie das *Lösen* dieser Problemstellungen (vgl. Dubs, 1985, 2001, 2011, 2013). Die anzuwendenden wissenschaftlichen Kategorien ergeben sich aus der Wirtschaftswissenschaft (Volkswirtschaftslehre, Betriebswirtschaftslehre) und anderen Sozialwissenschaften (z.B.

Rechtswissenschaft, Politikwissenschaft (vgl. Dubs, 2011, 2014; Kaminski et al., 2017; Kaminski et al., 2008; Kruber, 2000, 2005; G. Seeber et al., 2012; B. Weber, 2013, 2014).

Für Kompetenz gibt es kein allgemeines und einheitliches Konzept – und entsprechend entstanden in den letzten Jahren unzählige Definitionen in verschiedenen alltäglichen und wissenschaftlichen Kontexten. In der Bildungsforschung sind kognitive Kompetenzkonzepte verbreitet, jedoch findet sich auch hier keine einheitliche und allgemein akzeptierte Kompetenzdefinition. In empirischen Studien zu Kompetenzmodellierung und -messung werden am häufigsten Definitionen nach dem Kompetenzkonzept der „kontextspezifischen kognitiven Leistungsdispositionen“ *sensu* Weinert (Weinert, 2001a, S. 46) formuliert. Dieses Kompetenzkonzept wurde auch in der vorliegenden Dissertation für die Definition der wirtschaftsbürgerlichen Kompetenz verwendet.

Die in der Wirtschaftspädagogik gängigen Konzepte der *economic literacy* bzw. der ökonomischen Kompetenz beziehen sich traditionellerweise auf eine fachwissenschaftliche Systematik, i.d.R. volkswirtschaftliche Themen. Dabei werden komplexe wirtschaftliche Problemstellungen und deren kontroverse Lösungsansätze vernachlässigt, was jedoch für eine Bildung des allgemeinen Wirtschafts- und Gesellschaftsverständnisses (die sog. «wirtschaftsbürgerliche Bildung») gefordert wird. Das neuere Konzept der wirtschaftsbürgerlichen Kompetenz eignet sich am besten für die in dieser Dissertation angestrebte Kompetenzmodellierung.

Für die Modellierung der wirtschaftsbürgerlichen Kompetenz wurde zunächst die bisherige Definition der WBK (Eberle, 2015; Eberle et al., 2016) kritisch beleuchtet, insbesondere die mehrdeutige und missverständliche Bezeichnung «wirtschaftsbürgerlich» sowie die vage und fragmentarische Beschreibung der Domäne. Für die vorliegende Dissertation wurde, aufgrund der reklamierten Unschärfe, die Definition der WBK modifiziert.

Wirtschaftsbürgerliche Kompetenz bezieht sich auf ökonomisch geprägte Anforderungssituationen im persönlich-finanziellen, im beruflich-unternehmerischen und im gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereich. Die Bewältigung dieser Anforderungssituationen erfolgt durch kognitive Prozesse der Informationsverarbeitung und des Problemlösens unter Einbezug individueller Dispositionen. Sie ist im Ergebnis beobachtbar durch das standardisierte Erfassen von fundiertem Verstehen, systematischem Vergleichen, differenziertem Beurteilen und begründetem Entscheiden.

Basierend auf der obigen modifizierten Definition wurde in der vorliegenden Dissertation ein neues Rahmenmodell für die wirtschaftsbürgerliche Kompetenz entworfen (WBK-Modell), das für Schüler*innen auf der Sekundarstufe II an allgemeinbildenden und berufsbildenden Schulen gelten soll. Der Kern des WBK-Modells sind *ökonomisch geprägte Anforderungssituationen*, darüber werden eine Strukturebene mit *Lebensbereichen* und eine Prozessebene mit *Kognitionsprozessen*

aufgespannt. In den drei Lebensbereichen (s1) persönlich-finanziell, (s2) beruflich-unternehmerisch und (s3) gesamtgesellschaftlich/gesamtwirtschaftlich hat der Mensch unterschiedliche Rollen inne (z.B. Verbraucher*in, Erwerbstätige*r, Staatsbürger*in) und ist mit verschiedenen ökonomisch geprägten Anforderungssituationen konfrontiert. Zu diesen Anforderungssituationen gehören bspw. Geld ausgeben und ansparen, eine Versicherung abschliessen, eine Steuererklärung ausfüllen, sich für eine Berufsausbildung entscheiden, ein eigenes Unternehmen gründen, über wirtschafts-, sozial- und umweltpolitische Fragen entscheiden.

Die informationsverarbeitenden Kognitionsprozesse wurden in drei Stufen zerlegt (vgl. Marzano & Kendall, 2007): (k1) Informationserinnerung (wiedererkennen/wiedergeben), (k2) Informationsverarbeitung (verstehen/anwenden) und (k3) Informationsnutzung (vergleichen/evaluieren/entscheiden). Für die problemlösenden Kognitionsprozesse wurden vier Phasen festgelegt (vgl. Edelmann & Wittmann, 2012, S. 190f.; Gollob et al., 2007; Massing, 1999): (p1) Problem identifizieren und formulieren, (p2) Problemsituation multiperspektiv analysieren, (p3) alternative Lösungsansätze suchen, vergleichen und beurteilen, (p4) eine Lösung implementieren, evaluieren und ggf. modifizieren.

Der gesamtgesellschaftliche/gesamtwirtschaftliche Lebensbereich ist fachwissenschaftlich und fachdidaktisch als Überschneidungsbereich von Ökonomie und Politik zu sehen, in dem sich diverse Politikfelder verorten lassen, z.B. Sozialpolitik, Umweltpolitik, Finanzpolitik. Der politische Problemlöseprozess für gesamtgesellschaftliche/gesamtwirtschaftliche Problemsituationen lässt sich mit dem Modell des Politikzyklus abbilden. Im Schweizer Kontext ist die gesellschaftliche Verantwortung des Einzelnen sowohl in der Bundesverfassung wie auch in den Lehrplänen auf der Sekundarstufe II verankert und hat eine erhebliche praktische Bedeutung, z.B. für die politische Teilhabe. Für den Schweizer Kontext wurden mittels einer umfassenden empirischen Domänenanalyse reale Politikfelder identifiziert (z.B. Energie & Umwelt, Finanzen & Steuern, soziale Sicherheit & Vorsorge) und konkrete sozioökonomische Problemsituationen extrahiert (z.B. Altersvorsorge, Agrarhandel, Energieversorgung, Gehaltsgerechtigkeit, Krankenversicherung, Service Public, Staatsverschuldung, Strassenverkehr). Diese theoretisch und empirisch begründete Beschreibung des gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereichs war für die Konstruktion bzw. Revision des wirtschaftsbürgerlichen Kompetenztests (WBK-Test) notwendig, insbesondere für die Inhaltsspezifikationen des WBK-T2.

13.1.2 Zur Testentwicklung: WBK-T2

Nachfolgend werden die Ergebnisse zu den Leitfragen (5) und (6) zusammengefasst.

Der Test zur wirtschaftsbürgerlichen Kompetenz ist ein psychologischer Leistungstest. Seine erste Version (WBK-T1) wurde im Forschungsprojekt CoBALIT entwickelt: Die Version für die Schweiz enthielt acht Problemsituationen mit insgesamt 63 Items (Version CH), die für Deutschland

adaptierte Version 46 Items (Version DE). Für die Datenerhebung bei kaufmännischen Lernenden wurden jedoch nur zwei Problemsituationen der Version CH und sechs Problemsituationen der adaptierten Version DE eingesetzt. Im Rahmen des Forschungsprojekts WBKgym und insbesondere für diese Dissertation wurde der WBK-T1 evaluiert und umfassend revidiert. Bei der Revision wurden inhaltliche/fachdidaktische Kriterien der Konstrukt-Repräsentativität und -Adäquanz sowie psychometrische Kennzahlen (Item-Infit, Item-Trennschärfe, Item-Schwierigkeit) berücksichtigt, um die Items des WBK-T1 in die drei Kategorien „beibehalten“, „modifizieren“, „ausschliessen“ einzuteilen. Aus der Revision entstand der WBK-T2 mit vier Problemsituationen und insgesamt 32 Items.

Der WBK-T2 ist eine Operationalisierung der *sozioökonomischen Facette der WBK* (WBK-soek). Er beinhaltet sozioökonomische Problemsituationen des gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereichs. Die Testergebnisse sollen kriteriumsorientiert interpretiert und für diagnostische Zwecke verwendet werden. Der WBK-T2 adressiert Schüler*innen der Sekundarstufe II: Als Zielgruppe fungieren Schüler*innen mit potenziellem curricularem Vorwissen in WuR/WuG, als Kontrastgruppe Schüler*innen ohne oder mit einem sehr geringen curricularen Vorwissen. Inwiefern diese Aufteilung in Zielgruppe und Kontrastgruppe gerechtfertigt ist, bedarf weiterer Untersuchungen, die im Rahmen dieser Dissertation nicht zu leisten sind.

Für die *Inhaltsbereiche* des WBK-T2 (Problemsituationen) wurde auf die umfassende Domänenanalyse des gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereichs zurückgegriffen und die dort identifizierten Problemsituationen anhand fünf Kriterien weiter selektiert: Repräsentativität, Wissenschaftlichkeit, Komplexität/Kontroversität/Multiperspektivität, Bekanntheit und Adaptierbarkeit. Schliesslich wurden für den WBK-T2 vier Inhaltsbereiche, sog. Problemsituationen ausgewählt: Altersvorsorge (AHV), Energieversorgung (ENE), Staatsverschuldung (STA), Managervergütungen (MAN). Diese Inhaltsbereiche des WBK-T2 (Problemsituationen) können als Facetten des Konstrukts WBK-soek betrachtet werden, aber sollen beliebig erweiterbar und austauschbar sein. Die Inhaltsbereiche des WBK-T2 müssen vor jedem neuen Testeinsatz bezüglich der erwähnten Kriterien überprüft werden. Insbesondere muss die Relevanz der sozioökonomischen Problemsituationen für die Zielgruppe hinsichtlich institutioneller und politischer Aspekte (z.B. Schulform/Bildungsgang, Branche/Beruf, Kanton, Land) validiert werden. Es könnten neue sozioökonomische Problemsituationen entwickelt und im WBK-T2 integriert werden, z.B. Arbeitslosigkeit, Umwelt- und Klimaschutz.

Die Reihenfolge der Problemsituationen im WBK-T2 Testheft sollte sorgfältig abgewogen werden. Aufgrund der Erfahrungen im Projekt WBKgym empfiehlt sich für den weiteren Einsatz des WBK-T2 folgende Reihenfolge der Problemsituationen: AHV, ENE, STA, MAN. Die vorgeschlagene

Reihenfolge der Problemsituationen soll einer Demotivation und Frustration der Testpersonen während der Testung vorbeugen. Die Items der Problemsituation AHV entpuppten sich im Mittel als *relativ* einfacher, jene der Problemsituation MAN als *relativ* schwieriger. Die Themen AHV und ENE haben am meisten politische Relevanz (Vorlagen der eidgenössischen Volksabstimmungen) und curriculare Relevanz (Lerninhalte in Lehrplänen der Sekundarstufe II); die Problemsituation MAN hat am wenigsten.

Jede Problemsituation beginnt mit einem Einleitungstext, gefolgt ca. acht Items. Die *Inhaltselemente* des WBK-T2 (Items) wurden aufgrund einer fachdidaktischen Inhaltsanalyse jeder Problemsituation festgelegt. Die Items variieren im Kognitionsprozess: (k1) wiedererkennen/wiedergeben, (k2) verstehen/anwenden und (k3) vergleichen/beurteilen/entscheiden), wobei es mehr Items im unteren und mittleren Kognitionsbereich als im oberen gibt. Zudem variieren die Items im Antwortformat: (f1) Selected-Response, (f2) Short-Constructed-Response, (f3) Extended-Constructed-Response, wobei das Verhältnis von gebundenen und freien Antwortformaten ziemlich ausgeglichen ist. Der WBK-T2 enthält insgesamt 32 Items. Diese Testlänge ist angesichts des Gütekriteriums der Zumutbarkeit eine obere Grenze dar. Bei der Testdurchführung wurden dafür ca. 60 Minuten eingeplant.

Die Einleitungstexte des WBK-T2 müssen hinsichtlich ihrer politischen Aktualität vor jedem neuen Testeinsatz geprüft und ggf. angepasst werden. Insbesondere müssen aktuelle Informationen (Faktenblätter, Medientexte, Positionspapiere) recherchiert und die entsprechenden Textbausteine im Einleitungstext aktualisiert werden.

13.1.3 Zur evidenzbasierten Validierung

Nachfolgend werden die Ergebnisse zu den Leitfragen (7) bis (9) zusammengefasst. Die evidenzbasierte Validierung erfolgt entlang von drei Evidenzquellen (vgl. AERA et al., 2014): Testinhalt, interne Struktur und externe Kriterien.

Bei der Validierung des *Testinhalts* wurde die Relevanz und Adäquanz der Inhaltsbereiche (Problemsituationen) und der Inhaltselemente (Items) des WBK-T2 hinsichtlich des Konstrukts WBK-soek untersucht. Zum einen war diese Passung bei den Inhaltsspezifikationen durch die vorangegangene Domänenanalyse des gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereichs bereits berücksichtigt worden, zum anderen sollte sie durch externe Experten beurteilt werden. Für diese externe Beurteilung wurden Leitfaden-Interviews mit Experten geführt. Die Interviews wurden mittels zusammenfassender Inhaltsanalyse ausgewertet. Die Ergebnisse der Interviews flossen in die Revision des WBK-T1 ein, bei der ca. 60 % der Items modifiziert bzw. neukonstruiert wurden. Die Experten beurteilten jedes Item des WBK-T2 anhand der folgenden Kriterien: domänenspezifische Repräsentativität (bezüglich Problemsituation), fachliche Adäquanz (fachwissenschaftlich, fachdidaktisch), sprachliche Adäquanz und zielgruppenbezogene Adäquanz.

Die Items wurden von den Experten mehrheitlich als domänenspezifisch repräsentativ sowie fachlich und sprachlich adäquat beurteilt. Hingegen wurden die Items nur für die Zielgruppe WuR als überwiegend passend beurteilt, für die Kontrastgruppe Nicht-WuR aufgrund des fehlenden curricularen Vorwissens als überwiegend nicht passend.

Zudem wurden kognitive Interviews mit Proband*innen einer Zielsubgruppe geführt. Die Kriterien für die kognitiven Interviews waren sprachliche Klarheit (Fremdwörter, Fachbegriffe) und formelle Klarheit (Layout, Instruktion, Frageformulierung, Antwortstrukturierung/Antwortraum). Die Proband*innen gaben mit ihrem „lauten Denken“ Hinweise auf formelle und sprachliche Mängel im Testheft des WBK-T2. Auch diese Ergebnisse fanden Eingang in die Testrevision.

Für die Validierung der internen Struktur und der externen Kriterien wurde bei einer Stichprobe von 375 Gymnasialschüler*innen (42 % WuR, 58 % Nicht-WuR) im Deutschschweizer Kanton St.Gallen eine Testung mit dem WBK-T2 sowie mit einem Fragebogen zu Individual- und Kontextmerkmalen (WBK-Q1) durchgeführt. Die Kodierung der freien Schüler*innenantworten erfolgte durch zwei unabhängige Rater anhand eines detaillierten Kodierungsmanuals; die Inter-Rater-Reliabilität war substanziell (Spearman's $\rho \geq .86$, Cohens $\kappa \geq .75$). Für die Datenauswertung wurden zum einen probabilistische Analysen für das Partial-Credit-Rasch-Modell mit dem Programm ACER ConQuest 4 gerechnet, zum anderen deskriptiv- und inferenzstatistische Analysen mit dem Programm SPSS Version 25.

Bei der Validierung der *internen Struktur* wurden folgende Aspekte untersucht: Dimensionalität der WBK-soek bzw. faktorielle Struktur des WBK-T2, Parameterschätzung und -skalierung mittels MML-Methode, Schätz- bzw. Messgenauigkeit der Personenparameter und Personen-Testwerte, Gültigkeit des Partial-Credit-Rasch-Modells bzw. Item-Homogenität, Verteilung der Testwerte, Qualität der Items sowie Graduierung des WBK-T2.

Der WBK-T2 weist eine eindimensionale faktorielle Struktur auf, und zwar bezüglich Inhaltsbereichen, Kognitionsprozessen, Antwortformaten – bei ersterem allerdings weniger deutlich. Die Personen- und Itemparameter konnten mit der MML-Methode konsistent geschätzt und skaliert werden. Die Person-Item-Map zeigt, dass die Itemparameter ausgewogen über den ganzen Logit-Bereich streuen und dass 41 % der Personenparameter im negativen und 59 % im positiven Logit-Bereich liegen (Personenparameter $M_\theta = 0.525$, $VAR_\theta = 0.394$; restringierte Itemparameter $M_\sigma = 0$, $VAR_\sigma = 1.460$). Dies impliziert, dass der WBK-T2 für die Stichprobe WBKgym etwas zu einfach konzipiert ist. Entsprechend ist die Schätzgenauigkeit der einzelnen Personenparameter gemäss der Testinformationskurve im negativen Logit-Bereich grösser als im positiven Logit-Bereich. Die Schätzgenauigkeit der globalen Personenparameter und der Personen-Testwerte ist als befriedigend zu beurteilen (EAP/PV = .76, WLE = .74, Cronbachs $\alpha = .74$).

Die Item-Infits lagen für alle 32 Items im Toleranzintervall ($0.92 \leq wMNSQ \leq 1.17$), woraus sich die Annahme paralleler IC-Kurven des Rasch-Modells bestätigen lässt. Die Schwellenparameter der polytomen Items waren bei den allermeisten Items aufsteigend angeordnet, woraus sich die Annahme von polytom geordneten Antwortkategorien im Partial-Credit-Rasch-Modell bestätigen lässt. *Differential item functioning* wurde auf globaler Ebene bezüglich des Schwerpunktfachs (Fokusgruppe WuR, Referenzgruppe Nicht-WuR) und bezüglich des Geschlechts (Fokusgruppe Jungen, Referenzgruppe Mädchen) gefunden. Auf lokaler Ebene zeigte beim Schwerpunktfach nur ein Item einen grossen und signifikanten DIF-Effekt, beim Geschlecht waren es drei Items. Insgesamt ist das vorliegende DIF für den Testzwecks des WBK-T2 als unkritisch zu beurteilen.

Die Testwerte zeigen eine annähernde Normalverteilung. Die klassischen Item-Schwierigkeiten streuen über den ganzen Wertebereich, jedoch werden drei Items als zu einfach ($M_{adj} \geq .95$) bzw. zu schwierig ($M_{adj} \leq .05$) beurteilt. Die klassischen Item-Trennschärfen sind für die allermeisten Items als genügend bis gut zu beurteilen ($r_{it-total} \geq .20$).

Für die Gradierung des WBK-T2 wurden vier schwierigkeitsbestimmende Merkmale festgelegt und alle Items hinsichtlich der Merkmale von Experten eingeschätzt. Zwei schwierigkeitsbestimmende Merkmale („inhaltliche Komplexität“, „Informationsvorlage“) erwiesen sich als bedeutsam für die Prädiktion der Itemparameter und für die Bestimmung der Schwellen. Es konnten fünf Kompetenzniveaus beschrieben werden, die eine kriteriumsorientierte Interpretation der Personenfähigkeiten erlauben.

Bei der Validierung der Beziehung zu *externen Merkmalen* wurden zwei Aspekte berücksichtigt. Zum einen wurde die *kriteriale Validität* der WBK-soek für die Zielgruppe (WuR) und Kontrastgruppe (Nicht-WuR) sowie für verschiedene Zielsubgruppen (biologisches Geschlecht, soziokultureller Hintergrund, Schulform/Bildungsgang, Politiksystem/Land) untersucht. Zum anderen wurde die *divergente Validität* der WBK-soek zu externen Leistungs- und Interessensmerkmalen untersucht: Schulleistungen in WuR und Deutsch sowie Sachinteresse an sozioökonomischen Problemstellungen.

Schüler*innen der Zielgruppe WuR (im Schwerpunktfach WuR) verfügen über eine stark höhere WBK-soek als jene der Kontrastgruppe Nicht-WuR (in anderem Schwerpunktfach). Jungen haben eine signifikant höhere WBK-soek als Mädchen; dieser Effekt zeigt sich ausgeprägt in der Gruppe Nicht-WuR, jedoch nicht in der Gruppe WuR. Schüler*innen mit Erstsprache Deutsch haben eine moderat höhere WBK-soek als jene mit Zweitsprache Deutsch, Schüler*innen mit Schweizer Staatsbürgerschaft eine schwach höhere WBK-soek als jene mit anderer Staatsbürgerschaft. Schweizer Gymnasiast*innen verfügen über eine moderat höhere WBK-soek als Schweizer kaufmännische Lernende. Schweizer Gymnasiast*innen haben eine stark höhere WBK-soek als

deutsche. Schweizer kaufmännische Lernende haben keine systematisch höhere WBK-soek als deutsche.

Es zeigt sich ein schwach positiver Zusammenhang zwischen WBK-soek und den Schulfachnoten WuR und Deutsch. Zudem zeigt sich ein schwach positiver Zusammenhang zwischen WBK-soek und dem Interesse an sozioökonomischen Problemstellungen.

Insgesamt deuten die untersuchten Validierungsaspekte darauf hin, dass die Testergebnisse des WBK-T2 valide hinsichtlich des Konstrukts WBK-soek interpretiert werden können. Allerdings scheint der WBK-T2 für die vorliegende Stichprobe WBKgym etwas zu einfach konstruiert. Dieser Befund ist jedoch aus Perspektive der Validierung als nicht problematisch zu werten, da die ausgewählte Stichprobe wohl den oberen Bereich der Deutschschweizer Gymnasiast*innen repräsentiert. Im schweizweiten Vergleich hat der Kanton St.Gallen für das Einführungsfach WuR eine sehr hohe Stundentafel (4 Jahreswochenstunden) (Eberle & Brüggenbrock, 2013, Anhang, S. 108ff.), was sich insbesondere im curricularen Vorwissen der Kontrastgruppe niederschlägt, sowie eine sehr geringe Maturitätsquote (13.2 % im Jahr 2011), was sich in höheren allgemeinen kognitiven Fähigkeiten der Stichprobe spiegelt (Eberle & Brüggenbrock, 2013, S. 102).

13.2 Empfehlungen für Testrevision des WBK-T2

Aufgrund der Ergebnisse zur evidenzbasierten Validierung, insbesondere der klassischen und probabilistischen Itemanalyse, empfiehlt es sich, vor einem erneuten Einsatz des WBK-T2 eine kleine Revision durchzuführen. Konkrete Vorschläge zu dieser Revision des WBK-T2, insbesondere zur Elimination und Modifikation bestehender Items und zur Konstruktion neuer Items finden sich in der „Dokumentation der Entwicklung und Revision des WBK-Tests“ (Ackermann, 2018a). Diese Vorschläge wurden bereits mit externen Experten diskutiert und sind nachfolgend zusammengefasst (vgl. Anhang Tabelle A- 5 und Tabelle A- 6).

- Die Problemsituation „Mangervergütungen“ (MAN) sollte inhaltlich breiter gefasst und in „Gehaltsgerechtigkeit“ umbenannt werden. So könnten weitere fachlich relevante Inhaltselemente integriert und entsprechende Items entwickelt werden, wie z.B. „Mindestlohn und Arbeitsmarkt“ (MAN_11n) und „Gehaltsunterschiede Frau/Mann“ (MAN_12n). Im WBK-T2 weist MAN eine vergleichsweise zu geringe Repräsentativität und Wissenschaftlichkeit auf. Auf eine umfassende Revision der Problemsituation musste im Projekt WBKgym aus ressourcentechnischen Gründen verzichtet werden; dies könnte jedoch in einem Folgeprojekt realisiert werden.
- Der Einleitungstext MAN sollte um zwei Fachbegriffe ergänzt werden, die das Bearbeiten eines Items (MAN_1m) auch ohne fachspezifisches Vorwissen ermöglicht.

- Bei zwei Selected-Response-Items sollten aufgrund inhaltlicher/fachdidaktischer Kriterien und psychometrischer Kennwerte die *Antwortoptionen modifiziert* werden (STA_6mm, MAN_9nm).
- Bei drei Items ist zu prüfen, ob sie gegebenenfalls aufgrund DIF bezüglich Schwerpunktfach und/oder Geschlecht ausgeschlossen werden (AHV_5, ENE_4m, ENE_9m).
- Zwei Items des WBK-T2 sollten aufgrund inhaltlicher/fachdidaktischer Kriterien und psychometrischer Kennzahlen *eliminiert* werden (AHV_2, AHV_13n).
- Zwei Items des WBK-T1 sollten aufgrund ihrer inhaltlichen Relevanz in den WBK-T2 *integriert* werden, allerdings leicht modifiziert (AHV_3m, AHV_9m).
- Beim Itemtypus II sollte die *Kodierung* von Score 1 (dichotom) zu Score 2 (polytom geordnet) modifiziert werden, um die Antworten der Testpersonen qualitativ differenzierter zu kodieren (fachliche Korrektheit, Argumentationsstruktur). Es handelt sich hierbei um vier Items (AHV_8, ENE_8, STA_6, MAN_5).
- Beim Itemtypus IV (SWOT-Tabelle) sollte geprüft werden, ob gegebenenfalls die *Kodierung* von Score 2 zu Score 3 modifiziert wird. Mit einer vierstufigen Kodierung lassen sich die Antworten der Testpersonen differenzierter kodieren, nämlich bezüglich quantitativer Aspekte (Begründungsumfang i.S.v. Anzahl Kriterien beantwortet) und qualitativer Aspekte (fachliche Korrektheit, Begründungsqualität). Es handelt sich hierbei um 4 Items (AHV_12m, ENE_11m, STA_8m, MAN_12n).
- Beim Itemtypus VI (Diagramm) sollte das *Antwortformat* zu einem „echten“ Selected-Response mit drei Antwortoptionen, d.h. mit drei Diagrammen modifiziert werden, um die komplexen Zusammenhänge kognitionsbezogen vernetzt, statt additiv abzufragen. Es handelt es sich hierbei um vier Items (AHV_14n, ENE_12nm, STA_10nm, MAN_10nm).

14 Limitationen und Perspektiven

Dieses Kapitel betont die Limitationen der vorliegenden Dissertation bezüglich ihrer dreifachen Zielsetzung (Kapitel 14.1). Es entwirft zudem Perspektiven für weitere Forschungsvorhaben, die sich mit dieser Dissertation auf tun (Kapitel 14.2).

14.1 Limitationen

Die Limitationen werden anhand der Schlagworte im Titel der Dissertationsschrift aufgezeigt: «Wirtschaftsbürgerliche Kompetenz Deutschschweizer Gymnasiastinnen und Gymnasiasten: Kompetenzmodellierung, Testentwicklung, evidenzbasierte Validierung».

«*Wirtschaftsbürgerliche Kompetenz*». Der für diese Dissertation konstruierte bzw. revidierte Test zur wirtschaftsbürgerlichen Kompetenz (WBK-T2) beinhaltet reale sozioökonomische Problemsituationen. Diese wurden basierend auf einer umfassenden Domänenanalyse des gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereichs identifiziert und mittels fünf Kriterien selektiert. Nichtsdestotrotz sind die Problemsituationen im WBK-T2 nicht konkret-real, sondern abstrakt-fiktiv zu lösen, *à la*: Wie würde ich in dieser Situation begründet entscheiden, wenn ich es müsste? Ob die Testpersonen in realen Situationen, d.h. im Vorfeld von politischen Abstimmungen, überhaupt und vergleichbar analysieren, reflektieren und argumentieren ist damit noch nicht erwiesen. So zeigen mitunter die Forschungsergebnisse aus der *behavioural economics* und der *economic psychology*, dass das Verhalten von Menschen nicht (nur) durch rationales Denken gesteuert ist, sondern auch durch soziale Normen und Emotionen. Es stellt sich also die grundsätzliche Frage, inwiefern die «Bewältigung ökonomisch geprägter Lebenssituationen» im Allgemeinen und die Bewältigung sozioökonomischer Problemsituationen im Speziellen mittels eines Leistungstest im *Paper-and-Pencil*-Verfahren simuliert werden kann. Diese Frage kann dahingehend zugespitzt werden, ob es sich bei der im WBK-T2 operationalisierten WBK-soek um ein Kompetenzkonzept *sensu* Weinert (Weinert, 2001a, S. 46) oder um ein *literacy*-Konzept *sensu* Walstad et al. (Walstad & Rebeck, 2001; Walstad et al., 2013b) handelt. Die Autorin lässt diese Frage in Anbetracht der anerkannten Unschärfe beider Begriffe und der guten Gründe für beide Konzepte – vorerst – offen.

«*Deutschschweizer Gymnasiast*innen*». Die in dieser Dissertation dargestellten und diskutierten Ergebnisse deuten darauf hin, dass die mit dem WBK-T2 erzeugten Testwerte bei den ausgewählten Gymnasiast*innen im Kanton St.Gallen valide hinsichtlich des Konstrukts WBK-soek interpretiert werden können. Allerdings handelt sich bei dieser Stichprobe um eine merkmalspezifisch-repräsentative Stichprobe, so dass die Ergebnisse nicht global, sondern nur

vorbehältlich dieser Merkmale in der Grundgesamtheit auf alle Gymnasiast*innen im Kanton St.Gallen übertragen werden können. Zudem sind die Ergebnisse dieser Dissertation aufgrund des föderalistischen Schulwesens in der Schweiz nicht vorbehaltlos vom Kanton St.Gallen auf andere Deutschschweizer Kantone übertragbar. In einer *Follow-up-Studie* müsste daher untersucht werden, ob sich die Ergebnisse zur evidenzbasierten Validität des Konstrukts WBK-soek, insbesondere die interne Struktur und die Beziehung zu externen Merkmalen, bei einer Vergleichsstichprobe aus anderen Kantonen als robust erweisen. Für eine weitere Validierung würden sich z.B. die Kantone Bern und Basel Stadt anbieten; beide erliessen ebenfalls einen kantonalen Lehrplan für gymnasiale Maturitätsschulen, aber weisen eine höhere Maturitätsquote (18.8 % bzw. 29.3 %) und eine tiefere Stundentafel für das Einführungsfach „Wirtschaft & Recht“ (2 Stunden) aus (vgl. Eberle & Brüggenschrock, 2013, Anhang).

Als weitere Limitation dieser Dissertation muss erwähnt werden, dass der WBK-T2 in deutscher Sprache konzipiert ist und nur in der deutschsprachigen Schweiz eingesetzt werden kann. Eine Übersetzung des WBK-T2 ins Französische oder Italienische ist bisher nicht erfolgt und auch nicht geplant. Dem stehen forschungsökonomische Gründe im Wege, da eine Übersetzung sehr zeit- und kostenintensiv ist. Auch wären bei einem WBK-T2 mit mehreren Sprachversionen die probabilistischen DIF-Analysen bezüglich der Sprachgruppen zu prüfen. Aus epistemologischer Perspektive ist dies zweifach bedauerlich, denn: Sozioökonomische Problemsituationen beziehen sich auf die gesamte Schweiz und nicht nur auf die Deutschschweiz; Wirtschaftsfächer werden auch an den *Collèges* der französischsprachigen Schweiz („économie et droit“) und an den *Licei* der italienischsprachigen Schweiz („scienze umane ed economiche“).

«Kompetenzmodellierung». Das Rahmenmodell der wirtschaftsbürgerlichen Kompetenz (WBK-Modell), das in dieser Dissertation entworfen wurde, strukturiert «wirtschaftsbürgerliche» Anforderungssituationen in drei Lebensbereiche: persönlich-finanziell, beruflich-unternehmerisch, gesamtgesellschaftlich/gesamtwirtschaftlich. In dieser Dissertation wurde die Domäne des gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereichs detailliert beschrieben und das Konstrukt sozioökonomische Facette der WBK (WBK-soek) mit dem Test zur wirtschaftsbürgerlichen Kompetenz (WBK-T2) operationalisiert sowie die Testergebnisse des WBK-T2 hinsichtlich des Konstrukts WBK-soek erstmals und umfassend validiert. Allerdings können die oben dargestellten und diskutierten Ergebnisse nicht hinsichtlich der gesamten WBK, d.h. allen drei Facetten interpretiert werden; dies war im Rahmen dieses Dissertationsvorhabens weder möglich noch beabsichtigt. Eine detaillierte Beschreibung des persönlich-finanziellen Lebensbereichs und des beruflich-unternehmerischen Lebensbereichs und eine Operationalisierung dieser beiden Facetten müsste jedoch in nachfolgenden Projekten ebenfalls angegangen werden, um das WBK-Modell

vollständig abzubilden und empirisch zu überprüfen. Dabei könnten bereits bestehende Leistungstests für *financial literacy* hinsichtlich ihrer Passung zum WBK-Modell geprüft und ggf. adaptiert werden (z.B. Förster, Happ, et al., 2017; Rudeloff, 2019; Walstad & Rebeck, 2016, 2017).

«*Testentwicklung*». Der WBK-T2 ist das Ergebnis einer umfassenden Revision des WBK-T1, der fachdidaktische und psychometrische Kriterien zugrunde gelegt wurden. Der WBK-T2 zeigt, wie erhofft und beabsichtigt, bessere probabilistische und klassische Item-Kennwerte als der WBK-T1. Trotzdem beinhaltet der WBK-T2 noch einige „querulante“ Items; diese müssten im Rahmen einer kleinen Testrevision eliminiert, modifiziert oder neu konstruiert werden – und schliesslich auch erprobt werden. Empfehlungen für die konkrete Revision dieser Items wurden bereits oben formuliert. Derzeit läuft eine von der Autorin initiierte Pilotierung mit diesen revidierten Items des WBK-T2.

«*Evidenzbasierte Validierung*». In dieser Dissertation wurde eine evidenzbasierte Validierung des Konstrukts WBK-soek angestrebt und dafür Evidenz aus drei Quellen zusammengetragen: Testinhalt, interne Struktur und Beziehung zu externen Merkmalen. Die ersten beiden Evidenzquellen wurden m.E. genügend ausgeschöpft, jedoch mussten bei der dritten Evidenzquelle forschungsökonomisch begründete Kompromisse eingegangen werden. So wurde für die konvergente und divergente Validierung behelfsmässig auf Schulfachnoten zurückgegriffen und nicht auf andere Leistungstests. Für eine „echte“ konvergente Validierung müssten jedoch die Ergebnisse des WBK-T2 mit den Ergebnissen eines verwandten Tests (z.B. WBT, OEOKMA-Test) verglichen werden, für eine „echte“ divergente Validierung mit jenen eines nicht-verwandten Tests (z.B. Sprachtest, Mathematiktest). Erst damit können zuverlässige Aussagen über die Beziehung des Konstrukts WBK-soek zu anderen Konstrukten gemacht werden.

Auf weitere Evidenzquellen wurde aus ressourcentechnischen Gründen verzichtet, wobei der Einbezug sämtlicher Evidenzquellen für die Validierung auch nicht zwingend ist. Im Hinblick auf die vorgeschlagene kleine Revision des WBK-T2 bietet nun aber die Evidenzquelle Denkprozesse ein bisher nicht vollständig ausgeschöpftes Potenzial, um die modifizierten und neu konstruierten Items empirisch zu untersuchen. Mittels kognitiver Interviews mit Testpersonen aus der Zielgruppe des WBK-T2 könnten die beobachteten Denkprozesse zu den betreffenden Items mit den erwarteten verglichen und nicht beabsichtigte „Stolpersteine“ erkannt werden.

14.2 Perspektiven

Ausgehend von dieser Dissertation tun sich vielerlei Perspektiven für weitere Forschungsvorhaben zur ökonomischen Bildung und zur ökonomischen Kompetenz auf.

Konzeption der ökonomischen Bildung. In der Schweiz wird die Fachdidaktik der ökonomischen Bildung bzw. die Wirtschaftsdidaktik traditionellerweise von der Unterrichtspraxis her gelehrt und gelernt. Dabei geht es hauptsächlich um fachdidaktisch verkleidete, allgemeindidaktische Fragen wie z.B. Unterrichtsvorbereitung, Unterrichtsgestaltung (Lehrmethoden, Lernformen) und Leistungsbeurteilung (vgl. Dubs, 1985, 2009, 2014; Euler & Hahn, 2014). Die Antworten auf diese Fragen haben sich m.W. seit den 1990er-Jahren nicht grundlegend verändert. Eine wissenschaftliche Auseinandersetzung in der und mit der Wirtschaftsdidaktik, wie sie z.B. in Deutschland passiert, fehlt hierzulande grösstenteils. Dies wäre m.E. aber dringend nötig, um eine theoretisch fundierte und differenzierte Konzeption für die Leitidee des «mündigen Wirtschaftsbürgers» und die Zielsetzung der «Bildung des allgemeinen Wirtschafts- und Gesellschaftsverständnisses» für die Sekundarstufe II zu entwickeln. So stellt sich mitunter die Frage, wie diese Zielsetzung erreicht und überprüft werden soll und *inwiefern* sie für die verschiedenen Schulformen (Maturitätsschulen, Berufsschulen) und Ausbildungsschwerpunkte (z.B. Einführung WuR, Schwerpunkt WuR) differenziert werden muss. Daraus ergeben sich wiederum Fragen der Entwicklung/Revision von Lehrplänen¹⁰⁹ (Auswahl und Begründung der Lernziele/Lerninhalte) in den verschiedenen Bildungsgängen der Sekundarstufe II sowie der Formulierung kompetenzorientierter Bildungsstandards für die verschiedenen Abschlüsse auf der Sekundarstufe II.

Kompetenzmessung in verschiedenen Schulformen. Das WBK-Modell soll für Schüler*innen an gymnasialen und beruflichen Schulen Anwendung finden. Der WBK-T2 wurde entsprechend für Schüler*innen am Ende der Sekundarstufe II konzipiert. Allerdings wurde der WBK-Test wurde bislang erst bei ausgewählten Stichproben in der Deutschschweiz eingesetzt: der WBK-T1 bei kaufmännischen Lernenden zweier Branchen (SWISSMEM und SPEDLOGSWISS, vgl. Projekt CoBALIT) und der WBK-T2 bei Gymnasiast*innen eines Kantons (St.Gallen, vgl. Projekt WBKgym). Der

¹⁰⁹ Für die kaufmännische Berufsbildung hat die SKKAB Anfang 2018 das Reformprojekt „Kaufleute 2022“ lanciert, bei dem die Bildungsverordnung und der Bildungsplan revidiert werden (SKKAB, 2018). In der ersten Phase des Projekts werden anhand einer Berufsfeldanalyse branchenübergreifende und branchenspezifische Kompetenzen identifiziert, die für die zukünftige flexible und digitale Arbeitswelt relevant sind. In der zweiten Phase wird ein didaktisches Rahmenkonzept für alle drei Lernorte – Ausbildungsbetrieb, überbetriebliche Kurse (üK), Berufsschule – entwickelt. Zur Diskussion steht mitunter die Auflösung der bisherigen allgemeinbildenden Lernbereiche (W&G, IKA, etc.) in einem allgemeinbildenden Lernbereich.

Für die gymnasialen Maturitätsschulen wird derzeit diskutiert, welche Punkte des Maturitäts-Anerkennungsreglements (MAR) und des Rahmenlehrplans für die Maturitätsschulen (RLP-MS) bei der nächsten Evaluation angegangen werden müssten.

optimierte WBK-T2 kann insofern als „Steilpass“ für weitere Untersuchungen zur WBK-soek mit dem WBK-T2 betrachtet werden. Zum Beispiel könnte mit einer gross angelegten Studie die WBK-soek bei verschiedenen Zielgruppen und Kontrastgruppen erfasst und schliesslich fachdidaktische und bildungspolitische Empfehlungen abgeleitet werden.

Kompetenzmessung auf verschiedenen Schulstufen. Mit Einführung des Lehrplan 21 in den Deutschschweizer Kantonen ab dem Schuljahr 2017/2018 wird der Kompetenzbereich „Wirtschaft, Arbeit, Haushalt“ (WAH) im 8. und 9. Schuljahr unterrichtet (vgl. D-EDK, 2014). Somit erwerben die Schüler*innen ökonomische und finanzielle Kompetenzen früher in ihrer Schulbiografie als bis anhin. Dies wirkt sich mitunter auf das curriculare Vorwissen beim Übertritt in die Sekundarstufe II aus. Der WBK-T2 könnte also ein zusätzliches Potenzial für die längsschnittliche Erfassung der WBK-soek entfalten, d.h. am Anfang der Sekundarstufe II und am Ende der Sekundarstufe II.

Kompetenzmessung im Längsschnitt (Kohorten-Design). Die «Bewältigung ökonomisch geprägter Lebenssituationen» nimmt mit Abschluss der Sekundarstufe II kein abruptes Ende, sondern setzt sich im ganzen Leben der Menschen fort. Mit dem WBK-T2 böte es sich an, die Entwicklung der WBK-soek über einen längeren Zeitraum zu untersuchen, z.B. fünf und zehn Jahre nach Abschluss der Sekundarstufe II. Methodisch denkbar wäre ein echtes Längsschnitt-Design mit den Testpersonen der Stichprobe WBKgym oder ein Quasi-Längsschnitt-Design (Kohorten-Design) mit einer neuen Stichprobe bei Studierenden an Hochschulen.

Kompetenzmessung in verschiedenen deutschsprachigen Ländern. Durch die Adaption des WBK-T1 und des WBK-T2 für Deutschland (Version DE) kann die WBK-soek auch bei deutschen Schüler*innen am Ende der Sekundarstufe II erfasst werden. Allerdings wurden der WBK-T1 und der WBK-T2 erst in ausgewählten Bundesländern und Schulformen eingesetzt (Schumann et al., 2017; Siegfried & Ackermann, under review). Eine Adaption für Österreich und Liechtenstein ist derzeit nicht geplant, da sehr zeitaufwändig und kostspielig, erscheint aber *prima vista* durchaus möglich. Mit dem adaptierten WBK-T2 ergibt sich also ein weiteres Potenzial für internationale Studien zur WBK-soek. Eine solche Studie wurde von der Autorin im Bundesland Hessen bereits initiiert und durchgeführt. Weitere Studien sind in den Bundesländern Bayern, Baden-Württemberg und Rheinland-Pfalz angedacht.

Lehr-Lern-Forschung zu problemorientierten Unterrichtsszenarien. In der Fachdidaktik der ökonomischen Bildung werden v.a. problem- und handlungsorientierte Unterrichtsszenarien vorgeschlagen und ihre Bedeutung im Hinblick auf die «Bewältigung ökonomisch geprägter

Lebenssituationen» betont (vgl. z.B. Albers, 1995; Dörig, 2003; Eberle, 2006a; Steinmann, 1995; A. Weber, 2007). Der Bildungsplan für die kaufmännische Grundbildung fordert z.B. auch für den Lernbereich „Wirtschaft & Gesellschaft“ eine sog. Handlungskompetenzorientierung (SKKAB, 2011a). Die Wirksamkeit solcher problemorientierter Lehr-Lern-Arrangements auf die Leistung bzw. Kompetenz der Schüler*innen ist m.W. für den gymnasialen und beruflichen Bildungsbereich in der Schweiz noch unzulänglich untersucht. Die im WBK-T2 enthaltenen sozioökonomischen Problemsituationen könnte also als Instrument für solche Interventionsstudien eingesetzt werden. Dafür könnten beispielsweise je ein traditionelles und *problemorientiertes Lehr-Lern-Arrangements* zu einem Testinhalt des WBK-T2 entwickelt und die Leistung der Schüler*innen im Pre-/Post-Testing mit dem WBK-T2 erfasst werden. Eine solche Interventionsstudie wurde mit der Problemsituation „Energieversorgung“ aus dem adaptierten WBK-T2 in Deutschland bereits durchgeführt (Siegfried & Ackermann, in progress), weitere Interventionsstudien sind angedacht.

Lehr-Lern-Forschung zu Socio Scientific Issues und Socio Scientific Argumentation. In der Fachdidaktik der ökonomischen Bildung werden ausgehend vom Leitbild des «mündigen Wirtschaftsbürgers» Bildungsstandards und Kompetenzen formuliert, die sich u.a. auf das Verstehen, Reflektieren, Urteilen und Entscheiden beziehen. Mit Bezug auf die modellierte WBK-soek sollte in den Wirtschaftsfächern das differenzierte und reflektierte Argumentieren der Schüler*innen zu sozioökonomischen Problemstellungen nicht nur gezielt gefördert, sondern auch empirisch untersucht werden. Ausgewählte Items des WBK-T2 könnten als Instrument verwendet werden, um die *schriftliche Argumentation* der Schüler*innen in kurzen Essays zu analysieren. Eine erste explorative Studie zur Argumentationsstruktur zu sozioökonomischen Problemsituationen wurde mit den Daten des Projekt WBKgym von der Autorin bereits durchgeführt (Ackermann & Kavadarli, in progress). Zudem könnte *mündliche Argumentation* der Schüler*innen in dafür konzipierten Lehr-Lern-Arrangements aufgezeichnet und ausgewertet werden, sog. *dialogic teaching and learning*.

Mit diesen vielfältigen Forschungsperspektiven zeigt sich auch, warum die mit dieser Dissertation angestrebte Kompetenzmodellierung, Testentwicklung und evidenzbasierte Validierung unabdingbar war. Ohne Evidenz für eine valide Interpretation der WBK-T2 Ergebnisse hinsichtlich der WBK-soek kann nämlich weder der Test noch das Modell in der fachdidaktischen Forschung und Praxis „seriös“ verwendet werden. Welche der skizzierten Forschungsvorhaben zukünftig umgesetzt werden, hängt nicht zuletzt von den Forschungsambitionen der Autorin, von synergetischen Forschungsk Kooperationen und von akquirierten Forschungsmitteln ab.

LITERATURVERZEICHNIS

- Ackermann, N. (2018a). *Dokumentation der Entwicklung und Revision des Tests zur Wirtschaftsbürgerlichen Kompetenz (WBK-Test)*. Unveröffentlicht, Universität Zürich, Institut für Erziehungswissenschaft.
- Ackermann, N. (2018b). *Dokumentation des Fragebogens zu Individual- und Kontextfaktoren der Wirtschaftsbürgerlichen Kompetenz (WBK-Q1): Skalen- und Item-Spezifikationen, Skalen-Kennwerte, Kodierungsmanual*. Unveröffentlicht, Universität Zürich, Institut für Erziehungswissenschaft.
- Ackermann, N. (2018c). *Dokumentation des revidierten Tests zur Wirtschaftsbürgerlichen Kompetenz (WBK-T2): Item-Spezifikationen, Item-Kennwerte, Kodierungsmanual*. Unveröffentlicht, Universität Zürich, Institut für Erziehungswissenschaft.
- Ackermann, N., & Eberle, F. (2016). Financial Literacy in Switzerland. In C. Aprea, E. Wuttke, K. Breuer, N. K. Koh, P. Davies, B. Greimel-Fuhrmann, & J. S. Lopus (Eds.), *International Handbook of Financial Literacy* (pp. 341-355). Singapore: Springer.
- Ackermann, N., & Kavadarli, B. (2019). *Argumentation zu sozioökonomischen Problemsituationen: Eine explorative Studie der schriftlichen Argumentationsstruktur von Schüler*innen in einem Leistungstest*. Paper presented at the GEBF Tagung 25.-27.02.2019, Köln.
- Ackermann, N., & Kavadarli, B. (in progress). *Argumentation on Socio-Economic Issues: An Exploratory Study of Students' Written Argumentation Structure in a Performance Test*.
- Ackermann, N., & Peng, P. (2018). *Wirtschaftsbürgerliche Kompetenz Deutschschweizer Gymnasiast*innen: Die Bedeutung schulischer und ausserschulischer Kontextfaktoren*. Paper presented at the BWP Tagung, 03.-05.09.2018, Frankfurt a.M.
- Ackermann, N., & Peng, P. (in progress). *Wirtschaftsbürgerliche Kompetenz Deutschschweizer Gymnasiast*innen: Die Bedeutung schulischer und ausserschulischer Kontextfaktoren*.
- Ackermann, N., Ruoss, T., & Flury, C. (2018). Warum fördern sie ökonomische Bildung? Aktivitäten, Argumente und Handlungslegitimationen von Akteuren am Beispiel der Schweiz. *bwp@ Berufs- und Wirtschaftspädagogik – online*, 35, 1-19. Retrieved from [http://www.bwpat.de/ausgabe35/ackermann et al bwpat35.pdf](http://www.bwpat.de/ausgabe35/ackermann_et al bwpat35.pdf) (13.12.2018)
- Ackermann, N., & Siegfried, C. (under review). 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 (CSEE), Special Issue "Gender-specific Differences in Economic Competence"*.
- Adams, R. J., & Osses, A. (2016). ACER ConQuest Command Reference: Supplement and Update for Section 4 of the ConQuest Manual Version 2. Retrieved from <https://www.acer.org/files/Command-Reference.pdf> (31.08.2018)
- Adams, R. J., Wilson, M. R., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit. *Applied Psychological Measurement*, 21, 1-24.
- Adams, R. J., & Wu, M. L. (2010a). ACER ConQuest Tutorial 2: Modelling Polytomously Scored Items With The Rating Scale and Partial Credit Models. Retrieved from <https://www.acer.org/files/Conquest-Tutorial-2-RatingScaleAndPartialCreditModels.pdf> (31.08.2018)
- Adams, R. J., & Wu, M. L. (2010b). ACER ConQuest Tutorial 6: Differential Item Functioning. Retrieved from <https://www.acer.org/files/Conquest-Tutorial-6-DifferentialItemFunctioning.pdf> (31.08.2018)
- Adams, R. J., & Wu, M. L. (2010c). ACER ConQuest Tutorial 7: Multidimensional Models. Retrieved from <https://www.acer.org/files/Conquest-Tutorial-7-MultidimensionalModels.pdf> (31.08.2018)

- Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). ACER ConQuest: Generalised Item Response Modelling Software (Version 4.0). Camberwell, Victoria: ACER Imprint. Retrieved from <https://www.acer.org/conquest> (31.08.2018)
- AERA, APA, & NCME. (1985). *Standards for Educational and Psychological Testing (4th edition)*. Washington, DC: American Psychological Association (APA).
- AERA, APA, & NCME. (1999). *Standards for Educational and Psychological Testing (5th edition)*. Washington, DC: American Psychological Association (APA).
- AERA, APA, & NCME. (2014). *Standards for Educational and Psychological Testing (6th edition)*. Washington, DC: American Educational Research Association (AERA).
- Albers, H.-J. (1988). Ökonomische Bildung und Allgemeinbildung. In BföB (Ed.), *Ökonomische Bildung: Aufgabe für die Zukunft* (pp. 1-15). Bergisch Gladbach: Verlag Thomas Hobein.
- Albers, H.-J. (1995). Handlungsorientierung und ökonomische Bildung. In H.-J. Albers (Ed.), *Handlungsorientierung und ökonomische Bildung* (pp. 1-22). Bergisch Gladbach: Verlag Thomas Hobein.
- Andersen, E. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, 38, 123-140.
- Anderson, J. R. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.
- Anderson, J. R. (1993). *Rules of the Mind*. Hillsdale, NJ: Lawrence Erlbaum.
- Anderson, J. R. (1996). ACT: A Simple Theory of Complex Cognition. *American Psychologist*, 51(4), 355-365.
- Anderson, L. W., & Krathwohl, D. R. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Addison-Wesley.
- APA. (1954). Technical Recommendations for Psychological Tests and Diagnostic Techniques. *Psychological Bulletin*, 51(2), 1-38.
- APA, AERA, & NCME. (1966). *Standards for Educational and Psychological Tests and Manuals (2nd edition)*. Washington, DC: American Psychological Association (APA).
- APA, AERA, & NCME. (1974). *Standards for Educational and Psychological Testing (3rd edition)*. Washington, DC: American Psychological Association (APA).
- Apra, C. (2014). Finanzielle Allgemeinbildung: Entwurf einer bildungstheoretisch verankerten Konzeptualisierung. *Zeitschrift für Didaktik der Gesellschaftswissenschaften*, 5, Themenschwerpunkt Bildung, 68-89.
- Apra, C., Leumann, S., & Gerber, C. (2014). Finanzielle Allgemeinbildung bei Berufslernenden. *Soziale Sicherheit CHSS*(1), 22-24.
- Apra, C., & Wuttke, E. (2016). Financial Literacy of Adolescents and Young Adults: Setting the Course for a Competence-Oriented Assessment Instrument. In C. Apra, E. Wuttke, K. Breuer, N. K. Koh, P. Davies, B. Greimel-Fuhrmann, & J. S. Lopus (Eds.), *International Handbook of Financial Literacy* (pp. 397-414). Singapore: Springer.
- Apra, C., Wuttke, E., Leumann, S., & Heumann, M. (2015). Kompetenzfacetten von Financial Literacy: Sichtweisen verschiedener Akteure. In J. Seifried, S. Seeber, & B. Ziegler (Eds.), *Jahrbuch der berufs- und wirtschaftspädagogischen Forschung 2015* (pp. 11-22). Opladen: Budrich.
- Baumert, J., Bos, W., & Lehmann, R. (2000). *TIMSS/III Dritte Internationale Mathematik- und Naturwissenschaftsstudie: Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn (Band 1)*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft*, 9(4), 469-520.
- Beaton, E., & Allen, N. (1992). Interpreting Scales through scale anchoring. *Journal of Educational Statistics*, 17, 191-204.

- Beck, B., & Klieme, E. (Eds.). (2007). *Sprachliche Kompetenzen: Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)*. Weinheim: Beltz Verlag.
- Beck, K. (1989). Ökonomische Bildung - Zur Anatomie eines wirtschaftspädagogischen Begriffs. *Zeitschrift für Berufs- und Wirtschaftspädagogik*, 85(7), 579-596.
- Beck, K. (1993). Dimensionen der ökonomischen Bildung: Messinstrumente und Befunde. Abschlussbericht zum DFG-Projekt: Wirtschaftskundlicher Bildung-Test (WBT). Normierung und internationaler Vergleich. Nürnberg: Universität Erlangen-Nürnberg.
- Beck, K., & Krumm, V. (1989). *Economic Literacy in German Speaking Countries and the United States. First Steps to a Comparative Study*. Paper presented at the Annual meeting of AERA, San Francisco.
- Beck, K., & Krumm, V. (1998). *Wirtschaftskundlicher Bildungs-Test (WBT)*. Göttingen: Hogrefe.
- Beck, K., Landenberger, M., & Oser, F. (Eds.). (2016). *Technologiebasierte Kompetenzmessung in der beruflichen Bildung: Ergebnisse aus der BMBF-Förderinitiative ASCOT*. Bielefeld: wbv.
- Benner, D., & Brüggem, F. (2010). Bildsamkeit/Bildung. In D. Benner & J. Oelkers (Eds.), *Historisches Wörterbuch der Pädagogik* (pp. 174-215). Weinheim: Beltz Verlag.
- Bern. (2005). *Lehrplan gymnasialer Bildungsgang, 9.-12. Schuljahr, in Kraft seit 01.08.2006*. Bern: Erziehungsdirektion des Kantons Bern.
- BFS. (2017). Eidgenössische Volksabstimmungen 1848-2017: Detaillierte Ergebnisse (Stand am 23.11.2017). Neuchâtel: Bundesamt für Statistik.
- Bijedic, T. (2013). 'Unternehmerisch handeln macht Schule': Legitimation, Voraussetzungen und Ergebnisse einer Entrepreneurship Education in der Sekundarstufe II. *Zeitschrift für ökonomische Bildung (ZföB)*, 1, 44-72.
- Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie*, 223(1), 3-13. doi:10.1027/2151-2604/a000194
- Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives. The classification of educational goals. Handbook I: Cognitive domain*. New York: Longmans.
- BMBF. (2012). Vocational skills and competencies made visible: The ASCOT research initiative. Bonn: Federal Ministry of Education and Research (BMBF).
- Bortz, J., Lienert, G. A., & Boehnke, K. (2008). *Verteilungsfreie Methoden in der Biostatistik (3. Auflage)*. Heidelberg: Springer.
- Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler (7., vollständig überarbeitete und erweiterte Auflage)*. Berlin: Springer.
- Brandlmaier, E., Frank-Hermann, P., Korunka, C., Plessnig, A., Schopf, C., & Tamegger, K. (2006). *Ökonomische Bildung von Schüler/innen Allgemeinbildender Höherer Schulen: Modellentwicklung, Entwicklung eines Messinstruments, ausgewählte Ergebnisse*. Wien: WUV.
- Brown, M., & Graf, R. (2013). Financial Literacy and Retirement Planning in Switzerland. *Numeracy*, 6(2), 6. doi:10.5038/1936-4660.6.2.6
- Brückner, S., Förster, M., Zlatkin-Troitschanskaia, O., & Walstad, W. B. (2015). 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.
- Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion (3. aktualisierte und erweiterte Auflage)*. München: Pearson Studium.
- Bünger, C. (2013). *Die offene Frage der Mündigkeit: Studien zur Politizität der Bildung*. Paderborn: Schöningh.
- CAP. (2015). Comparative Agendas Project (CAP) Codebook. Retrieved from <http://www.comparativeagendas.net/codebook> (07.02.2018)

- Capaul, R., & Steingruber, D. (2013). *Betriebswirtschaft verstehen: das St. Galler Management-Modell (2. Auflage)*. Berlin: Cornelsen.
- CED. (1961). Economic education in the schools: A report of the national task force on economic education. New York: Committee for Economic Development (CED).
- CEE. (2010). Voluntary national content standards in economics. New York: Council for Economic Education (CEE).
- CEE. (2013). National Standards for Financial Literacy: Council for Economic Education (CEE).
- Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
- CIIP. (2010). Plan d'études romand (PER): Conférence intercantonale de l'instruction publique de la Suisse romande et du Tessin (CIIP).
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd edition)*. Hillsdale, N.J.: Lawrence Earlbaum Associates.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.
- Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.
- D-EDK. (2014). Lehrplan 21 (Freigegebene Vorlage vom 31.10.2014, bereinigte Fassung vom 29.02.2016): Deutschschweizer Erziehungsdirektoren-Konferenz.
- Dauenhauer, E. (2001). *Kategoriale Wirtschaftsdidaktik, Band 1: Anregungen zur inhaltlichen Neugestaltung (3. Auflage)*. Münchweiler: Walthari-Verlag.
- Davies, P. (2015). Toward a framework for financial literacy in the context of democracy. *Journal of Curriculum Studies*, 47(2), 300-316.
- De Swert, K. (2012). Calculating inter-coder reliability in media content analysis using Krippendorff's Alpha (Manual): University of Amsterdam.
- DEGÖB. (2004). Kompetenzen der ökonomischen Bildung für allgemein bildende Schulen und Bildungsstandards für den mittleren Schulabschluss. *Journal of Social Science Education*, 5(3). Retrieved from http://degoeb.de/uploads/degoeb/04_DEGOEB_Sekundarstufe-I.pdf (26.05.2018)
- DEGÖB. (2006). Kompetenzen der ökonomischen Bildung für allgemein bildende Schulen und Bildungsstandards für den Grundschulabschluss. Retrieved from http://degoeb.de/uploads/degoeb/06_DEGOEB_Grundschole.pdf (26.05.2018)
- DEGÖB. (2009). Kompetenzen der ökonomischen Bildung für allgemein bildende Schulen und Bildungsstandards für den Abschluss der gymnasialen Oberstufe. Retrieved from http://degoeb.de/uploads/degoeb/09_DEGOEB_Abitur.pdf (26.05.2018)
- Dinh, A., Haves, M., & Retzmann, T. (2018). Krisen kompetent bewältigen? Entwicklung eines Kompetenzmodells für die Entrepreneurship Education. In E. Wittmann, D. Frommberger, & B. Ziegler (Eds.), *Jahrbuch der berufs- und wirtschaftspädagogischen Forschung 2018* (pp. 81-94). Opladen: Verlag Barbara Budrich.
- Dörig, R. (2003). *Handlungsorientierter Unterricht*. Stuttgart: WiKu-Verlag.
- Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften (5., vollständig überarbeitete, aktualisierte und erweiterte Auflage)*. Berlin: Springer VS.
- Dörner, D. (1979). *Problemlösen als Informationsverarbeitung*. Stuttgart: Kohlhammer.
- Dörner, D., Kreuzig, H. W., Reither, F., & Ständel, T. (Eds.). (1983). *Lohhausen: Vom Umgang mit Unbestimmtheit und Komplexität*. Bern: Huber.

- Draba, R. E. (1977). The identification and interpretation of item bias: University of Chicago, Educational Statistics Laboratory, Memo 25.
- Draxler, D. (2005). *Aufgabendesign und basismodellorientierter Physikunterricht*. (Unv. Dissertationsschrift), Universität Duisburg-Essen.
- Dubs, R. (1985). *Kleine Unterrichtslehre für den Lernbereich Wirtschaft, Recht, Staat und Gesellschaft*. Aarau: Verlag Sauerländer.
- Dubs, R. (1997). Das Grundlagenfach «Wirtschaft und Recht» an Gymnasien. *Schweizerische Zeitschrift für kaufmännisches Bildungswesen*, 91(2), 101-122.
- Dubs, R. (1998). *Volkswirtschaftslehre: Wirtschaftsbürgerkunde für Mittelschulen und zum Selbststudium (7. aktualisierte Auflage)*. Bern: Haupt.
- Dubs, R. (2001). Wirtschaftsbürgerliche Bildung: Überlegungen zu einem alten Postulat. *sowi-onlinejournal*, 2/2001. Retrieved from <https://www.sowi-online.de/sites/default/files/wirtschaftsbuergerliche-bildung-dubs.pdf>
- Dubs, R. (2004). Bildungsstandards - ein erfolgversprechender Paradigmawechsel? Ein Umsetzungsversuch als Diskussionsgrundlage im Fach Volkswirtschaftslehre. In M. Wosnitza, A. Frey, & R. S. Jäger (Eds.), *Lernprozess, Lernumgebung und Lerndiagnostik. Wissenschaftliche Beiträge zum Lernen im 21. Jahrhundert* (pp. 38-55). Landau: Empirische Pädagogik.
- Dubs, R. (2007). Bildungsstandards für die Fächer Wirtschaft und Recht. In P. Labudde (Ed.), *Bildungsstandards am Gymnasium? Korsett oder Katalysator?* (pp. 305-314). Bern: hep Verlag.
- Dubs, R. (2009). *Lehrerverhalten: Ein Beitrag zur Interaktion von Lehrenden und Lernenden im Unterricht*. Zürich: Verlag SKV.
- Dubs, R. (2011). Die Bedeutung der wirtschaftlichen Bildung in einer Demokratie. In L. Ludwig, H. Luckas, F. Hamburger, & S. Aufenanger (Eds.), *Bildung in der Demokratie II: Tendenzen-Diskurse-Praktiken* (pp. 191-206). Opladen: Verlag Barbara Budrich.
- Dubs, R. (2013). Ökonomische Allgemeinbildung in der Sekundarstufe II. In T. Retzmann (Ed.), *Ökonomische Allgemeinbildung in der Sekundarstufe II: Konzepte, Analysen und empirische Befunde* (pp. 13-25). Schwalbach/Ts.: Wochenschau Verlag.
- Dubs, R. (2014). *Unterrichtsplanung in der Praxis - Ein Handbuch für den Lernbereich Wirtschaft* (1st ed.). Stuttgart: Franz Steiner Verlag.
- Duden. (2019). *Duden Wörterbuch*. Berlin: Verlag Bibliographisches Institut GmbH.
- Eberle, F. (2006a). Anwendungs- und problemorientierter Unterricht (APU): Ein Unterrichtsmodell für Gymnasien. *NetzWerk: die Zeitschrift der Wirtschaftsbildung Schweiz*(3), 20-30.
- Eberle, F. (2006b). Zur Bedeutung von Wirtschaft und Recht in der gymnasialen Bildung. *Gymnasium Helveticum*(3), 16-23.
- Eberle, F. (2007). Das Fundament von Bildungsstandards an Schweizer Gymnasien. In P. Labudde (Ed.), *Bildungsstandards am Gymnasium: Korsett oder Katalysator?* (pp. 37-45). Bern: hep Verlag.
- Eberle, F. (2015). Die Förderung ökonomischer Kompetenzen zwischen normativem Anspruch und empirischer Rationalität: Am Beispiel der Schweizer Sekundarstufe II. *Empirische Pädagogik*, 29(1), 10-34.
- Eberle, F., & Brüggemann, C. (2013). *Bildung am Gymnasium*. Bern: EDK.
- Eberle, F., Gehrler, K., Jaggi, B., Kottonau, J., Oepke, M., & Pflüger, M. (2008). Evaluation der Mauritätsreform 1995 (EVAMAR): Schlussbericht zur Phase II. Bern: Staatssekretariat für Bildung und Forschung SBF.
- Eberle, F., Schumann, S., Ackermann, N., Jüttler, A., & Kaufmann, E. (2017a). *Forschungsinitiative ASCOT, Verbundprojekt CoBALIT: Praxisbericht für die Branche SPEDLOGSWISS*. Unveröffentlicht, Universität Zürich, Institut für Erziehungswissenschaften. Zürich/Konstanz.

- Eberle, F., Schumann, S., Ackermann, N., Jüttler, A., & Kaufmann, E. (2017b). *Forschungsinitiative ASCOT, Verbundprojekt CoBALIT: Praxisbericht für die Branche SWISSMEM*. Unveröffentlicht, Universität Zürich, Institut für Erziehungswissenschaft.
- 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 (Eds.), *Technologiebasierte Kompetenzmessung in der beruflichen Bildung: Ergebnisse aus der BMBF-Förderinitiative ASCOT* (pp. 93-117). Bielefeld: wbv.
- Eberle, F., Schumann, S., Oepke, M., Müller, C., Pflüger, M., Hesske, S., . . . Blum, R. (2009). Der lange Weg zur Verbesserung der Unterrichtsqualität: Ein Rückblick auf die Interventionsstudie APU. *NetzWerk: die Zeitschrift der Wirtschaftsbildung Schweiz*(1), 15-23.
- Edelmann, W., & Wittmann, S. (2012). *Lernpsychologie (7., vollständig überarbeitete Auflage)*. Weinheim: Beltz.
- EDK. (1994). Rahmenlehrplan für die Maturitätsschulen vom 9. Juni 1994. Empfehlung an die Kantone gemäss Art. 3 des Schulkonkordats vom 29. Oktober 1970. Mit Handreichung zur Umsetzung. Bern: Schweizerische Konferenz der kantonalen Erziehungsdirektoren.
- Eid, M., Gollwitzer, M., & Schmitt, M. (2011). *Statistik und Forschungsmethoden (2., korrigierte Auflage)*. Weinheim: Beltz.
- Eid, M., & Schmidt, K. (2014). *Testtheorie und Testkonstruktion*. Göttingen: Hogrefe.
- Embretson, S. E. (1999). Issues in the Measurement of Cognitive Ability. In S. E. Embretson & S. L. Hershberger (Eds.), *The New Rules of Measurement* (pp. 1-15). Mahwah, NJ: Lawrence Erlbaum.
- Engartner, T., & Krisanthan, B. (2013). Ökonomische Bildung im sozialwissenschaftlichen Kontext - oder: Aspekte eines Konzepts sozio-ökonomischer Bildung. *Gesellschaft Wirtschaft Politik (GWP)*, 62(2), 243-256.
- Euler, D., & Hahn, A. (2014). *Wirtschaftsdidaktik (3., aktualisierte Auflage)*. Bern: Haupt Verlag.
- Findeisen, S. (2017). *Fachdidaktische Kompetenzen angehender Lehrpersonen: Eine Untersuchung zum Erklären im Rechnungswesen*. Wiesbaden: Springer.
- Fischer, A., & Zurstrassen, B. (Eds.). (2014). *Sozioökonomische Bildung* (Vol. 1436). Bonn: Bundeszentrale für politische Bildung.
- Fischer, H. E., & Draxler, D. (2001). Aufgaben und Naturwissenschaftlicher Unterricht. *Der mathematische und naturwissenschaftliche Unterricht*, 54(7), 388-393.
- Fischer, H. E., & Draxler, D. (2007). Konstruktion und Bewertung von Physikaufgaben. In E. Kircher, R. Girwidz, & P. Häussler (Eds.), *Physikdidaktik - Theorie und Praxis* (pp. 639-655). Berlin: Springer VS.
- Fischer, H. E., Glemitz, I., Kauertz, A., & Sumfleth, E. (2007). Auf Wissen aufbauen - kumulatives Lernen in Chemie und Physik. In E. Kircher, R. Girwidz, & P. Häussler (Eds.), *Physikdidaktik - Theorie und Praxis* (pp. 657-678). Berlin: Springer VS.
- Förster, M., Brückner, S., Happ, R., Beck, K., & Zlatkin-Troitschanskaia. (2017). Strukturanalyse eines kognitiven Messinstruments im Multiple Choice-Format. *Zeitschrift für Berufs- und Wirtschaftspädagogik*, 113(3), 366-396.
- Förster, M., & Happ, R. (under review). The connection between gender, interest in economic topics and economic knowledge among students of vocational schools. *Citizenship, Social and Economics Education (CSEE), Special Issue "Gender-specific Differences in Economic Competence"*.
- Förster, M., Happ, R., & Molerov, D. (2017). Using the US test of financial literacy in Germany: adaptation and validation. *The Journal of Economic Education*, 48(2), 123-135. doi:10.1080/00220485.2017.1285737

- Förster, M., Happ, R., & Walstad, W. B. (2019). Relations between young adults' knowledge and understanding, experiences, and information behavior in personal finance matters. *Empirical Research in Vocational Education and Training*, 11(2), 1-21. doi:10.1186/s40461-019-0077-z
- Forstmoser, P., & Vogt, H.-U. (2012). *Einführung in das Recht (5., vollständig überarbeitete und stark erweiterte Auflage)*. Bern: Stämpfli Verlag.
- Freundlinger, A. (1992). Wirtschaftskenntnisse von Maturanden. Wien: Schriftenreihe des Instituts für Bildungsforschung und Wirtschaft, Band 88.
- Frey, A. (2012). Adaptives Testen. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion (2., aktualisierte und überarbeitete Auflage)* (pp. 275-293). Heidelberg: Springer.
- Frey, A., Hartig, J., & Rupp, A. A. (2009). Booklet Designs in Large-Scale Assessments of Student Achievement: Theory and Practice. *Educational Measurement: Issues and Practice*, 28(39-53).
- Fromm, S. (2012). *Datenanalyse mit SPSS für Fortgeschrittene 2: Multivariate Verfahren für Querschnittsdaten* (2 ed.). Heidelberg: Springer VS.
- Fromm, S., Akremi, L., & Baur, N. (2010). *Datenanalyse mit SPSS für Fortgeschrittene: Datenaufbereitung und uni- und bivariate Statistik*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Gagné, R. M., & Briggs, L. (1974). *Principles of Instructional Design*. New York, NY: Holt, Rinehart & Winston.
- Ganzeboom, H. B. G. (2010). International Standard Classification of Occupations ISCO-08 with ISEI-08 scores. Retrieved from http://www.harryganzeboom.nl/ISCO08/isco08_with_isei.pdf (20.11.2017)
- Geisinger, K. F. (1992). The metamorphosis of test validation. *Educational Psychologist*, 27, 197-222.
- Gelman, R., & Greeno, J. G. (1989). On the Nature of Competence: Principles for Understanding in a Domain. In L. B. Resnick (Ed.), *Knowing, Learning, and Instruction: Essays in Honor of Robert Glaser* (pp. 125-186). Hillsdale, NJ: Lawrence Erlbaum.
- Gilg, P., Raymond, F., & Bindschedler-Robert, D. (1966). Année politique suisse 1965 = Schweizerische Politik im Jahre 1965. *Annuaire suisse de science politique = Schweizerisches Jahrbuch für Politische Wissenschaft*, 6, 139-210.
- Gollob, R., Graf-Zumsteg, C., Bachmann, B., Gattiker, S., & Ziegler, B. (2007). *Politik und Demokratie – leben und lernen. Politische Bildung in der Schule, Grundlagen für die Aus- und Weiterbildung*. Bern: schulverlag.
- Greeno, J. G., Riley, M. S., & Gelman, R. (1984). Conceptual Competence and Children's Counting. *Cognitive Psychology*, 16, 94-143.
- Guggemos, J., & Schönlein, M. (2015). Modellierung von Kompetenzen in der beruflichen Bildung: Entwicklung und Validierung eines Kompetenzniveaumodells für das externe Rechnungswesen. *Zeitschrift für Berufs- und Wirtschaftspädagogik*, 111(4), 524-551.
- Gwet, K. L. (2014). *Handbook of Inter-Rater Reliability (4th edition)*. Gaithersburg: Advanced Analytics, LLC.
- Hansen, W. L., Bach, G. L., Calderwood, J. D., & Saunders, P. (1977). Master Curriculum Grade in Economics for the Nation's Schools: Part I: A framework for teaching economics: Basic concepts. New York: Joint Council of Economic Education (JCEE).
- Hartig, J. (2004). Methoden zur Bildung von Kompetenzstufenmodellen. In H. Moosbrugger, W. Rauch, & D. Frank (Eds.), *Qualitätssicherung im Bildungswesen*. Frankfurt a.M.
- Hartig, J. (2007). Skalierung und Definition von Kompetenzniveaus. In B. Beck & E. Klieme (Eds.), *Sprachliche Kompetenzen: Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)* (pp. 83-99). Weinheim: Beltz Verlag.

- Hartig, J. (2008). Kompetenzen als Ergebnisse von Bildungsprozessen. In N. Jude, J. Hartig, & E. Klieme (Eds.), *Kompetenzerfassung in pädagogischen Handlungsfeldern: Theorien, Konzepte und Methoden* (pp. 15-25). Berlin: Bundesministerium für Bildung und Forschung (BMBF).
- Hartig, J., Frey, A., & Jude, N. (2012). Validität. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion (2., aktualisierte und überarbeitete Auflage)* (pp. 143-171). Heidelberg: Springer.
- Hartig, J., & Jude, N. (2007). Empirische Erfassung von Kompetenzen und psychometrische Kompetenzmodelle. In J. Hartig & E. Klieme (Eds.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik: Eine Expertise im Auftrag des Bundesministeriums für Bildung und Forschung* (Vol. 20, pp. 17-36). Berlin: Bundesministerium für Bildung und Forschung (BMBF).
- Hartig, J., & Klieme, E. (2006). Kompetenz und Kompetenzdiagnostik. In K. Schweizer (Ed.), *Leistung und Leistungsdiagnostik* (pp. 127-143). Berlin: Springer VS.
- Hausheer, H., & Aepli-Müller, R. E. (2016). *Das Personenrecht des Schweizerischen Zivilgesetzbuches (4. Auflage)*. Bern: Stämpfli Verlag.
- Hayes, A. F., & Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, 1(1), 77-89. doi:10.1080/19312450709336664
- Hedtke, R. (2005). Gemeinsam und unterschieden: Zum Problem der Integration von politischer und ökonomischer Bildung. In D. Kahsnitz (Ed.), *Integration von politischer und ökonomischer Bildung?* (pp. 19-74). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Hedtke, R. (2006). Sozialwissenschaftliche ökonomische Bildung. In A. Fischer (Ed.), *Ökonomische Bildung - quo vadis?* (pp. 95-119). Bielefeld: Bertelsmann.
- Hedtke, R. (2014). Was ist sozio-ökonomische Bildung. In A. Fischer & B. Zurstrassen (Eds.), *Sozioökonomische Bildung* (pp. 81-127). Bonn: Bundeszentrale für politische Bildung.
- Hedtke, R. (2015). *Was ist und wozu Sozioökonomie?* Wiesbaden: Springer VS.
- Hedtke, R. (2019). *Das sozioökonomische Curriculum*. Frankfurt a.M.: Wochenschau Verlag.
- Holtsch, D. (2018). Zur professionellen Kompetenz von Lehrpersonen (Teilprojekt 3). In D. Holtsch & F. Eberle (Eds.), *Untersuchungen zu Lehr-Lernprozessen im kaufmännischen Bereich: Ergebnisse aus dem Leading House LINCA und Schlussfolgerungen für die Praxis* (pp. 129-158). Münster: Waxmann.
- Holtsch, D., & Eberle, F. (2018). *Untersuchungen zu Lehr-Lernprozessen im kaufmännischen Bereich: Ergebnisse aus dem Leading House LINCA und Schlussfolgerungen für die Praxis*. Münster: Waxmann.
- Hopkins, K. D., & Weeks, D. L. (1990). Tests for normality and measures of skewness and kurtosis: Their place in research reporting. *Educational and Psychological Measurement*, 50, 717-729.
- IBM. (2017). SPSS Statistics (Version 25): IBM Corporation.
- ILO. (2017). International Standard Classification of Occupations (ISCO): International Labour Organisation (ILO).
- Jonkisz, E., Moosbrugger, H., & Brandt, H. (2012). Planung und Entwicklung von Tests und Fragebogen. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion (2., aktualisierte und überarbeitete Auflage)* (pp. 27-74). Heidelberg: Springer.
- Jüttler, A., & Schumann, S. (2016). Effects of Students Socio-cultural Background on Economic Competencies at the End of Upper Secondary Education. In E. Wuttke, J. Seifried, & S. Schumann (Eds.), *Economic Competence and Financial Literacy of Young Adults: Status and Challenges* (pp. 121-148). Opladen: Verlag Barbara Budrich.
- Kamininski, H. (2017). *Fachdidaktik der ökonomischen Bildung*. Paderborn: Ferdinand Schöningh.

- Kaminski, H. (1996). *Ökonomische Bildung und Gymnasium: Ziele, Inhalte, Lernkonzepte des Ökonomieunterrichts*. Neuwied: Luchterhand Verlag.
- Kaminski, H., Brettschneider, V., & Schnell, C. (2017). *Fachdidaktik der ökonomischen Bildung*. Paderborn: Ferdinand Schöningh.
- Kaminski, H., Eggert, K., & Burkard, K.-J. (2008). *Konzeption für die ökonomische Bildung als Allgemeinbildung von der Primarstufe bis zur Sekundarstufe II*. Berlin: Bundesverband der deutschen Banken.
- Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin*, 112, 527-535.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50, 1-73.
- Katschnig, T., & Hanisch, G. (2005). Wirtschaftswissen von Maturanden im internationalen Vergleich: Eine empirische Studie in den Ländern Österreich, Deutschland, Tschechien und Ungarn. Wien: Institut für Erziehungswissenschaften (Universität Wien) und Ludwig-Boltzmann-Institut für Schulentwicklung und international-vergleichende Schulforschung.
- Kauertz, A. (2007). *Schwierigkeitserzeugende Merkmale physikalischer Leistungstestaufgaben*. (Unv. Dissertationsschrift), Universität Duisburg-Essen.
- Kauertz, A., Löffler, P., & Fischer, H. E. (2015). Physikaufgaben. In E. Kircher, R. Girwidz, & P. Häussler (Eds.), *Physikdidaktik - Theorie und Praxis (3. Auflage)* (pp. 451-475). Berlin, Heidelberg: Springer.
- Kavadarli, B. (2019). *Argumentation zu sozioökonomischen Problemsituationen: Analyse von schriftlichen Antworten zu Testaufgaben von Gymnasiastinnen und Gymnasiasten*. Masterarbeit. Unveröffentlicht, Universität Zürich, Institut für Erziehungswissenschaft.
- Kelava, A., & Moosbrugger, H. (2012). Deskriptivstatistische Evaluation von Items (Itemanalyse) und Testwertverteilungen. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion (2. aktualisierte und überarbeitete Auflage)* (pp. 75-102). Berlin: Springer VS.
- Klafki, W. (1985). *Neue Studien zur Bildungstheorie und Didaktik*. Weinheim: Beltz.
- Klafki, W. (1998). Schlüsselprobleme der modernen Welt und die Aufgaben der Schule: Grundlinien einer neuen Allgemeinbildungskonzeption in internationaler/interkultureller Perspektive. In I. Gogolin, M. Krüger-Potratz, & M. A. Meyer (Eds.), *Pluralität und Bildung* (pp. 235-249). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Klieme, E. (2000). Fachleistungen im voruniversitären Mathematik- und Physikunterricht: Theoretische Grundlagen, Kompetenzstufen und Unterrichtsschwerpunkte. In J. Baumert, W. Bos, & R. Lehmann (Eds.), *TIMSS III: Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn* (Vol. 2, pp. 57-128). Opladen: Leske & Budrich.
- Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., . . . Vollmar, H. J. (2007). *Zur Entwicklung nationaler Bildungsstandards: Eine Expertise*. Berlin: Bundesministerium für Bildung und Forschung (BMBF).
- Klieme, E., Maag Merki, K., & Hartig, J. (2007). Kompetenzbegriff und Bedeutung von Kompetenzen im Bildungswesen. In J. Hartig & E. Klieme (Eds.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik: Eine Expertise im Auftrag des Bundesministeriums für Bildung und Forschung* (pp. 5-15). Berlin: Bundesministerium für Bildung und Forschung (BMBF).
- Krapp, A. (1992). Interesse, Lernen und Leistung: Neue Forschungsansätze in der Pädagogischen Psychologie. *Zeitschrift für Pädagogik*, 38(5), 747-770.
- Krathwohl, D. R. (2002). A Revision of Bloom's Taxonomy: An Overview. *Theory into Practice*, 41(4), 212-218.

- Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30, 61-70.
- Kruber, K.-P. (1995). Wirtschaftspolitisches Denken lernen an und in Modellen. In H.-J. Albers (Ed.), *Handlungsorientierung in der ökonomischen Bildung* (Vol. 15, pp. 93-115). Bergisch Gladbach: Verlag Thomas Hobein.
- Kruber, K.-P. (2000). Kategoriale Wirtschaftsdidaktik: der Zugang zur ökonomischen Bildung. *Gegenwartskunde*, 49(3), 285-295.
- Kruber, K.-P. (2005). Ökonomische und politische Bildung - der mehrperspektivische Zugriff auf Wirtschaft und Politik. In D. Kahsnitz (Ed.), *Integration von politischer und ökonomischer Bildung?* (pp. 75-109). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Kuckartz, U., Rädiker, S., Ebert, T., & Schehl, J. (2013). *Statistik: Eine verständliche Einführung* (2., überarbeitete Auflage). Wiesbaden: Springer VS.
- Kuhn, C. (2014). *Fachdidaktisches Wissen von Lehrkräften im kaufmännisch-verwaltenden Bereich. Modellbasierte Testentwicklung und Validierung* (Vol. 2). Landau: Verlag Empirische Pädagogik.
- Kuhn, C., Happ, R., Zlatkin-Troitschanskaia, O., Beck, K., Förster, M., & Preuße, D. (2014). Kompetenzentwicklung angehender Lehrkräfte im kaufmännisch-verwaltenden Bereich – Erfassung und Zusammenhänge von Fachwissen und fachdidaktischem Wissen. *Zeitschrift für Erziehungswissenschaft*, 17(1), 149–167.
- Kutscha, G. (2014). Ökonomie an Gymnasien unter dem Anspruch des Bildungsprinzips: Diskursgeschichtlicher Rückblick und Zielperspektiven für die sozio-ökonomische Bildung. In A. Fischer & B. Zurstrassen (Eds.), *Sozioökonomische Bildung* (pp. 65–80). Bonn: Bundeszentrale für politische Bildung.
- Langeheine, R., & Rost, J. (Eds.). (1988). *Latent trait and latent class models*. New York: Plenum.
- Le, L. (2012). ACER ConQuest Note 5: Item point-biserial discrimination. Retrieved from <https://www.acer.org/files/Conquest-Notes-5-ItemPointBiserialDiscrimination.pdf> (08.07.2018)
- Lehmann, R., & Seeber, S. (2007). *ULME III. Untersuchung von Leistungen, Motivation und Einstellungen der Schülerinnen und Schüler in den Abschlussklassen der Berufsschulen*. Hamburg: Behörde für Bildung und Sport.
- Lenhard, W., & Lenhard, A. (2014). Signifikanztests bei Korrelationen. Retrieved from <https://www.psychometrica.de/korrelation.html> (30.06.2018)
- Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Auflage). Weinheim: Beltz.
- Lusardi, A., & Mitchell, O. S. (2014). The economic importance of financial literacy: Theory and evidence. *Journal of Economic Literature*, 52(1), 5-44.
- Maag Merki, K. (2007). Bildungsstandards - Konzept und Begrifflichkeiten. In P. Labudde (Ed.), *Bildungsstandards am Gymnasium: Korsett oder Katalysator?* (pp. 17-25). Bern: hep Verlag.
- Maag Merki, K. (2009). Kompetenz. In S. Andresen, R. Casale, T. Gabriel, R. Horlacher, S. L. Klee, & J. Oelkers (Eds.), *Handwörterbuch Erziehungswissenschaft* (pp. 492-506). Weinheim, Basel: Beltz Verlag.
- Macaskill, G., & Adams, R. J. (2016). ACER ConQuest Note 8: Score Estimation and Generalised Partial Credit Models. Retrieved from https://www.acer.org/files/Note_8--The_ConQuest_4_Model.pdf (31.08.2018)
- Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The package eRm for the application of IRT models in R. *Journal of Statistical Software*, 20(9).
- Mair, P., & Hatzinger, R. (2009). erm: Extended Rasch modeling (R package version 0.11-0).
- Marzano, R. J., & Kendall, J. S. (2007). *The new taxonomy of educational objectives* (2nd edition). Thousand Oaks, CA: Corwin Press.

- Massing, P. (1995). Wege zum Politischen. In P. Massing & G. Weisseno (Eds.), *Politik als Kern der politischen Bildung* (pp. 61-98). Opladen: Verlag Barbara Budrich.
- Massing, P. (1999). Politikzyklus. In D. Richter & G. Weisseno (Eds.), *Lexikon der politischen Bildung, Bd. 1: Didaktik und Schule*. Schwalbach/Ts.: Wochenschau Verlag.
- Masters, G. N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika*, 47(2), 149-174.
- May, H. (1978). *Arbeitslehre: Wirtschaftswissenschaftliche und wirtschaftsdidaktische Grundlagen*. München: Reinhardt Verlag.
- May, H. (2011). Ökonomische Bildung als Allgemeinbildung. *Aus Politik und Zeitgeschichte (APuZ)*(12), 3-9.
- May, H., & Albers, H.-J. (Eds.). (2008). *Handbuch zur ökonomischen Bildung (9., völlig überarbeitete und aktualisierte Auflage)*. München: Oldenbourg.
- Mayring, P. (2015). *Qualitative Inhaltsanalyse (12. Auflage)*. Weinheim: Beltz Verlag.
- McClelland, D. C. (1973). Testing for Competence rather than for «Intelligence». *American Psychologist*, 28(1), 1-14.
- Mentele, S., Heinzer, S., Lekic, C., Lenggenhager, M., Holtsch, D., & Eberle, F. (2014). *Entwicklung eines computerbasierten Instruments LINCA zur Erfassung des kaufmännischen Wissens und Könnens von Lernenden in der deutschsprachigen Schweiz*. Unveröffentlicht, Universität Zürich, Institut für Erziehungswissenschaft.
- Mertens, D. (1974). Schlüsselqualifikationen: Thesen zur Schulung für eine moderne Gesellschaft. *Mitteilungen aus der Arbeitsmarkt- und Berufsforschung*, 7(1), 36-43.
- Messick, S. J. (1980). Test validity and the ethics of assessment. *American Psychologist*, 35(11), 1012-1027.
- Messick, S. J. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & I. Braun (Eds.), *Test validity* (Vol. Lawrence Erlbaum Associates, Inc., pp. 33-48): Hillsdale, NJ, England.
- Messick, S. J. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). Washinton, DC: American Council on Education and National Council on Measurement in Education.
- Messick, S. J. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749. doi:10.1037/0003-066X.50.9.741
- Metzger, C., Waibel, R., Henning, C., Hodel, M., & Luzi, R. (1993). *Anspruchsniveau von Lernzielen und Prüfungen im kognitiven Bereich: Inhaltsniveaus, Prozessionsniveaus, Reproduktion und Transfer*. St. Gallen Universität St. Gallen, Institut für Wirtschaftspädagogik.
- Michaelis, C. (2017). *Kompetenzentwicklung zum nachhaltigen Wirtschaften: Eine Längsschnittstudie in der kaufmännischen Ausbildung* (Vol. 21). Frankfurt a.M.: Peter Lang.
- Moosbrugger, H. (2012a). Item-Response-Theorie (IRT). In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion (2., aktualisierte und überarbeitete Auflage)* (pp. 227-274). Heidelberg: Springer.
- Moosbrugger, H. (2012b). Klassische Testtheorie (KTT). In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion (2., aktualisierte und überarbeitete Auflage)* (pp. 103-117). Heidelberg: Springer.
- Moosbrugger, H., & Höfling, V. (2012). Standards für psychologisches Testen. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion (2., aktualisierte und überarbeitete Auflage)* (pp. 203-224). Heidelberg: Springer.
- Moosbrugger, H., & Kelava, A. (2012a). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion (2. aktualisierte und überarbeitete Auflage)* (pp. 7-26). Berlin: Springer VS.

- Moosbrugger, H., & Kelava, A. (Eds.). (2012b). *Testtheorie und Fragebogenkonstruktion (2. aktualisierte und überarbeitete Auflage)*. Berlin: Springer VS.
- Moosbrugger, H., & Schermelleh-Engel, K. (2012). Exploratorische (EFA) und Konfirmatorische Faktorenanalyse (CFA). In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion (2. aktualisierte und überarbeitete Auflage)* (pp. 325-343). Berlin: Springer VS.
- Morris, M. H. (2013). A Competency-Based Perspective on Entrepreneurship Education: Conceptual and Empirical Insights. *Small Business Management*, 5(3), 352-369.
- Müller, K., Fürstenau, B., & Witt, R. (2007). Ökonomische Kompetenz sächsischer Gymnasiasten. *Zeitschrift für Berufs- und Wirtschaftspädagogik*, 103(2), 227-247.
- Newton, P., & Shaw, S. (2013). Standards for talking and thinking about validity. *Psychological Methods*, 18(3), 301-319.
- Nickolaus, R., Gschwendtner, T., & Geißel, B. (2008). Entwicklung und Modellierung beruflicher Fachkompetenz in der gewerblich-technischen Grundbildung. *Zeitschrift für Berufs- und Wirtschaftspädagogik*, 104(1), 48-73.
- Norris, N. (1991). The Trouble with Competence. *Cambridge Journal of Education*, 21(3), 330-341.
- Ochs, D., & Steinmann, B. (1978). Beitrag der Ökonomie zu einem sozialwissenschaftlichen Curriculum. In E. Forndran, H. J. Hummel, & H. Süßmuth (Eds.), *Studiengang Sozialwissenschaften: Zur Definition eines Faches* (pp. 186-223). Düsseldorf: Schwann-Bagel.
- OECD. (2005a). Improving Financial Literacy: Analysis of Issues and Policies. Paris: OECD Publishing.
- OECD. (2005b). PISA 2003 Technical Report. Retrieved from <http://www.oecd.org/education/school/programmeforminternationalstudentassessmentpisa/35188570.pdf> (31.08.2018)
- OECD. (2005c). Recommendation on Principles and Good practices for Financial Education and Awareness: Recommendation of the Council. Retrieved from <http://www.oecd.org/finance/financial-education/35108560.pdf> (19.07.2018)
- OECD. (2012). PISA 2012 Financial Literacy Assessment Framework. Retrieved from <https://www.oecd.org/pisa/pisaproducts/46962580.pdf> (10.07.2018)
- OECD. (2013). PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy. Retrieved from <https://doi.org/10.1787/9789264190511-en> (10.07.2018)
- OECD. (2014a). PISA 2012 Results: Students and Money (Volume VI): Financial Literacy Skills for the 21st Century. Retrieved from <https://doi.org/10.1787/9789264208094-en> (10.07.2018)
- OECD. (2014b). PISA 2012 Technical Report. Retrieved from <https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf> (31.08.2018)
- OECD. (2017a). PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving (revised edition). Paris: OECD Publishing.
- OECD. (2017b). PISA 2015 Financial Literacy Framework. In OECD (Ed.), *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving* (pp. 81-102). Paris: OECD Publishing.
- Oser, F. (1986). Moral Education and Values Education: The Discourse Perspective. In M. C. Wittrock (Ed.), *Handbook of Research on Teaching* (pp. 917-941). New York: Macmillan.
- Oser, F., & Volery, T. (2012). 'Sense of failure' and 'sense of success' among entrepreneurs: The identification and promotion of neglected twin entrepreneurial competencies. *Empirical Research in Vocational Education and Training*, 4(1), 27-44.
- Paek, I., & Wilson, M. R. (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.
- Peng, P. (2018). *Die Bedeutung schulischer und ausserschulischer Kontextfaktoren für die wirtschaftsbürgerliche Kompetenz Deutschschweizer Gymnasiastinnen und Gymnasiasten*. Masterarbeit. Unveröffentlicht, Universität Zürich, Institut für Erziehungswissenschaft.
- Pinto, L. E. (2013). When politics trump evidence: financial literacy education narratives following the global financial crisis. *Journal of Education Policy*, 28(1), 95-120.
- Prenzel, M., Krapp, A., & Schiefele, H. (1986). Grundzüge einer pädagogischen Interessentheorie. *Zeitschrift für Pädagogik*, 32, 163-173.
- R-Team. (2010). R: A language and environment for statistical computing. Vienna, Austria: R Development Core Team.
- Rasch, G. W. (1960). Probabilistic models for some intelligence and attainment tests. Kopenhagen: The Danish Institute for Educational Research.
- Reetz, L. (1984). *Wirtschaftsdidaktik: eine Einführung in Theorie und Praxis wirtschaftsberuflicher Curriculumentwicklung und Unterrichtsgestaltung*. Bad Heilbrunn/Obb.: Klinkhardt.
- Reetz, L. (1990). Zur Bedeutung der Schlüsselqualifikationen in der Berufsbildung. In L. Reetz & T. Reitmann (Eds.), *Schlüsselqualifikationen* (pp. 16-35). Hamburg: Feldhaus.
- Reetz, L. (1999). Zum Zusammenhang von Schlüsselqualifikationen – Kompetenzen – Bildung. In T. Tramm, D. Sembill, F. Klauser, & E. G. John (Eds.), *Professionalisierung kaufmännischer Berufsbildung* (pp. 32-51). Frankfurt: Peter Lang.
- Reetz, L. (2003). Prinzipien der Ermittlung, Auswahl und Begründung relevanter Lernziele und Inhalte. In F.-J. Kaiser & H. Kaminski (Eds.), *Wirtschaftsdidaktik* (pp. 99-124). Bad Heilbrunn/Obb.: Klinkhardt.
- Retzmann, T. (2005). Nationale Bildungsstandards für ökonomische Bildung: Theoretische Grundlagen und offene Forschungsfragen. In B. O. Weitz (Ed.), *Standards in der ökonomischen Bildung* (pp. 51-72). Bergisch Gladbach.
- Retzmann, T., & Hausmann, V. (2012). Wie lässt sich unternehmerisches Denken messen? Überlegungen zur Konstruktion eines standardisierten Tests. In T. Retzmann (Ed.), *Entrepreneurship und Arbeitnehmerorientierung: Leitbilder und Konzepte für die ökonomische Bildung in der Schule* (pp. 50-65). Schwalbach/Ts.: Wochenschau Verlag.
- Retzmann, T., Seeber, G., Remmele, B., & Jongebloed, H.-C. (2010). Ökonomische Bildung an allgemeinbildenden Schulen. Bildungsstandards, Standards für die Lehrerbildung: Studie im Auftrag des Gemeinschaftsausschlusses der deutschen gewerblichen Wirtschaft. Berlin.
- Riklin, B. (in progress). *Betriebswirtschaftliche Bildung und Entrepreneurship Education am Gymnasium*. Dissertation. Universität Zürich, Institut für Erziehungswissenschaft.
- Robinson, S. B. (1967). *Bildungsreform als Revision des Curriculum*. Neuwied: Luchterhand Verlag.
- Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, 24(2), 3-13.
- Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion, 2. überarbeitete und erweiterte Auflage*. Bern: Huber.
- Roth, H. (1971). *Pädagogische Anthropologie. Band 2: Entwicklung und Erziehung. Grundlagen einer Entwicklungspädagogik*. Hannover: Schroedel.
- Rudeloff, M. (2019). *Der Einfluss informeller Lerngelegenheiten auf die Finanzkompetenz von Lernenden am Ende der Sekundarstufe I* (Vol. Springer): Wiesbaden.
- Rudeloff, M., Brahm, T., & Pumptow, M. (under review). Does gender matter for the use of learning opportunities? Potential explanation for the gender gap in financial literacy. *Citizenship, Social and Economics Education (CSEE), Special Issue "Gender-specific Differences in Economic Competence"*.

- Saunders, P., & Gilliard, J. (1995). A framework for teaching basic economic concepts: With scope and sequence guidelines, K-12. New York: National Council on Economic Education (NCEE).
- SBFI. (2006). Berufliche Grundbildung: Rahmenlehrplan für den allgemeinbildenden Unterricht vom 13.12.2006. Bern: Staatssekretariat für Bildung, Forschung und Innovation.
- SBFI. (2012). Rahmenlehrplan für die Berufsmaturität vom 18.12.2012. Bern: Staatssekretariat für Bildung, Forschung und Innovation.
- Schermelleh-Engel, K., & Schweizer, K. (2012). Multitrait-Multimethod-Analysen. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion (2. aktualisierte und überarbeitete Auflage)* (pp. 345-362). Berlin: Springer VS.
- Schermelleh-Engel, K., & Werner, C. S. (2012). Methoden der Reliabilitätsbestimmung. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion (2., aktualisierte und überarbeitete Auflage)* (pp. 119-142). Springer: Heidelberg.
- Schiefele, U., Krapp, A., & Schreyer, I. (1993). Metaanalyse des Zusammenhangs von Interesse und schulischer Leistung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 10, 120-148.
- Schumann, S., & Eberle, F. (2011). Bedeutung und Verwendung schwierigkeitsbestimmender Aufgabenmerkmale für die Erfassung ökonomischer und beruflicher Kompetenzen. In U. Fasshauer, B. Fürstenau, & E. Wuttke (Eds.), *Grundlagenforschung zum Dualen System und Kompetenzentwicklung in der Lehrerbildung* (pp. 77-89). Opladen: Barbara Budrich.
- Schumann, S., & Eberle, F. (2014a). Ökonomische Kompetenzen von Lernenden am Ende der Sekundarstufe II. *Zeitschrift für Erziehungswissenschaften*, 17(1), 103-126. doi:10.1007/s11618-013-0459-0
- Schumann, S., & Eberle, F. (2014b). Wirtschafts-, Mathematik- und Deutschkenntnisse Deutschschweizer Lernender am Ende der Berufsmaturität und des Gymnasiums. In F. Eberle, B. Schneider-Taylor, & D. Bosse (Eds.), *Abitur und Matura zwischen Hochschulvorbereitung und Berufsorientierung* (pp. 215-229). Wiesbaden: Springer VS.
- Schumann, S., Eberle, F., & Oepke, M. (2013). Ökonomisches Wissen und Können am Ende der Sekundarstufe II: Effekte der Bildungsgang-, Klassen- und Geschlechtszugehörigkeit. In U. Fasshauer, B. Fürstenau, & E. Wuttke (Eds.), *Jahrbuch der berufs- und wirtschaftspädagogischen Forschung 2013* (pp. 35-46). Opladen: Verlag Barbara Budrich.
- Schumann, S., Eberle, F., Oepke, M., Pflüger, M., Gruber, C., & Pezzotta, D. (2010). *Inhaltsauswahl für den Test zur Erfassung ökonomischen Wissens und Könnens im Projekt «Ökonomische Kompetenzen von Maturandinnen und Maturanden (OEKOMA)»*. Universität Zürich, Institut für Erziehungswissenschaft (ehem. Institut für Gymnasial- und Berufspädagogik).
- 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*, 9:13, 1-21. doi:10.1186/s40461-017-0056-1
- Schumann, S., Oepke, M., & Eberle, F. (2011). Über welche ökonomischen Kompetenzen verfügen Maturandinnen und Maturanden? Hintergrund, Fragestellungen, Design und Methode des Schweizer Forschungsprojekts OEKOMA im Überblick. In U. Fasshauer, J. Aff, B. Fürstenau, & E. Wuttke (Eds.), *Lehr-Lernforschung und Professionalisierung: Perspektiven der Berufsbildungsforschung* (Vol. 31, pp. 51-63). Opladen: Verlag Barbara Budrich.
- Schürkmann, S. (2017). *FILS: Financial Literacy Study - Validierung und Analyse einer schülerorientierten Financial Literacy*. Berlin: De Gruyter Oldenbourg.
- Schürkmann, S., & Schuhen, M. (2013). Kompetenzmessung im Bereich financial literacy: Ergebnisse zum Umgang mit Online-Rechnern aus der FILS-Studie. *Zeitschrift für ökonomische Bildung (ZföB)*, 1, 73-89. doi:10.7808/zfoeb.1.1.64

- Seeber, G., Retzmann, T., Remmele, B., & Jongebloed, H.-C. (2012). *Bildungsstandards der ökonomischen Allgemeinbildung: Kompetenzmodell, Aufgaben, Handlungsempfehlungen*. Schwalbach/Ts.: Wochenschau Verlag.
- Seeber, S., Schumann, S., & Nickolaus, R. (2015). Ökonomische Kompetenzen: Konzeptuelle Grundlagen und empirische Befunde. In G. Weissenö & C. Schelle (Eds.), *Empirische Forschung in gesellschaftswissenschaftlichen Fachdidaktiken – Ergebnisse und Perspektiven* (pp. 169-184). Wiesbaden: Springer VS.
- Shavelson, R. (2012). *Assessing College Learning: The Collegiate Learning Assessment*. Paper presented at the Herbsttagung der DGfE-Sektion Berufs- und Wirtschaftspädagogik, Paderborn. Keynote retrieved from
- Siegfried, C., & Ackermann, N. (in progress). Effekte einer problemorientierten Intervention zur Ökosteuer auf die wirtschaftsbürgerliche Kompetenz von Wirtschaftsgymnasiast*innen im Bundesland Hessen (Arbeitstitel).
- Siegfried, C., & Ackermann, N. (under review). Die Bedeutung schulischer Lerngelegenheiten für die wirtschaftsbürgerliche Kompetenz: Ein Vergleich von Gymnasialschülerinnen und -schülern in der Schweiz und in Deutschland. *Zeitschrift für ökonomische Bildung (ZföB)*.
- SKKAB. (2011a). Bildungsplan Kauffrau/Kaufmann EFZ vom 26. September 2011 für die betrieblich organisierte Grundbildung (Stand am 1. Juni 2018): Staatssekretariat für Bildung, Forschung und Innovation; Schweizerische Konferenz der kaufmännischen Ausbildungs- und Prüfungsbranchen.
- SKKAB. (2011b). Leistungszielkatalog Wirtschaft und Gesellschaft W&G (E-Profil) vom 26. September 2011 (Stand am 1. Januar 2017): Staatssekretariat für Bildung, Forschung und Innovation; Schweizerische Konferenz der kaufmännischen Ausbildungs- und Prüfungsbranchen.
- SKKAB. (2018). Kaufleute 2022: Fit für die digitale Arbeitswelt. Bern, 13. Februar 2018 [Press release]
- Soper, J. C. (1979). *The Test of Economic Literacy: Discussion guide and rationale*. New York: Joint Council of Economic Education (JCEE).
- Soper, J. C., & Walstad, W. B. (1987). *Test of Economic Literacy (2nd edition)*. New York: Joint Council on Economic Education (JCEE).
- St.Gallen. (2006). *Lehrplan für das Gymnasium im Kanton St. Gallen (Lehrplan 2008/10), in Vollzug ab 01.08.2006*. St.Gallen: Erziehungsdirektion des Kantons St.Gallen.
- St.Gallen. (2018). *Mittelschulen des Kantons St.Gallen 2018/2019*. St.Gallen: Bildungsdepartement des Kantons St.Gallen Retrieved from https://www.schule.sg.ch/home/.../mittelschulen_kanton_st-gallen_2018-19.pdf (15.04.2019).
- Steinmann, B. (1995). Handlungsorientierte Methoden als unverzichtbares Element ökonomischer Bildung. In H.-J. Albers (Ed.), *Handlungsorientierung und ökonomische Bildung* (Vol. 15, pp. 151-167). Bergisch Gladbach: Verlag Thomas Hobein.
- Steinmann, B. (Ed.) (1997). *Das Konzept ‚Qualifizierung für Lebenssituationen‘ im Rahmen der ökonomischen Bildung heute*. Bergisch Gladbach: Verlag Thomas Hobein.
- Sticca, F., Rohr-Mentele, S., & Forster-Heinzer, S. (2018). Kompetenzentwicklung der kaufmännischen Lernenden. In D. Holtsch & F. Eberle (Eds.), *Untersuchungen zu Lehr-Lernprozessen im kaufmännischen Bereich: Ergebnisse aus dem Leading House LINCA und Schlussfolgerungen für die Praxis* (pp. 59-87). Münster: Waxmann.
- Strobl, C. (2010). *Das Rasch-Modell: Eine verständliche Einführung für Studium und Praxis* (Vol. 2). München: Rainer Hampp Verlag.
- Sun, X. (2012). ACER ConQuest Note 7: Using Mantel-Haenszel Statistics. Retrieved from <https://www.acer.org/files/Conquest-Notes-7-UsingMantel-HaenszelStatistics.pdf> (08.07.2018)

- Tafner, G. (2016). Die Unterscheidung von Ökonomie und Ökonomik als die Crux der Ökonomischen Bildung. In H. Arndt (Ed.), *Das Theorie-Praxis-Verhältnis in der ökonomischen Bildung* (pp. 30-42). Schwalbach/Ts.: Wochenschau Verlag.
- Tafner, G. (2018). Ökonomische Bildung ist sozioökonomische Bildung: Grundlagen der Didaktik einer reflexiven Wirtschaftspädagogik. In T. Engartner, C. Fridrich, S. Graupe, R. Hedtke, & G. Tafner (Eds.), *Sozioökonomische Bildung und Wissenschaft: Entwicklungslinien und Perspektiven* (pp. 109-140). Wiesbaden: Springer VS.
- Thissen, D. (1996). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, 47(2), 175-186. doi:<https://doi.org/10.1007/BF02296273>
- Tramm, T., & Reetz, L. (2010). Berufliche Curriculumsentwicklung zwischen Persönlichkeits-, Situations- und Wissenschaftsbezug. In R. Nickolaus, G. Pätzold, H. Reinisch, & T. Tramm (Eds.), *Handbuch Berufs- und Wirtschaftspädagogik* (pp. 220-226). Bad Heilbrunn: Verlag Julius Klinkhardt.
- Tschannen, P. (2016). *Staatsrecht der Schweizerischen Eidgenossenschaft (4. Auflage)*. Bern: Stämpfli Verlag.
- Ulrich, P. (1993). Wirtschaftsethik als Beitrag zur Bildung mündiger Wirtschaftsbürger. *Ethica - Wissenschaft und Verantwortung*, 1, 227-250.
- Ulrich, P. (2001). Wirtschaftsbürgerkunde als Orientierung im politisch-ökonomischen Denken. *sowi-onlinejournal*, 2. Retrieved from <https://www.sowi-online.de/sites/default/files/wirtschaftsbuergerkunde-ulrich.pdf> (11.07.2018)
- Wagenschein, M. (1973). *Verstehen lernen: Genetisch-Sokratisch-Exemplarisch*. Weinheim: Beltz.
- Walstad, W. B. (2005). *Economic Education in U.S. High Schools*. Manuscript.
- Walstad, W. B., & Rebeck, K. (2001). Test of Economic Literacy (3rd edition). New York: National Council on Economic Education (NCEE).
- Walstad, W. B., & Rebeck, K. (2016). Test of Financial Literacy: Examiner's Manual. New York: Council for Economic Education.
- Walstad, W. B., & Rebeck, K. (2017). The Test of Financial Literacy: Development and Measurement Characteristics. *The Journal of Economic Education*, 48(2), 113-122. doi:10.1080/00220485.2017.1285739
- Walstad, W. B., Rebeck, K., & Butters, R. B. (2013a). The Test of Economic Literacy: Development and Results. *The Journal of Economic Education*, 44(3), 298-309.
- Walstad, W. B., Rebeck, K., & Butters, R. B. (2013b). Test of Economic Literacy: Examiner's Manual (4th edition): Council for Economic Education (CEE).
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response models. *Psychometrika*, 54(3), 427-450. doi:<https://doi.org/10.1007/BF02294627>
- Weber, A. (2005). Problem-Based Learning – Ansatz zur Verknüpfung von Theorie und Praxis. *Beträge zur Lehrerbildung*, 23(1), 94-104.
- Weber, A. (2007). *Problem-Based Learning: Ein Handbuch für die Ausbildung auf der Sekundarstufe II und der Tertiärstufe (2., überarbeitete Auflage)*. Bern: hep.
- Weber, B. (2005). Bildungsstandards, Qualifikationserwartungen und Kerncurricula: Stand und Entwicklungsperspektiven der ökonomischen Bildung. In DEGÖB (Ed.), *Standards in der ökonomischen Bildung* (pp. 17-50). Bergisch Gladbach: Verlag Thomas Hobein.
- Weber, B. (2013). Zwischen Subjekt, Lebenswelt, Wissenschaft und Verantwortung: Ökonomische Bildung im Spannungsfeld der Interessen. *GW-Unterricht*, 132(4), 5-16.
- Weber, B. (2014). Grundzüge einer Didaktik sozio-ökonomischer Allgemeinbildung. In A. Fischer & B. Zurstrassen (Eds.), *Sozioökonomische Bildung* (pp. 128-154). Bonn: Bundeszentrale für politische Bildung.

- Weber, S., Draxler, C., Bley, S., Wiethe-Körprich, M., Weiß, C., & Gürer, C. (2016). Large scale assessments in der kaufmännischen Berufsbildung – Intrapreneurship (CoBALIT). In K. Beck, M. Landenberger, & F. Oser (Eds.), *Technologiebasierte Kompetenzmessung in der beruflichen Bildung* (pp. 75-92). Bielefeld: wbv.
- Weber, S., Oser, F. K., Achtenhagen, F., Fretschner, M., & Trost, S. (2014). *Becoming an Entrepreneur*. Rotterdam: Sense Publishers.
- Weinert, F. E. (1999). *Konzepte der Kompetenz*. Paris: OECD.
- Weinert, F. E. (2001a). Concept of Competence: A Conceptual Clarification. In D. S. Rychen & L. H. Salganik (Eds.), *Defining and selecting key competencies: Theoretical and Conceptual Foundations (DeSeCo)* (pp. 45-65). Seattle: Hogrefe & Huber.
- Weinert, F. E. (2001b). Vergleichende Leistungsmessung in Schulen - Eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Ed.), *Leistungsmessungen in Schulen* (pp. 17-31). Weinheim: Beltz.
- Weißeno, G., Detjen, J., Juchler, I., Massing, P., & Richter, D. (2010). *Konzepte der Politik - ein Kompetenzmodell*. Bonn: Bundeszentrale für politische Bildung (BpB).
- Wilson, M. R. (2005). *Constructing measures. An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Winther, E. (2010). *Kompetenzmodellierung und Kompetenzmessung in der beruflichen Bildung*. Bielefeld.
- Winther, E., & Achtenhagen, F. (2008). Kompetenzstrukturmodell für die kaufmännische Ausbildung. *Zeitschrift für Berufs- und Wirtschaftspädagogik*, 104(4), 511-538.
- Winther, E., & Achtenhagen, F. (2009a). Simulationsaufgaben als innovatives Testverfahren für Industriekaufleute im Rahmen eines VET-LSA. *Wirtschaft und Erziehung (wue)*, 61(10), 317-325.
- Winther, E., & Achtenhagen, F. (2009b). Skalen und Stufen kaufmännischer Kompetenz. *Zeitschrift für Berufs- und Wirtschaftspädagogik*, 4, 521-556.
- Winther, E., Seeber, S., Festner, D., Sangmeister, J., & Liedtke, M. (2016). Large scale assessments in der kaufmännischen Berufsbildung – Das Unternehmensassessment ALUSIM (CoBALIT). In K. Beck, M. Landenberger, & F. Oser (Eds.), *Technologiebasierte Kompetenzmessung in der beruflichen Bildung* (pp. 55-73). Bielefeld: wbv.
- Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe.
- Wirtz, M. A. (2004). Über das Problem fehlender Werte: Wie der Einfluss fehlender Informationen auf Analyseergebnisse entdeckt und reduziert werden kann. *Die Rehabilitation*, 43(2), 109-115.
- Witt, R. (2006). Kompetenzstufenmodelle zur Messung ökonomischer Kompetenz. In G. Minnameier & E. Wuttke (Eds.), *Berufs- und wirtschaftspädagogische Grundlagenforschung: Lehr-Lern-Prozesse und Kompetenzdiagnostik* (pp. 407-419). Frankfurt a.M.: Lang.
- Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest: Generalised Item Response Modelling Software - Manual Version 2.0*. Camberwell, Victoria: ACER Press.
- Zabeck, J. (1991). Schlüsselqualifikationen. Ein Schlüssel für eine antizipative Berufsbildung. In F. Achtenhagen (Ed.), *Duales System zwischen Tradition und Innovation*. Köln: Institut für Berufs- und Wirtschaftspädagogik.
- ZfD. (2015a). Der Politikzyklus. *Politische Bildung Schweiz*. Retrieved from <http://politischebildung.ch/fuer-lehrpersonen/didaktik-und-methoden/politikzyklus> (06.04.2019)
- ZfD. (2015b). *Politikzyklus.ch: „Wie sollen Asylverfahren geregelt sein?“ Ein Lernspiel für die ausgehende Sekundarstufe I. Handreichung für die Lehrperson*. Zentrum Politische Bildung und Geschichtsdidaktik der PH FHNW und Zentrum für Demokratie Aarau.

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An Empirical Bayes Approach to Mantel-Haenszel DIF Analysis. *Journal of Educational Measurement*, 36(1), 1-28.

Rechtstexte

- Aufnahmereglement der Mittelschule vom 25. Juni 2011 (Stand 1. August 2016), SR 215.110. St.Gallen: Erziehungsrat des Kantons St.Gallen. Online: <https://www.gesetzessammlung.sg.ch/frontend/versions/1590> (01.05.2019).
- Bundesgesetz über die Berufsbildung (Berufsbildungsgesetz, BBG) vom 13. Dezember 2002 (Stand am 1. Januar 2019), SR 412.10. Bern: Bundesversammlung der Schweizerischen Eidgenossenschaft. Online: <https://www.admin.ch/opc/de/classified-compilation/20001860/index.html> (31.03.2019).
- Bundesverfassung der Schweizerischen Eidgenossenschaft vom 18. April 1999 (Stand am 1. Januar 2018), SR 101. Online <https://www.admin.ch/opc/de/classified-compilation/19995395/index.html> (30.03.2019).
- Interkantonale Vereinbarung über die Harmonisierung der obligatorischen Schule (HarmoS-Konkordat) vom 14.06.2007, SR 1.2. Bern: Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK). Online: http://edudoc.ch/record/24711/files/HarmoS_d.pdf (08.07.2018).
- Reglement über die Anerkennung von gymnasialen Maturitätsausweisen (Maturitäts-Anerkennungsreglement, MAR) vom 16.01.1995 bzw. 15.02.2015, SR 4.2.1.1. Bern: Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK) und Schweizerischer Bundesrat. Online: https://edudoc.ch/record/38112/files/VO_MAR_d.pdf (08.07.2018).
- Schweizerisches Zivilgesetzbuch vom 10. Dezember 1907 (Stand am 1. Januar 2019), SR 210. Bern: Bundesversammlung der Schweizerischen Eidgenossenschaft. Online: <https://www.admin.ch/opc/de/classified-compilation/19070042/index.html> (30.03.2019).
- Verordnung des SBFI über die berufliche Grundbildung Kauffrau/Kaufmann mit eidgenössischem Fähigkeitszeugnis (EFZ) vom 26. September 2011 (Stand am 1. Mai 2017), SR 412.101.221.73. Bern: Staatssekretariat für Bildung, Forschung und Innovation (SBFI).
- Verordnung über die Anerkennung von gymnasialen Maturitätsausweisen (Maturitäts-Anerkennungsverordnung, MAV) vom 15. Februar 1995 (Stand am 1. Januar 2013), SR 413.11. Bern: Schweizerischer Bundesrat.
- Verordnung über die Berufsbildung (Berufsbildungsverordnung, BBV) vom 19. November 2003 (Stand am 1. Februar 2019), SR 412.101. Bern: Schweizerischer Bundesrat.
- Verordnung über die eidgenössische Berufsmaturität (Berufsmaturitätsverordnung, BMV) vom 24. Juni 2009 (Stand am 23. August 2016), SR 412.103.1. Bern: Schweizerischer Bundesrat.
- Verordnung über die schweizerische Maturitätsprüfung vom 7. Dezember 1998 (Stand am 1. Januar 2013), SR. 413.12. Bern: Schweizerischer Bundesrat.

ANHANG

A1 Testentwicklung des WBK-T2

A1.1 Domänenanalyse des gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereichs

Tabelle A- 1. Domänenanalyse des gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereichs

Politikfeld (alphabetisch)	Sozioökonomische Problemsituation	Schlüsselwörter	Eidgenössische Abstimmungsvorlagen: Originaltitel (Abstimmungsdatum, chronologisch)
Arbeit	Gehaltsgerechtigkeit (Managervergütungen)	Arbeitszeiten, Ferienwochen gerechte Löhne, Mindestlohn, Mangerlöhne	Volksinitiative «Für den Schutz fairer Löhne (Mindestlohn-Initiative)» (18.05.2014) Volksinitiative «1:12 – Für gerechte Löhne» (14.11.2013) Volksinitiative «Gegen die Abzockerei (Abzocker-Initiative)» (03.03.2013) Volksinitiative «6 Wochen Ferien für alle» (11.03.2012) Volksinitiative «für eine kürzere Arbeitszeit» (03.03.2002)
Bildung & Kultur	---	Lehrstellen, Stipendien, Grundbildung, Kulturförderung, Minarette	---
Energie & Umwelt	Energieversorgung	Atomenergie, Atomausstieg, erneuerbare Energien, Energemarkt, Energieabgabe	Volksinitiative «Für den geordneten Ausstieg aus der Atomenergie (Atomausstiegsinitiative)» (27.11.2016) Elektrizitätsmarktgesetz (22.09.2002) Volksinitiative «für einen Verfassungsartikel über eine «Energienkungsabgabe für die Umwelt» (24.09.2000)
	Umwelt- und Klimaschutz	Alpenschutz, Gewässerschutz, Luftschutz, Tierschutz, Pflanzenschutz, Biodiversität, Nachhaltigkeit	Volksinitiative "für einen autofreien Sonntag pro Jahreszeit - ein Versuch für vier Jahre (Sonntags-Initiative)" (18.05.2003) Volksinitiative "zum Schutze des Alpengebietes vor dem Transitverkehr" (20.02.1994) Bundesgesetz über den Schutz der Gewässer (Gewässerschutzgesetz, GSchG) (17.05.1992)

Tabelle A- 1. Domänenanalyse des gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereichs

Politikfeld (alphabetisch)	Sozioökonomische Problemsituation	Schlüsselwörter	Eidgenössische Abstimmungsvorlagen: Originaltitel (Abstimmungsdatum, chronologisch)
Finanzen & Steuern	Staatsverschuldung	Bundesfinanzen, Finanzordnung, Schuldenbremse	Bundesbeschluss über eine neue Finanzordnung (28.11.2004) Bundesbeschluss über eine Schuldenbremse (02.12.2001) Bundesbeschluss über Massnahmen zum Haushaltausgleich (07.06.1998)
	Steuerwettbewerb/ Finanzausgleich	Finanzausgleich, Steuerwettbewerb Verbrauchssteuern, Unternehmenssteuern, Wohneigentumssteuern Ehe- und Familiensteuern, Familienzulagen, Kinderzulagen	Bundesgesetz über steuerliche Massnahmen zur Stärkung der Wettbewerbsfähigkeit des Unternehmensstandorts Schweiz (Unternehmenssteuerreformgesetz III) (12.02.2017) Volksinitiative «Schluss mit den Steuerprivilegien für Millionäre (Abschaffung der Pauschalbesteuerung)» (30.11.2014) Volksinitiative «Für faire Steuern. Stopp dem Missbrauch beim Steuerwettbewerb (Steuergerechtigkeits-Initiative)» (28.11.2010) Bundesbeschluss zur Neugestaltung des Finanzausgleichs und der Aufgabenteilung zwischen Bund und Kantonen (NFA) (28.11.2004)
Forschung & Technologie	Biotechnologie	Fortpflanzungsmedizin, Transplantationsmedizin, Gentechnologie, Tierversuche	Bundesbeschluss über die Änderung der Verfassungsbestimmung zur Fortpflanzungsmedizin und Gentechnologie im Humanbereich (14.06.2015) Volksinitiative «zur Abschaffung der Tierversuche» (07.03.1993)
Gesellschaft & Ethik	---	Behinderte Menschen, gleichgeschlechtliche Partnerschaften Drogenkonsum Schwangerschaftsabbruch, Sexualdelikte, Gewaltdelikte	Volksinitiative «für eine vernünftige Hanf-Politik mit wirksamem Jugendschutz» (30.11.2008) Bundesgesetz über die eingetragene Partnerschaft gleichgeschlechtlicher Paare (Partnerschaftsgesetz, PartG) (05.06.2005) Volksinitiative «Gleiche Rechte für Behinderte» (18.05.2003)
Gesundheit	Krankenversicherung	Krankenversicherung, öffentliche Krankenkasse Gesundheitskosten, Medikamentenpreise	Volksinitiative «Für eine öffentliche Krankenkasse» (28.09.2014) Volksinitiative «für tiefere Arzneimittelpreise» (04.03.2001) Volksinitiative «für tiefere Spitalkosten» (26.11.2000) Bundesgesetz über die Krankenversicherung (KVG) (04.12.1994)

Tabelle A- 1. Domänenanalyse des gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereichs

Politikfeld (alphabetisch)	Sozioökonomische Problemsituation	Schlüsselwörter	Eidgenössische Abstimmungsvorlagen: Originaltitel (Abstimmungsdatum, chronologisch)
Internationale Beziehungen & wirtschaftliche Integration	Bilaterale Abkommen Schweiz/EU (Rahmenabkommen)	Bretton Woods Europäischer Wirtschaftsraum (EWR), Europäische Gemeinschaft (EG), Europäische Union (EU) United Nations (UNO)	<i>Volksinitiative «für den Beitritt der Schweiz zur Organisation der Vereinten Nationen (UNO)» (03.03.2002)</i> <i>Volksinitiative «Ja zu Europa» (04.03.2001)</i>
Landwirtschaft	Agrarhandel	Freihandel, Bilaterale Abkommen (Landwirtschaft)	<i>Volksinitiative «Keine Spekulation mit Nahrungsmitteln!» (28.02.2016)</i> <i>Bundesbeschluss über die Genehmigung und Umsetzung des Protokolls über die Ausdehnung des Freizügigkeitsabkommens auf die neuen EG-Mitgliedstaaten zwischen der Schweizerischen Eidgenossenschaft einerseits und der Europäischen Gemeinschaft und ihren Mitgliedstaaten andererseits sowie über die Genehmigung der Revision der flankierenden Massnahmen zur Personenfreizügigkeit (25.09.2005)</i> <i>Bundesbeschluss über die Genehmigung der sektoriellen Abkommen zwischen der Schweizerischen Eidgenossenschaft einerseits und der Europäischen Gemeinschaft sowie gegebenenfalls ihren Mitgliedstaaten oder der Europäischen Atomgemeinschaft andererseits (21.05.2000)</i> <i>Bundesbeschluss über den Europäischen Wirtschaftsraum (EWR) (06.12.1992)</i>
	Agrarproduktion	ökologische Landwirtschaft, Nahrungsmittelpreise (Getreide, Milch)	<i>Volksinitiative «für Lebensmittel aus gentechnikfreier Landwirtschaft» (27.11.2005)</i> <i>Volksinitiative «für preisgünstige Nahrungsmittel und ökologische Bauernhöfe» (27.09.1998)</i> <i>Bundesbeschluss über die Volksinitiative "für eine umweltgerechte und leistungsfähige bäuerliche Landwirtschaft" (Gegenentwurf) (12.03.1995)</i>
Migration & Integration	Flüchtlingsströme	Asylwesen, Ausländerwesen, Zuwanderung, Ausschaffung Einbürgerung	<i>Bundesbeschluss über die erleichterte Einbürgerung von Personen der dritten Ausländergeneration (12.02.2017)</i> <i>Änderung des Asylgesetzes (AsylG) (05.06.2016)</i> <i>Volksinitiative «Gegen Masseneinwanderung» (09.02.2014)</i> <i>Volksinitiative «Für die Ausschaffung krimineller Ausländer (Ausschaffungsinitiative)» (28.11.2010)</i>

Tabelle A- 1. Domänenanalyse des gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereichs

Politikfeld (alphabetisch)	Sozioökonomische Problemsituation	Schlüsselwörter	Eidgenössische Abstimmungsvorlagen: Originaltitel (Abstimmungsdatum, chronologisch)
Öffentliche Sicherheit & Ordnung	---	Armee, Waffen, Nachrichtendienst Wehrdienst, Zivildienst, Friedensdienst	Bundesgesetz über den Nachrichtendienst (Nachrichtendienstgesetz, NDG) (25.09.2016) Volksinitiative «Ja zur Aufhebung der Wehrpflicht» (22.09.2013) Volksinitiative «Für den Schutz vor Waffengewalt» (13.02.2011)
Raumplanung	Wohnungsmarkt und Mietpreise	Grundstückserwerb, Zweitwohnungen, Raumplanung Mietpreise, Bausparen	Änderung des Bundesgesetzes über die Raumplanung (Raumplanungsgesetz, RPG) (03.03.2013) Volksinitiative «Schluss mit uferlosem Bau von Zweitwohnungen!» (11.03.2012) Volksinitiative «Ja zu fairen Mieten» (18.05.2003)
Soziale Sicherheit & Vorsorge	Altersvorsorge	Alters- und Hinterlassenversicherung (AHV), berufliche Vorsorge (BV), Finanzierung (Mehrwertsteuer, Erbchaftssteuer), Rentenalter	Volksinitiative «AHVplus: für eine starke AHV» (25.09.2016) Volksinitiative «Für ein flexibles AHV-Alter» (30.11.2008) Bundesgesetz über die Alters- und Hinterlassenversicherung (11. AHV-Revision) (16.05.2004) Bundesbeschluss über die Finanzierung der AHV/IV durch Anhebung der Mehrwertsteuersätze (16.05.2004) Volksinitiative "zum Ausbau von AHV und IV" (25.06.1995)
	Arbeitslosigkeit	Arbeitslosenversicherung (ALV)	Bundesgesetzes über die obligatorische Arbeitslosenversicherung und die Insolvenzenschädigung (Arbeitslosenversicherungsgesetz, AVIG) (19.03.2010) Bundesbeschluss über die Finanzierung der Arbeitslosenversicherung (13.12.1996) Bundesbeschluss über Massnahmen in der Arbeitslosenversicherung (26.09.1993)
Verfassung & Grundrechte	---	Stimm- und Wahlrecht, Volksinitiative Staatsverträge	

Tabelle A- 1. Domänenanalyse des gesamtgesellschaftlichen/gesamtwirtschaftlichen Lebensbereichs

Politikfeld (alphabetisch)	Sozioökonomische Problemsituation	Schlüsselwörter	Eidgenössische Abstimmungsvorlagen: Originaltitel (Abstimmungsdatum, chronologisch)
Verkehr	Strassenverkehr	Autobahn, Strassenbau, Strassenverkehr Nationalstrassenabgabe („Vignette“), Schwerverkehrsabgabe, Leistungsabhängige Schwerverkehrsabgabe (LSVA) Neue Eisenbahn- Alpentransversale (NEAT)	Änderung des Bundesgesetzes über den Strassentransitverkehr im Alpengebiet (STVG) (Sanierung Gotthard-Strassentunnel) (28.02.2016) Änderung des Bundesgesetzes über die Abgabe für die Benützung von Nationalstrassen (Nationalstrassenabgabegesetz, NSAG) (24.11.2013) Bundesbeschluss über die Einführung einer leistungs- oder verbrauchsabhängigen Schwerverkehrsabgabe (20.02.1994) Bundesbeschluss über den Bau der schweizerischen Eisenbahn-Alpentransversale (Alpentransit-Beschluss) (27.09.1992) Volksinitiative "Stopp dem Beton - für eine Begrenzung des Strassenbaus!" (01.04.1990)
	öffentlicher Verkehr	öffentlicher Verkehr, Eisenbahn, Finanzierung	Bundesbeschluss über die Finanzierung und den Ausbau der Eisenbahninfrastruktur (direkter Gegenentwurf zur Volksinitiative «Für den öffentlichen Verkehr») (09.02.2014) Bundesbeschluss über Bau und Finanzierung von Infrastrukturvorhaben des öffentlichen Verkehrs (29.11.1998) Volksinitiative "zur Förderung des öffentlichen Verkehrs" (03.03.1991)
Verwaltung	---	Bundesrat, Bundesversammlung, Kantone, Gemeinden Rechtspflege, Verwaltungsorganisation	Volksinitiative "für Beschleunigung der direkten Demokratie (Behandlungsfristen für Volksinitiativen in Form eines ausgearbeiteten Entwurfs)" (12.03.2000) Bundesbeschluss über die Reform der Justiz (12.03.2000) Regierungs- und Verwaltungsorganisationsgesetz (RVOG) (09.06.1996)
Wettbewerb	Service Public	Post, Telekom, Rundfunk (Radio und Fernsehen)	Volksinitiative «Pro Service public» (05.06.2016) Änderung des Bundesgesetzes über Radio und Fernsehen (RTVG) (14.06.2015) Volksinitiative "Postdienste für alle" (26.09.2004)
	Strukturpolitik	Spielbanken, Geldspiele, Buchpreisbindung Kriegsmaterialausfuhr, Pulverregal	Bundesgesetz über die Buchpreisbindung (BuPG) (11.03.2012) Volksinitiative "Für ein Verbot von Kriegsmaterial-Exporten" (29.11.2009) Bundesbeschluss über die Aufhebung des Spielbankenverbots (07.03.1993)

Quelle: BFS (2017); Inhaltsanalyse nach adaptiertem Kategoriensystem von Gilg, Raymond, & Bindschedler (1966) und CAP (2105); eigene Kategorisierung und Darstellung.

A1.2 Kriteriengeleitete Beurteilung sozioökonomischer Problemsituationen für WBK-T2

Tabelle A- 2. Beurteilung sozioökonomischer Problemsituationen für WBK-T2 nach den Kriterien Repräsentativität, Wissenschaftlichkeit und Bekanntheit

Sozioökonomische Problemsituationen	Repräsentativität (politische Relevanz): Eidgenössische Abstimmungsvorlagen (Anzahl, Schlüsselwörter)	Wissenschaftlichkeit (fachliche Struktur): Inhaltsstruktur der Wirtschaftswissenschaften auf Tertiärstufe (Code)	Bekanntheit (curriculare Relevanz): Lerninhalte/Lernziele in Lehrplänen auf Sekundarstufe II
Gehaltsgerechtigkeit (Managervergütungen)	# 3 Vorlagen Arbeitszeiten, Ferienwochen, gerechte Löhne, Mindestlohn, Mangerlöhne	VWL, Grundlagen: Produktionsfaktoren (v12); Wirtschaftskreislauf (v16) VWL, Mikro: Angebot & Nachfrage (v21) VWL, Makro: Arbeitsmarkt (v34) BWL, Grundlagen: Soziale Umwelt (b17) BWL, Unternehmungsbereiche: Personal (b37)	VWL: Wirtschaftspolitik (sozialer Ausgleich), Sozialpolitik (Einkommensverteilung) VWL: Produktionsfaktoren (Arbeit), Preisbildung, Marktwirtschaft (Preisgrenzen, Mindestlohn) BWL/FRW: Personalwesen (Anreizsysteme, Lohnformen), Lohnbuchhaltung RECHT/STAAT: Grundrechte (insb. Art. 8 BV) GESCHICHTE: Industrialisierung, Rollen Patron/Arbeiter, „Fabrikgesetz“ PHILOSOPHIE: Ethik
Energieversorgung	# 12 Vorlagen Atomenergie, Atomausstieg, Energiemarkt, Energieabgabe, erneuerbare Energien, Nachhaltigkeit	VWL, Grundlagen: Produktionsfaktoren (v12); Wirtschaftssektoren, Institutionen (v15) VWL, Makro: Angebot & Nachfrage (v21); Wirtschaftsentwicklung (v35); Wirtschaftspolitik (v37) BWL, Grundlagen: Unternehmenstypologie (b12); Technologische Umwelt (b15)	VWL: Wirtschaftspolitik (Umweltqualität), Wirtschaftswachstum (nachhaltige Entwicklung, Ökonomie und Ökologie) VWL: internationale Arbeitsteilung (Globalisierung) VWL: Marktwirtschaft (Marktversagen, Umweltsteuer, Emissionszertifikate) RECHT/STAAT: Aufgaben Bund/Kantone (Umwelt, Art. 73 ff. BV und Art. 76 ff. BV; Energie, Art. 89 ff. BV) GEOGRAFIE: Rohstoffvorkommen, Ökosysteme

Tabelle A- 2. Beurteilung sozioökonomischer Problemsituationen für WBK-T2 nach den Kriterien Repräsentativität, Wissenschaftlichkeit und Bekanntheit

Sozioökonomische Problemsituationen	Repräsentativität (politische Relevanz): Eidgenössische Abstimmungsvorlagen (Anzahl, Schlüsselwörter)	Wissenschaftlichkeit (fachliche Struktur): Inhaltsstruktur der Wirtschaftswissenschaften auf Tertiärstufe (Code)	Bekanntheit (curriculare Relevanz: Lerninhalte/Lernziele in Lehrplänen auf Sekundarstufe II)
Umwelt- und Klimaschutz	# 5 Vorlagen Alpenschutz, Gewässerschutz, Luftschutz, Tierschutz, Pflanzenschutz, Biodiversität	VWL, Grundlagen: Produktionsfaktoren (v12) VWL, Mikro: Angebot & Nachfrage (v21); Rolle des Staates (v23) VWL, Makro: Wirtschaftsentwicklung (v35); Wirtschaftspolitik (v37) BWL, Grundlagen: Ökologische Umwelt (b16)	VWL: Wirtschaftspolitik (Umweltqualität); Wirtschaftswachstum (nachhaltige Entwicklung, Ökonomie und Ökologie); internationale Arbeitsteilung (Globalisierung); Marktwirtschaft (Marktversagen, Umweltsteuer, Emissionszertifikate) RECHT/STAAT: Aufgaben Bund/Kantone (Umwelt, Art. 73 ff. BV und Art. 76 ff. BV)
Staatsverschuldung	# 7 Vorlagen Bundesfinanzen, Finanzordnung, Schuldenbremse	VWL, Grundlagen: Wirtschaftssystem/-ordnung (v13) VWL, Makro: Volkswirtschaftliche Gesamtrechnung (v31); Konjunktur (v32); Wirtschaftsentwicklung (v35); Öffentliche Finanzen (v36)	VWL: Staatsverschuldung (Ausgaben, Einnahmen, Defizit, Verschuldung); Konjunkturpolitik (keynesianische Konzeption, Fiskalpolitik)
Steuerwettbewerb/Finanzausgleich	# 10 Vorlagen Finanzausgleich, Steuerwettbewerb, Verbrauchssteuern, Unternehmenssteuern, etc.	VWL, Grundlagen: Wirtschaftssystem/-ordnung (v13); Wirtschaftskreislauf (v16) VWL, Makro: Öffentliche Finanzen (v36) BWL, Grundlagen: Ökonomische Umwelt (b14), Soziale Umwelt (b17)	RECHT/STAAT: Steuern (Steuerarten, Steuerhoheit)
Biotechnologie	# 11 Vorlagen Fortpflanzungsmedizin, Stammzellenforschung, Transplantationsmedizin, Gentechnologie, Tierversuche	VWL, Grundlagen: Wirtschaftssystem/-ordnung (v13); VWL, Makro: Wirtschaftsentwicklung (v35)	RECHT/STAAT: Aufgaben Bund und Kantone (Gesundheit, Art. 118b ff. BV) PHILOSOPHIE: Ethik

Tabelle A- 2. Beurteilung sozioökonomischer Problemsituationen für WBK-T2 nach den Kriterien Repräsentativität, Wissenschaftlichkeit und Bekanntheit

Sozioökonomische Problemsituationen	Repräsentativität (politische Relevanz): Eidgenössische Abstimmungsvorlagen (Anzahl, Schlüsselwörter)	Wissenschaftlichkeit (fachliche Struktur): Inhaltsstruktur der Wirtschaftswissenschaften auf Tertiärstufe (Code)	Bekanntheit (curriculare Relevanz: Lerninhalte/Lernziele in Lehrplänen auf Sekundarstufe II)
Krankenversicherung	# 14 Vorlagen Krankenversicherung, öffentliche Krankenkasse, Gesundheitskosten, Medikamentenpreise	VWL, Grundlagen: Wirtschaftssystem/-ordnung (v13); Wirtschaftssectoren/Institutionen (v15) VWL, Mikro: Wettbewerb/Marktformen (v22); Rolle des Staates (v23) BWL, Grundlagen: Unternehmenstypologie (b12)	VWL: Wirtschaftspolitik (sozialer Ausgleich), Sozialpolitik (Einkommensverteilung) BWL: Privatversicherungen (Krankenversicherung) RECHT/STAAT: Aufgaben Bund/Kantone (Gesundheit, Art. 117 ff. BV)
Bilaterale Abkommen Schweiz/EU (Rahmenabkommen)	# 12 Vorlagen Bretton Woods, Europäischer Wirtschaftsraum (EWR), Europäische Gemeinschaft (EG), Europäische Union (EU), UNO	VWL, IntWB: Aussenwirtschaft (v42) VWL, IntWB: Internationale Organisationen (v44)	VWL: internationale Arbeitsteilung (Globalisierung, Freihandel, WTO, Schweiz und Europa) RECHT/STAAT: Staatsverträge, multilaterale und bilaterale Abkommen; Aufgaben Bund/Kantone (Beziehungen zum Ausland, Art. 54 ff. BV) GESCHICHTE: „Europa“ als politische Idee
Agrarhandel	# 3 Vorlagen (indirekt) Bilaterale Abkommen bis 1999 (u.a. Freihandel, Zollerleichterung), Bilaterale Abkommen I ab 1999 (Landwirtschaft), Bilaterale Abkommen II ab 2004 (Landwirtschaftliche Verarbeitungsprodukte)	VWL, Grundlagen: Arbeitsteilung/Tausch/Geld (v14); Wirtschaftskreislauf (v16) VWL, Mikro: Angebot & Nachfrage (v21) VWL, IntWB: Weltwirtschaft (v41); Aussenwirtschaft (v42)	VWL: Wirtschaftspolitik (aussenwirtschaftliches Gleichgewicht); Internationale Arbeitsteilung (Globalisierung, Freihandel, WTO, europäische Integration), Zahlungsbilanz (Warenhandel); Marktwirtschaft (Preisgrenzen, Subventionen) RECHT/STAAT: Aussenpolitik (bilaterale Abkommen CH/EU); Aufgaben Bund/Kantone (Landwirtschaft und Ernährungssicherheit, Art. 104 f. BV) GEOGRAFIE: Klimazonen, Anbauregionen, Nahrungsmittelproduktion und -handel

Tabelle A- 2. Beurteilung sozioökonomischer Problemsituationen für WBK-T2 nach den Kriterien Repräsentativität, Wissenschaftlichkeit und Bekanntheit

Sozioökonomische Problemsituationen	Repräsentativität (politische Relevanz): Eidgenössische Abstimmungsvorlagen (Anzahl, Schlüsselwörter)	Wissenschaftlichkeit (fachliche Struktur): Inhaltsstruktur der Wirtschaftswissenschaften auf Tertiärstufe (Code)	Bekanntheit (curriculare Relevanz): Lerninhalte/Lernziele in Lehrplänen auf Sekundarstufe II
Agrarproduktion	# 12 Vorlagen ökologische Landwirtschaft, Ernährungssicherheit, Nahrungsmittelpreise (Getreide, Milch)	VWL, Grundlagen: Wirtschaftssectoren/Institutionen (v15); Wirtschaftskreislauf (v16) VWL, Mikro: Wettbewerb/Marktformen (v22), Rolle des Staates (v23) VWL, Makro: Wirtschaftspolitik (v37) BWL, Grundlagen: Unternehmenstypologie (b12)	RECHT/STAAT: Aufgaben Bund/Kantone (Landwirtschaft und Ernährungssicherheit, Art. 104 f. BV)
Flüchtlingsströme	# 15 Vorlagen Asylwesen, Ausländerwesen, Zuwanderung, Ausschaffung	VWL, IntWB: Weltwirtschaft (v41)	VWL: Internationale Arbeitsteilung (Globalisierung) RECHT/STAAT: Grundrechte, Staatsverträge (EMRK, Genfer Flüchtlingskonvention); Aufgaben Bund/Kantone (Aufenthalt und Niederlassung von Ausländer*innen, Art. 121 ff. BV)
Wohnungsmarkt und Mietpreise	# 11 Vorlagen Grundstückwerb, Zweitwohnungen, Mietpreise, Bausparen, Raumplanung	VWL, Grundlagen: Produktionsfaktoren (v12); Wirtschaftssystem/-ordnung (v13) VWL, Mikro: Angebot & Nachfrage (v21); Wettbewerb/Marktformen (v22); Rolle des Staates (v23) VWL, Makro: Wirtschaftspolitik (v37)	VWL: Wirtschaftspolitik (sozialer Ausgleich); Produktionsfaktoren (Boden), Marktwirtschaft RECHT/STAAT: Mietrecht; Aufgaben Bund/Kantone (Raumplanung, Art. 75 und 75b BV; Wohnen, Art. 108 f. BV) GEOGRAFIE: Raumplanung, Stadtplanung
Altersvorsorge	# 17 Vorlagen Alters- und Hinterlassenenversicherung (AHV), berufliche Vorsorge (BV), Rentenalter, Finanzierung (Mehrwertsteuer, Erbschaftssteuer, Goldreserven)	VWL, Grundlagen: Wirtschaftssystem/-ordnung (v13) VWL, Mikro: Rolle des Staates (v23) VWL, Makro: Wirtschaftspolitik (v37); Öffentliche Finanzen (v36) BWL, Grundlagen: Soziale Umwelt (b17)	VWL: Wirtschaftspolitik (sozialer Ausgleich, Vollbeschäftigung); Sozialpolitik (Sozialversicherungen, Vorsorge); Arbeitslosigkeit RECHT/STAAT: Sozialversicherungen (z.B. AHV, ALV), Aufgaben Bund/Kantone (Arbeit und soziale Sicherheit, Art. 110 ff. BV)

Tabelle A- 2. Beurteilung sozioökonomischer Problemsituationen für WBK-T2 nach den Kriterien Repräsentativität, Wissenschaftlichkeit und Bekanntheit

Sozioökonomische Problemsituationen	Repräsentativität (politische Relevanz): Eidgenössische Abstimmungsvorlagen (Anzahl, Schlüsselwörter)	Wissenschaftlichkeit (fachliche Struktur): Inhaltsstruktur der Wirtschaftswissenschaften auf Tertiärstufe (Code)	Bekanntheit (curriculare Relevanz: Lerninhalte/Lernziele in Lehrplänen auf Sekundarstufe II)
Arbeitslosigkeit	# 6 Vorlagen Arbeitslosenversicherung (ALV), Massnahmen, Finanzierung	VWL, Grundlagen: Produktionsfaktoren (v12) VWL, Mikro: Angebot & Nachfrage (v21) VWL, Makro: Arbeitsmarkt (v34); Wirtschaftsentwicklung (v35) BWL, Grundlagen: Ökonomische Umwelt (b14)	VWL: Wirtschaftspolitik (Vollbeschäftigung); Arbeitslosigkeit RECHT/STAAT: Sozialversicherungen (z.B. AHV, ALV); Aufgaben Bund/Kantone (Arbeit und soziale Sicherheit, Art. 110 ff. BV)
Strassenverkehr	# 16 Vorlagen Autobahn, Strassenbau, Strassenverkehr, Nationalstrassenabgabe („Vignette“), Schwerverkehrsabgabe, LSV, NEAT, Tempo 30	VWL, Grundlagen: Wirtschaftssectoren/Institutionen (v15) VWL, Mikro: Angebot & Nachfrage (v21) VWL, Makro: Wirtschaftspolitik (v37)	VWL: Marktwirtschaft (Marktversagen, Umweltsteuer) RECHT/STAAT: Aufgaben Bund und Kantone (öffentliche Werke und Verkehr, Art. 81 BV) GEOGRAFIE: Infrastruktur
Öffentlicher Verkehr	# 3 Vorlagen öffentlicher Verkehr, Finanzierung	VWL, Grundlagen: Wirtschaftssystem/-ordnung (v13) VWL, Mikro: Wettbewerb/Marktformen (v22); Rolle des Staates (v23) VWL, Makro: Wirtschaftspolitik (v37) BWL, Grundlagen: Unternehmenstypologie (b12)	RECHT/STAAT: Aufgaben Bund und Kantone (öffentliche Werke und Verkehr, Art. 81 BV)
Service Public	# 3 Vorlagen Post, Telekom, Rundfunk (Radio und Fernsehen)	VWL, Grundlagen: Wirtschaftssystem/-ordnung (v13); Wirtschaftssectoren/Institutionen (v15) VWL, Mikro: Wettbewerb/Marktformen (v22); Rolle des Staates (v23) VWL, Makro: Wirtschaftspolitik (v37) BWL, Grundlagen: Unternehmenstypologie (b12)	VWL: Wirtschaftspolitik; Marktwirtschaft (Marktversagen, staatliche Monopole) RECHT/STAAT: Aufgaben Bund und Kantone (Kommunikation, Art. 92 f. BV; Wirtschaft, Art. 94 ff. BV) GESCHICHTE: von Agrargesellschaft zu Dienstleistungsgesellschaft

Tabelle A- 2. Beurteilung sozioökonomischer Problemsituationen für WBK-T2 nach den Kriterien Repräsentativität, Wissenschaftlichkeit und Bekanntheit

Sozioökonomische Problemsituationen	Repräsentativität (politische Relevanz): Eidgenössische Abstimmungsvorlagen (Anzahl, Schlüsselwörter)	Wissenschaftlichkeit (fachliche Struktur): Inhaltsstruktur der Wirtschaftswissenschaften auf Tertiärstufe (Code)	Bekanntheit (curriculare Relevanz: Lerninhalte/Lernziele in Lehrplänen auf Sekundarstufe II)
Strukturpolitik	# 8 Vorlagen Spielbanken, Geldspiele, Buchpreisbindung, Kriegsmaterialausfuhr, Pulverregal	VWL, Grundlagen: Wirtschaftssystem/-ordnung (v13) VWL, Mikro: Wettbewerb/Marktformen (v22) VWL, Makro: Wirtschaftspolitik (v37)	VWL: Wirtschaftspolitik; Preisbildung, Marktwirtschaft (Marktversagen); Strukturwandel

Hinweis: VWL = Volkswirtschaftslehre, BWL = Betriebswirtschaftslehre, RECHT/STAAT = Rechts- und Staatskunde. Mikro = Mikroökonomie, Makro = Makroökonomie, IntWB = Internationale Wirtschaftsbeziehungen.

Quelle: Repräsentativität: BFS (2017); adaptiertes Kategoriensystem von Gilg, Raymond, & Bindschedler (1966) und CAP (2105). Wissenschaftlichkeit: Kategoriensystem von Schumann et al. (2010). Bekanntheit: Erziehungsdirektion des Kantons Bern (2005), Erziehungsdirektion des Kantons St.Gallen (2006), SBFI (2006, 2012), SBFI & SKKAB (2011a, 2011b); eigenes Kategorien- und Bewertungssystem.

A1.3 Testrevision des WBK-T1

Tabelle A- 3. Testrevision des WBK-T1 exemplarisch anhand der Problemsituationen Altersvorsorge (AHV)

Item-Nr.	Inhaltselement	Kognitionsprozess	Antwortformat	Score	Status nach Revision: Begründung
AHV_1	Finanzielle Risiken	k1: wiedererkennen/ wiedergeben	f1: SR (4 Antwortoptionen)	1	
AHV_1m	Finanzielle Risiken	k1: wiedererkennen/ wiedergeben	f2: short CR	2	Modifiziert: Format
AHV_2	Beitragsbeginn AHV-Kasse	k1: wiedererkennen/ wiedergeben	f1: SR (4 Antwortoptionen)	1	Unverändert
AHV_3	AHV-Leistungsempfänger	k1: wiedererkennen/ wiedergeben	f1: SR (4 Antwortoptionen)	1	Eliminiert: Psychometrische Qualität
AHV_4	AHV-Beitragssatz	k1: wiedererkennen/ wiedergeben	f1: SR (4 Antwortoptionen)	1	Eliminiert: Psychometrische Qualität
AHV_5	Versicherungsprinzip	k1: wiedererkennen/ wiedergeben	f1: SR (4 Antwortoptionen)	1	Unverändert
AHV_6	Umverteilungscharakter	k1: wiedererkennen/ wiedergeben	f1: SR (4 Antwortoptionen)	1	Unverändert
AHV_7	Generationenvertrag	k1: wiedererkennen/ wiedergeben	f1: SR (4 Antwortoptionen)	1	Unverändert
AHV_8	Mehrwertsteuer und privater Konsum	k2: verstehen/anwenden	f3: extended CR	1	Unverändert
AHV_9	Mehrwertsteuer und Unternehmen	k2: verstehen/anwenden	f3: extended CR	1	Eliminiert: Format (Typus doppelt vorhanden)
AHV_10	Mehrwertsteuer und Staatshaushalt	k2: verstehen/anwenden	f3: extended CR	1	Eliminiert: Inhalt (fachlich inadäquat)
AHV_11	Rententalter und Arbeitsmarkt	k2: verstehen/anwenden	f3: extended CR	4	
AHV_11m	Rententalter und Arbeitsmarkt	k2: verstehen/anwenden	f3: extended CR	2	Modifiziert: Psychometrische Qualität, Umfang (Frageformulierung, Antwortstruktur), Kodierung

Tabelle A- 3. Testrevision des WBK-T1 exemplarisch anhand der Problemsituationen Altersvorsorge (AHV)

Item-Nr.	Inhaltselement	Kognitionsprozess	Antwortformat	Score	Status nach Revision: Begründung
AHV_12	AHV-Finanzierung: Lohnbeiträge vs. Altersrenten	k3: vergleichen/ beurteilen/entscheiden	f3: extended CR	5	
AHV_12m	AHV-Finanzierung: Lohnbeiträge vs. Altersrenten	k3: vergleichen/ beurteilen/entscheiden	f3: extended CR	2	Modifiziert: Psychometrische Qualität, Darstellung (Frageformulierung, Umfang, Antwortstruktur), Kodierung
AHV_13n	AHV-Kasse: Beiträge und Leistungen	k2: verstehen/anwenden	f1: SR (3 Antwortoptionen)	4	Neu: Inhalt, Format/Darstellung (Diagramm)

Hinweis: rot = eliminiert, orange = modifiziert (z.B. AHV_1m), grün = neu konstruiert (z.B. AHV_13n).

Kognitionsprozess: k1 = Kognitionsstufe 1 (wiedererkennen/wiedergeben), k2 = Kognitionsstufe 2 (verstehen/anwenden), k3 = Kognitionsstufe 3 (vergleichen/beurteilen/entscheiden).

Antwortformat: f1 = Format 1 (Selected-Response, Multiple-Choice), f2 = Format 2 (Short-Constructed-Response, kurze Antwort), f3 = Format 3 (Extended-Constructed-Response, ausführliche Antwort). SR = Selected-Response, CR = Constructed-Response.

Quelle: Ackermann (2018a).

A1.4 Item-Spezifikationen des WBK-T2

Tabelle A- 4. Item-Spezifikationen des WBK-T2 nach Inhalt, Format und Kodierung

Item-Nr.	Inhaltselement	Itemtypus	Kognitions- prozess	Darstellungs- form	Antwortformat	Antwort- kategorien (Scoring)	Status nach Revision: Begründung
Problemsituation: Altersvorsorge							
AHV_1m	Finanzielle Risiken	V	k1	Text	f2	Dichotom (0/1) → Σ 2	Modifiziert: Format
AHV_2	Beitragsbeginn AHV-Kasse	I	k1	Text	f1 (4 Antwortoptionen)	Dichotom (0/1)	Unverändert
AHV_3	AHV-Leistungsempfänger	I	k1	Text	f1 (4 Antwortoptionen)	Dichotom (0/1)	Eliminiert: Item-Kennwerte
AHV_4	AHV-Beitragssatz	I	k1	Text	f1 (4 Antwortoptionen)	Dichotom (0/1)	Eliminiert: Item-Kennwerte
AHV_5	Versicherungsprinzip	I	k1	Text	f1 (4 Antwortoptionen)	Dichotom (0/1)	Unverändert
AHV_6	Umverteilungscharakter	I	k1	Text	f1 (4 Antwortoptionen)	Dichotom (0/1)	Unverändert
AHV_7	Generationenvertrag	I	k1	Text	f1 (4 Antwortoptionen)	Dichotom (0/1)	Unverändert
AHV_8	Mehrwertsteuer und privater Konsum	II	k2	Text	f3	Dichotom (0/1)	Unverändert
AHV_9	Mehrwertsteuer und Unternehmen	II	k2	Text	f3	Dichotom (0/1)	Eliminiert: Format (doppelt vorhanden)
AHV_10	Mehrwertsteuer und Staatshaushalt	II	k2	Text	f3	Dichotom (0/1)	Eliminiert: Inhalt (fachlich inadäquat)

Tabelle A- 4. Item-Spezifikationen des WBK-T2 nach Inhalt, Format und Kodierung

Item-Nr.	Inhaltselement	Itemtypus	Kognitionsprozess	Darstellungsform	Antwortformat	Antwortkategorien (Scoring)	Status nach Revision: Begründung
AHV_11m	Rentenalter und Arbeitsmarkt	III	K2	Text	f3	Polytom (0/1/2)	Modifiziert: Umfang (Frage, Antwortstruktur), Kodierung
AHV_12m	AHV-Finanzierung: Lohnbeiträge vs. Altersrenten	IV	k3	Tabelle	f3	Polytom (0/1/2)	Modifiziert: Umfang (Frage, Antwortstruktur), Kodierung
AHV_13n	AHV-Kasse: Beiträge und Leistungen	VI	k2	Diagramm	f1 (3 Antwortoptionen)	Dichotom (0/1) → Σ 4	Neu: Inhalt, Format (Diagramm)
Problemsituation: Energieversorgung							
ENE_1	Energieeffizienz	I	k1	Text	f1 (4 Antwortoptionen)	Dichotom (0/1)	Unverändert
ENE_2m	Nicht-erneuerbare Energieträger	V	k1	Text	f2	Dichotom (0/1) → Σ 2	Modifiziert: Inhalt (Frage), Format
ENE_3	Energieproduzierende Anlagen	I	k1	Text	f2	Dichotom (0/1)	Eliminiert: Inhalt (nicht repräsentativ)
ENE_4m	Lenkungsabgabe Energie	I	k1	Text	f1 (4 Antwortoptionen)	Dichotom (0/1)	Modifiziert: Inhalt (Frage, Antwortoption)
ENE_5	Energieverbrauch Sektoren	I	k1	Text	f1 (4 Antwortoptionen)	Dichotom (0/1)	Eliminiert: Inhalt (nicht repräsentativ)
ENE_6	Energieverbrauch Energieträger	I	k1	Text	f1 (4 Antwortoptionen)	Dichotom (0/1)	Eliminiert: Inhalt (nicht repräsentativ)

Tabelle A- 4. Item-Spezifikationen des WBK-T2 nach Inhalt, Format und Kodierung

Item-Nr.	Inhaltselement	Itemtypus	Kognitionsprozess	Darstellungsform	Antwortformat	Antwortkategorien (Scoring)	Status nach Revision: Begründung
ENE_7	Elektrizitätsproduktion Kernenergie	I	k1	Text	f1 (4 Antwortoptionen)	Dichotom (0/1)	Eliminiert: Inhalt (nicht repräsentativ), Item-Kennwerte
ENE_8	Atomausstieg und Elektrizitätspreis	II	k2	Text	f3	Dichotom (0/1)	Unverändert
ENE_9m	Atomausstieg und private Investitionen	I	k1	Text	f2	Dichotom (0/1)	Modifiziert: Format
ENE_10	Energieeffizienz und private Haushalte	III	k2	Text	f3	Polytom (0/1/2)	Eliminiert: Inhalt (nicht repräsentativ)
ENE_11m	Erneuerbare Energien: Wind, Solar, Biogas	IV	k3	Tabelle	f3	Polytom (0/1/2)	Modifiziert: Inhalt (Varianten), Umfang (Frage, Antwortstruktur), Kodierung
ENE_12n	Energieabgabe und Energieverbrauch	VI	k2	Diagramm	f1 (3 Antwortoptionen)	Dichotom (0/1) → Σ 4	Neu: Inhalt, Format (Diagramm)
Problemsituation: Staatsverschuldung							
STA_1	Staatsdefizit	I	k1	Text	f1 (4 Antwortoptionen)	Dichotom (0/1)	Unverändert
STA_2	Schuldenquote	I	k1	Text	f1 (4 Antwortoptionen)	Dichotom (0/1)	Eliminiert: Item-Kennwerte
STA_3	Schuldenquote Euro-Raum	I	k1	Text	f1 (4 Antwortoptionen)	Dichotom (0/1)	Unverändert
STA_4	Ausgabengebiete Bund	I	k1	Text	f1 (4 Antwortoptionen)	Dichotom (0/1)	Unverändert
STA_5	Staatsausgaben und privater Konsum	II	k2	Text	f3	Dichotom (0/1)	Unverändert

Tabelle A- 4. Item-Spezifikationen des WBK-T2 nach Inhalt, Format und Kodierung

Item-Nr.	Inhaltselement	Itemtypus	Kognitionsprozess	Darstellungsform	Antwortformat	Antwortkategorien (Scoring)	Status nach Revision: Begründung
STA_6m	Finanzpolitik und private Investitionen	I	k1	Text	f3	Dichotom (0/1)	Modifiziert: Inhalt (Fachbegriff), Format
STA_7m	Schuldenbremse und Gesamtwirtschaft	III	k2	Text	f3	Polytom (0/1/2)	Modifiziert: Umfang (Frage, Antwortstruktur), Kodierung
STA_8m	Staatsverschuldung: Staatsausgaben vs. Steuereinnahmen	IV	k3	Tabelle	f3	Polytom (0/1/2)	Modifiziert: Umfang (Frage, Antwortstruktur), Kodierung
STA_9n	Staatliche Einnahmequellen	V	k1	Text	f2	Dichotom (0/1) → Σ 2	Neu: Inhalt, Format
STA_10n	Staatsverschuldung und Schuldenquote	VI	k2	Diagramm	f1 (3 Antwortoptionen)	Dichotom (0/1) → Σ 4	Neu: Inhalt, Format (Diagramm)
Problemsituation: Managervergütung							
MAN_1m	Variable Vergütungskomponenten	V	k1	Text	f2	Dichotom (0/1) → Σ 2	Modifiziert: Inhalt (Frage), Format
MAN_2	Managergehälter vs. Mitarbeitergehälter	I	k1	Text	f1 (4 Antwortoptionen)	Dichotom (0/1)	Unverändert
MAN_3	Managervergütungen Entscheid	I	k1	Text	f1 (4 Antwortoptionen)	Dichotom (0/1)	Eliminiert: Inhalt (fachlich inadäquat), Item-Kennwerte
MAN_4m	Aktionärsrechte	V	k1	Text	f2	Dichotom (0/1) → Σ 2	Modifiziert: Format
MAN_5	Reingewinne und Managervergütungen	II	k2	Text	f3	Dichotom (0/1)	Unverändert

Tabelle A- 4. Item-Spezifikationen des WBK-T2 nach Inhalt, Format und Kodierung

Item-Nr.	Inhaltselement	Itemtypus	Kognitions- prozess	Darstellungs- form	Antwortformat	Antwort- kategorien (Scoring)	Status nach Revision: Begründung
MAN_6	Reingewinne und Managerlohnsumme	II	k2	Text	f3	Dichotom (0/1)	Eliminiert: Inhalt (fachlich inadäquat), Item-Kennwerte
MAN_7	Managervergütungen und Unternehmenserfolg	II	k2	Text	f3	Dichotom (0/1)	Eliminiert: Inhalt (nicht repräsentativ)
MAN_8m	Überhöhte Managervergütungen	III	k2	Text	f3	Polytom (0/1/2)	Modifiziert: Inhalt (Fachbegriff), Umfang (Frage, Antwortstruktur)
MAN_9n	Intrinsische Motivation	I	k1	Text	f1 (4 Antwortoptionen)	Dichotom (0/1)	Neu: Inhalt (Fachbegriff), Format
MAN_10n	Managerleistung und Managervergütung	VI	k2	Diagramm	f1 (3 Antwortoptionen)	Dichotom (0/1) → Σ 4	Neu: Inhalt, Format (Diagramm)

Hinweis: rot = eliminiert, orange = modifiziert (z.B. AHV_1m), grün = neu konstruiert (z.B. AHV_13n).

Quelle: Ackermann (2018c).

A1.5 Empfehlungen für Item-Revision im WBK-T2

Tabelle A- 5. Empfehlungen für Item-Revision im WBK-T2 aufgrund klassischer und probabilistischer Itemanalyse sowie fachdidaktischer Expertenbeurteilung

Testteil, Itemtypus	Item-Nr.	Revision	Konkretisierung	Begründung
AHV	AHV_1m	Evtl. modifizieren	Psychometrie: Underfit (= 1.12)	
	AHV_2	Ausschliessen	Inhalt: Fachkonzept („vollendetes 17. Altersjahr“) in Frage Psychometrie: klassische Trennschärfe (< .20)	- Nur mit fachspezifischem Wissen (JUS) oder durch Raten lösbar, nicht konform mit Konstrukt WBK-soek - Geringe inhaltliche Relevanz für Konstrukt WBK-soek
	AHV_3 > AHV_3m	Modifizieren und aufnehmen	Inhalt: Formulierung der Frage Format: Typus I (gebunden) > Typus V (frei)	Inhaltliche Relevanz (modifiziert AHV_3 und ersetzt AHV_13n)
	AHV_5	Evtl. ausschliessen	Psychometrie: klassische Schwierigkeit (> 0.90), DIF Geschlecht (C– zugunsten Mädchen)	Diskrimination aufgrund des Geschlechts vermeiden
	AHV_9 > AHV_9m	Modifizieren und aufnehmen	Inhalt: Formulierung von Antwortoptionen Format: Typus II (frei) > Typus I (gebunden)	Hohe inhaltliche Relevanz für Konstrukt WBK-soek
	AHV_13n	Ausschliessen	Inhalt/Format: durch Raten statt Wissen lösbar Psychometrie: klassische Schwierigkeit (> 0.90), klassische Trennschärfe (< .20), Underfit (= 1.17), Schwellenparameter (ungeordnet)	- Durch Raten lösbar - Inhaltliche Relevanz für Konstrukt WBK-soek: Inhalt integrieren in AHV_3
	AHV_14n	Neukonstruieren	Inhalt: neue Inhaltselemente auf Makroebene (Technologie/Medizin, Demografie, AHV-Kasse) Format: „echtes“ Selected Response mit 3 Antwortoptionen (Diagramme)	- Komplexe Zusammenhänge adäquat abbilden, Inhaltselemente vernetzt statt additiv abfragen - Problemsituation inhaltlich ausweiten

Tabelle A- 5. Empfehlungen für Item-Revision im WBK-T2 aufgrund klassischer und probabilistischer Itemanalyse sowie fachdidaktischer Expertenbeurteilung

Testteil, Itemtypus	Item-Nr.	Revision	Konkretisierung	Begründung
ENE	ENE_4m	Evtl. ausschliessen	Psychometrie: DIF Geschlecht (C+ zugunsten Jungen), DIF Schwerpunkt (C+ zugunsten WuR)	▪ Diskrimination aufgrund des Geschlechts und des Schwerpunktfachs vermeiden
	ENE_9m	Evtl. ausschliessen	Psychometrie: DIF Geschlecht (C+ zugunsten Jungen)	▪ Diskrimination aufgrund des Geschlechts vermeiden
	ENE_12n > ENE_12nm	Modifizieren	Format: „echtes“ Selected Response mit 3 Antwortoptionen (Diagramme) Psychometrie: Schwellenparameter (ungeordnet)	▪ Komplexe Zusammenhänge adäquat abbilden, Inhaltselemente vernetzt statt additiv abfragen
STA	STA_3	Evtl. modifizieren	Psychometrie: klassische Schwierigkeit (> 0.90), klassische Trennschärfe (< .20)	
	STA_6m > STA_6mm	Modifizieren	Inhalt: Fachbegriffe/Fremdwörter in Frage und Antwortoptionen („expansiv“, „stimulieren“) Psychometrie: klassische Trennschärfe (< .20)	▪ Nur mit fachspezifischem Wissen (VWL) oder durch Raten lösbar, nicht konform mit Konstrukt WBK-soek
	STA_10n > STA_10nm	Modifizieren	Format: „echtes“ Selected Response mit 3 Antwortoptionen (Diagramme) Psychometrie: Schwellenparameter (ungeordnet)	▪ Komplexe Zusammenhänge adäquat abbilden, Inhaltselemente vernetzt statt additiv abfragen

Tabelle A- 5. Empfehlungen für Item-Revision im WBK-T2 aufgrund klassischer und probabilistischer Itemanalyse sowie fachdidaktischer Expertenbeurteilung

Testteil, Itemtypus	Item-Nr.	Revision	Konkretisierung	Begründung
MAN	Problemsituation MAN	Ausweiten	Inhalt: thematisch ausweiten („Gehaltsgerechtigkeit“) Einleitungstext thematisch anpassen Items zu neuen Inhaltselementen konstruieren: Mindestlohn und Arbeitsmarkt (Typus III), Gehaltsunterschiede Frau/Mann (Typus IV)	Problemsituation zu eng gefasst für genügend repräsentative und adäquate Inhaltselemente
	Einleitungstext MAN (i.V.m. MAN_1m)	Modifizieren	Inhalt: Fachbegriffe („Leistungszuschlag“, „Erfolgsbeteiligung“) im Einleitungstext integrieren	- Ohne Modifikation: Item MAN_1m ist sonst nur mit fachspezifischem Vorwissen (BWL) lösbar, nicht konform mit Konstrukt WBK-soek - Mit Modifikation: Item MAN_1m ist teilweise mit Infos aus Text lösbar (Merkmal: info2)
	MAN_8m	Ausschliessen	Inhalt: Textverständnis vs. Fachverständnis	- Frage zu eng am Einleitungstext - SuS-Antworten zu repetitiv und zu kurz („Aktionärsrechte“, „1:12“)
	MAN_9n > MAN_9nm	Modifizieren	Inhalt: fachliche Kohärenz/Präzision der Antwortoptionen Psychometrie: klassische Schwierigkeit (< 0.10), klassische Trennschärfe (< .20)	Differenzierung zwischen guten/schlechten SuS erhöhen
	MAN_10n > MAN_10nm	Modifizieren	Format: „echtes“ Selected Response mit 3 Antwortoptionen (Diagramme) Psychometrie: Underfit (= 1.10)	Komplexe Zusammenhänge adäquat abbilden, Inhaltselemente vernetzt statt additiv abfragen
	MAN_11n	Neukonstruieren und aufnehmen	Inhalt: neues Inhaltselement (Mindestlohn und Arbeitsmarkt)	Problemsituation inhaltlich ausweiten
	MAN_12n	Neukonstruieren und aufnehmen	Inhalt: neues Inhaltselement (Gehaltsunterschiede Frau/Mann)	Problemsituation inhaltlich ausweiten

Tabelle A- 5. Empfehlungen für Item-Revision im WBK-T2 aufgrund klassischer und probabilistischer Itemanalyse sowie fachdidaktischer Expertenbeurteilung

Testteil, Itemtypus	Item-Nr.	Revision	Konkretisierung	Begründung
Typus II	AHV_8 ENE_8 STA_6 MAN_5	Modifizieren	Kodierung: von dichotom (Score 1) zu polytom (Score 2)	▪ SuS-Antworten qualitativ differenzierter kodieren (fachliche Korrektheit, Argumentationsstruktur)
Typus IV	AHV_12m ENE_11m STA_8m MAN_12n	Modifizieren	Kodierung: von polytom (Score 2) zu polytom (Score 3)	▪ SuS-Antworten qualitativ (fachliche Korrektheit, Begründungsqualität) und quantitativ (Begründungsumfang/Anzahl Kriterien) differenzierter kodieren

Quelle: Ackermann (2018a).

A1.6 Beispiele für Item-Revision im WBK-T2

Tabelle A- 6. Beispiele für Item-Revision im WBK-T2 aufgrund Itemanalyse und Expertenbeurteilung

Ursprüngliche Items (vgl. Testheft für Erhebung WBKgym)

STA_6m

Welchen Effekt hat eine expansive Finanzpolitik auf die Investitionen der Unternehmen?		
(1)	☐	Sie steigen, weil sie durch die erhöhte Geldmenge stimuliert werden.
(2)	☐	Sie sinken, weil sie durch die erhöhten Staatsausgaben verdrängt werden.
(3)	☒	Sie steigen, weil sie durch die erhöhte Nachfrage angekurbelt werden.
(4)	☐	Sie sinken, weil sie durch die erhöhten Steuern gebremst werden.

MAN_9n

Welcher Anreiz des Unternehmens beeinflusst die intrinsische (d.h. die innere) Motivation des Managers?		
(1)	☐	Anteil am Unternehmenserfolg
(2)	☒	Grad der Verantwortung
(3)	☐	Aussicht auf Karrieremöglichkeiten
(4)	☐	Höhe des Grundgehalts

Modifizierte bzw. neukonstruierte Items (vgl. Empfehlung zur Item-Revision)

STA_6mm

Welchen Effekt haben höhere Staatsausgaben auf die Investitionen der Unternehmen?		
(1)	☐	Sie steigen, weil sie durch die erhöhte Geldmenge angeregt werden.
(2)	☐	Sie sinken, weil sie durch die erhöhten staatlichen Unterstützungen verdrängt werden.
(3)	☒	Sie steigen, weil sie durch die erhöhte gesamtwirtschaftliche Nachfrage angekurbelt werden.
(4)	☐	Sie sinken, weil sie durch die erhöhten Steuern gebremst werden.

MAN_9mn

Was beeinflusst die intrinsische (d.h. die innere) Motivation des Managers?		
(1)	☐	Die Beteiligung am Unternehmensgewinn.
(2)	☒	Die Freude an der Arbeitstätigkeit.
(3)	☐	Die Aussicht auf Karrieremöglichkeiten.
(4)	☐	Die Höhe des Grundgehalts. / Die Beziehung zu Geschäftspartnern.

MAN_11n

Wie wirkt sich ein gesetzlicher Mindestlohn auf den Arbeitsmarkt aus? Erläutern Sie eine Auswirkung.	
Erläuterung:	

Tabelle A- 6. Beispiele für Item-Revision im WBK-T2 aufgrund Itemanalyse und Expertenbeurteilung

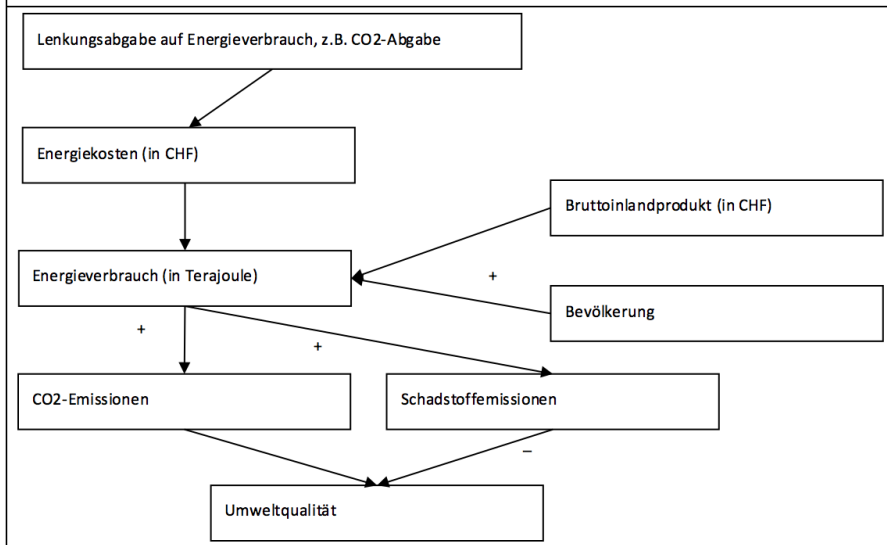
Ursprüngliche Items (vgl. Testheft für Erhebung WBKgytm)

ENE_12n

Welcher Zusammenhang besteht zwischen einer Lenkungsabgabe auf Energie, den Kosten des Energieverbrauchs und der Umweltqualität?

Ergänzen Sie im Diagramm bei den vier Pfeilen die folgenden Vorzeichen: +, 0, -.

Erläuterung: Das Diagramm zeigt die Beziehung (durch Pfeile mit Vorzeichen) zwischen zwei Begriffen (in Kästchen).
 0 : keine Beziehung zwischen X und Y, d.h. wenn X steigt, dann bleibt Y unverändert; wenn X sinkt, bleibt Y unverändert.
 + : positive Beziehung zwischen X und Y, d.h. wenn X steigt, dann steigt Y; wenn X sinkt, dann sinkt Y.
 - : negative Beziehung zwischen X und Y, d.h. wenn X steigt, sinkt Y; wenn X sinkt, dann steigt Y.

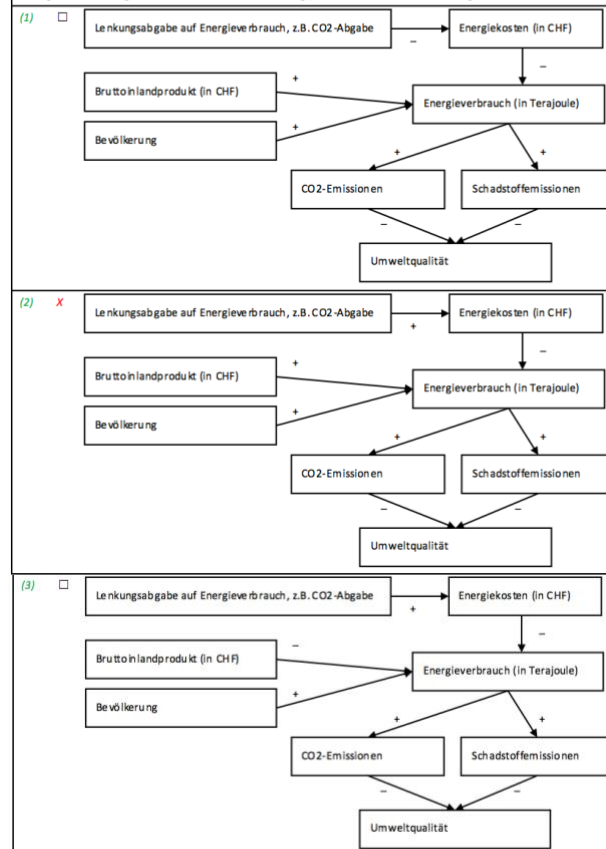


Modifizierte bzw. neukonstruierte Items (vgl. Empfehlung zur Item-Revision)

ENE_12nm

ENE_12nm
 Welcher Zusammenhang besteht zwischen einer Lenkungsabgabe auf Energie, den Kosten des Energieverbrauchs und der Umweltqualität? Kreuzen Sie das korrekte Diagramm an.

Erläuterung: Das Diagramm zeigt die Beziehung (durch Pfeile mit Vorzeichen) zwischen zwei Begriffen (in Kästchen).
 0 : keine Beziehung zwischen X und Y, d.h. wenn X steigt, dann bleibt Y unverändert; wenn X sinkt, bleibt Y unverändert.
 + : positive Beziehung zwischen X und Y, d.h. wenn X steigt, dann steigt Y; wenn X sinkt, dann sinkt Y.
 - : negative Beziehung zwischen X und Y, d.h. wenn X steigt, sinkt Y; wenn X sinkt, dann steigt Y.



A2 Testauswertung des WBK-T2

A2.1 Item-Kennwerte des WBK-T2

Tabelle A- 7. Item-Kennwerte des WBK-T2 nach klassischen und probabilistischen Analysen

Item-Nr.	Score	Klassische Kennwerte				Probabilistische Kennwerte		
		Lösungshäufigkeit (in %)	Item-Schwierigkeit: M_{adj} bzw. M_{cor} (SD)	Item-Trennschärfe: $r_{it-total}$ $r_{it-rest}$	Kategorien- Trennschärfe (r_{ptbis})	Itemparameter σ (in Logits) (SE σ)	Item-Infit (wMNSQ)	t-Statistik für wMNSQ
AHV_1m	---	58.4	0.68 (0.42)	0.31 0.19	---	-0.001 (0.048)	1.12 (0.89, 1.11)	2.1
	1	17.3	---	---	0.00	-0.26	1.01 (0.83, 1.17)	0.1
	2	58.4	---	---	0.16	0.26	1.10 (0.92, 1.08)	2.5
AHV_2	1	31.5	0.09 (0.62)	0.20 0.13	---	1.356 (0.059)	1.03 (0.91, 1.09)	0.6
AHV_5	1	97.3	0.96 (0.22)	0.20 0.18	---	-3.213 (0.067)	0.98 (0.41, 1.59)	0.0
AHV_6	1	82.1	0.76 (0.51)	0.34 0.29	---	-1.092 (0.062)	0.95 (0.83, 1.17)	-0.6
AHV_7	1	62.4	0.50 (0.65)	0.23 0.16	---	-0.017 (0.058)	1.02 (0.93, 1.07)	0.6
AHV_8	1	74.4	0.75 (0.43)	0.38 0.32	---	-0.611 (0.060)	0.95 (0.88, 1.12)	-0.9
AHV_11	---	58.9	0.71 (0.39)	0.39 0.28	---	-0.162 (0.050)	1.02 (0.88, 1.12)	0.4
	1	22.7	---	---	-0.10	-0.54	1.00 (0.87, 1.13)	0.0
	2	58.9	---	---	0.26	0.22	1.01 (0.92, 1.08)	0.3
AHV_12m	---	15.5	0.49 (0.29)	0.32 0.07	---	0.620 (0.055)	1.00 (0.86, 1.14)	0.1
	1	66.1	---	---	0.02	-0.93	1.00 (0.93, 1.07)	-0.0
	2	15.5	---	---	0.18	2.17	0.99 (0.81, 1.19)	-0.0

Tabelle A- 7. Item-Kennwerte des WBK-T2 nach klassischen und probabilistischen Analysen

Item-Nr.	Score	Klassische Kennwerte				Probabilistische Kennwerte		
		Lösungshäufigkeit (in %)	Item-Schwierigkeit: M _{adj} bzw. M _{cor} (SD)	Item-Trennschärfe: r _{it-total} r _{it-rest}	Kategorien- Trennschärfe (r _{ptbis})	Itemparameter σ (in Logits) (SE σ)	Item-Infit (wMNSQ)	t-Statistik für wMNSQ
AHV_13n	---	91.2	0.94 (0.22)	0.16 0.07	---	-1.229 (0.056)	1.17 (0.63, 1.37)	0.9
	1	0.8	---	---	0.00	-1.70	1.04 (0.00, 2.08)	0.3
	2	4.8	---	---	-0.10	-1.37	1.03 (0.60, 1.40)	0.2
	3	2.7	---	---	0.01	-0.91	1.01 (0.41, 1.59)	0.1
	4	91.2	---	---	0.07	-0.83	1.12 (0.73, 1.27)	0.9
ENE_1	1	77.1	0.73 (0.54)	0.31 0.25	---	-0.763 (0.060)	0.97 (0.87, 1.13)	-0.4
ENE_2m	---	39.5	0.62 (0.37)	0.40 0.30	---	0.163 (0.050)	1.02 (0.89, 1.11)	0.4
	1	39.5	---	---	0.02	-0.55	1.01 (0.95, 1.05)	0.3
	2	39.5	---	---	0.18	0.88	1.04 (0.93, 1.07)	1.2
ENE_4m	1	63.5	0.56 (0.63)	0.33 0.26	---	-0.065 (0.059)	0.98 (0.93, 1.07)	-0.5
ENE_8	1	67.5	0.71 (0.71)	0.44 0.38	---	-0.253 (0.059)	0.92 (0.91, 1.09)	-1.9
ENE_9m	1	82.7	0.82 (0.46)	0.40 0.35	---	-1.129 (0.062)	0.93 (0.83, 1.17)	-0.7
ENE_11m	---	26.9	0.46 (0.40)	0.44 0.33	---	0.764 (0.314)	1.00 (0.90, 1.10)	0.0
	1	32.5	---	---	0.02	0.23	0.99 (0.92, 1.08)	-0.1
	2	26.9	---	---	0.29	1.30	0.99 (0.89, 1.11)	-0.2
ENE_12n	---	25.3	0.52 (0.41)	0.60 0.46	---	0.159 (0.040)	0.96 (0.87, 1.13)	-0.6
	1	10.9	---	---	-0.19	-0.78	1.01 (0.76, 1.24)	0.1
	2	22.4	---	---	-0.12	-0.25	1.00 (0.87, 1.13)	0.0
	3	32.0	---	---	0.09	0.32	1.01 (0.92, 1.08)	0.4
	4	25.3	---	---	0.35	1.30	0.96 (0.88, 1.12)	-0.7
STA_1	1	72.5	0.63 (0.60)	0.23 0.17	---	-0.510 (0.060)	1.02 (0.89, 1.11)	0.3
STA_3	1	90.1	0.87 (0.40)	0.13 0.09	---	-1.803 (0.064)	1.02 (0.74, 1.26)	0.2

Tabelle A- 7. Item-Kennwerte des WBK-T2 nach klassischen und probabilistischen Analysen

Item-Nr.	Score	Klassische Kennwerte				Probabilistische Kennwerte		
		Lösungshäufigkeit (in %)	Item-Schwierigkeit: M _{adj} bzw. M _{cor} (SD)	Item-Trennschärfe: r _{it-total} r _{it-rest}	Kategorien- Trennschärfe (r _{ptbis})	Itemparameter σ (in Logits) (SE σ)	Item-Infit (wMNSQ)	t-Statistik für wMNSQ
STA_4	1	57.6	0.44 (0.66)	0.24 0.17	---	0.197 (0.058)	1.02 (0.94, 1.06)	0.8
STA_5	1	48.5	0.49 (0.50)	0.40 0.33	---	0.588 (0.058)	0.94 (0.95, 1.05)	-2.2
STA_6m	1	37.3	0.17 (0.65)	0.11 0.03	---	1.078 (0.058)	1.09 (0.93, 1.07)	2.4
STA_7	---	19.7	0.43 (0.37)	0.34 0.23	---	0.898 (0.051)	1.04 (0.89, 1.11)	0.8
	1	42.4	---	---	-0.15	0.10	1.03 (0.96, 1.04)	1.8
	2	19.7	---	---	0.31	1.69	0.95 (0.85, 1.15)	-0.6
STA_8m	---	4.0	0.29 (0.29)	0.42 0.34	---	1.877 (0.056)	0.95 (0.88, 1.12)	-0.9
	1	48.8	---	---	0.21	0.42	0.99 (0.96, 1.04)	-0.5
	2	4.0	---	---	0.22	3.34	0.97 (0.54, 1.46)	-0.0
STA_9n	---	43.7	0.70 (0.28)	0.47 0.40	---	-0.714 (0.056)	0.93 (0.88, 1.12)	-1.2
	1	51.7	---	---	-0.26	-2.20	0.96 (0.96, 1.04)	-2.0
	2	43.7	---	---	0.36	0.77	0.94 (0.94, 1.06)	-2.2
STA_10n	---	45.1	0.74 (0.33)	0.47 0.34	---	-0.513 (0.046)	1.01 (0.83, 1.17)	0.1
	1	3.2	---	---	-0.27	-1.41	0.94 (0.49, 1.51)	-0.2
	2	15.2	---	---	-0.08	-1.00	1.04 (0.81, 1.19)	0.4
	3	34.4	---	---	0.00	-0.36	1.02 (0.93, 1.07)	0.6
	4	45.1	---	---	0.32	0.66	1.05 (0.93, 1.07)	1.4
MAN_1m	---	2.7	0.13 (0.25)	0.34 0.27	---	2.422 (0.058)	0.98 (0.80, 1.20)	-0.1
	1	19.5	---	---	0.23	1.71	0.98 (0.85, 1.15)	-0.2
	2	2.7	---	---	0.14	3.14	1.02 (0.42, 1.58)	0.2
MAN_2	1	70.9	0.62 (0.60)	0.25 0.18	---	-0.425 (0.059)	1.01 (0.90, 1.10)	0.3
MAN_4m	---	47.2	0.71 (0.32)	0.46 0.37	---	-0.264 (0.052)	0.95 (0.88, 1.12)	-0.7

Tabelle A- 7. Item-Kennwerte des WBK-T2 nach klassischen und probabilistischen Analysen

Item-Nr.	Score	Klassische Kennwerte				Probabilistische Kennwerte		
		Lösungshäufigkeit (in %)	Item-Schwierigkeit: M_{adj} bzw. M_{cor} (SD)	Item-Trennschärfe: $r_{it-total}$ $r_{it-rest}$	Kategorien- Trennschärfe (r_{ptbis})	Itemparameter σ (in Logits) (SE σ)	Item-Infit (wMNSQ)	t-Statistik für wMNSQ
	1	41.1	---	---	-0.21	-1.14	0.98 (0.95, 1.05)	-0.7
	2	47.2	---	---	0.36	0.61	0.95 (0.94, 1.06)	-1.7
MAN_5	1	52.8	0.53 (0.50)	0.29 0.22	---	0.406 (0.058)	1.00 (0.95, 1.05)	0.1
MAN_8	---	51.5	0.72 (0.33)	0.46 0.36	---	-0.198 (0.051)	0.97 (0.88, 1.12)	-0.4
	1	33.9	---	---	-0.85	-0.85	1.00 (0.93, 1.07)	-0.1
	2	51.5	---	---	0.46	0.46	0.99 (0.94, 1.06)	-0.2
MAN_9n	1	7.5	-0.23 (0.35)	0.06 0.02	---	3.168 (0.065)	1.03 (0.68, 1.32)	0.2
MAN_10n	---	51.2	0.81 (0.24)	0.38 0.25	---	-0.735 (0.046)	1.10 (0.84, 1.16)	1.2
	1	4.8	---	---	-0.17	-1.98	1.00 (0.61, 1.39)	0.1
	2	14.1	---	---	-0.14	-1.04	1.02 (0.80, 1.20)	0.3
	3	28.8	---	---	-0.04	-0.36	1.01 (0.90, 1.10)	0.1
	4	51.2	---	---	0.22	0.45	1.05 (0.93, 1.07)	1.3

Hinweis: Alle Kennwerte basierend auf IRT-Analysen mit dem Partial-Credit-Rasch-Modell und mit ACER ConQuest Version 4 gerechnet (Adams, Wu, & Wilson, 2015).

Klassische Item-Schwierigkeit: „Normaler“ Mittelwert (M) für dichotome Items (M), adjustierter Mittelwert (M_{adj}) für polytome Items, korrigierter Mittelwert (M_{cor}) für Selected-Response-Items.

Klassische Item-Trennschärfe: Unkorrigierte Trennschärfe ($r_{item-total}$), korrigierte Trennschärfe ($r_{item-rest}$); Toleranzwert: $r_{item-total} < .30$, $r_{item-rest} < .30$. Klassische Kategorien-Trennschärfe: Punkt-biseriale Korrelation (r_{ptbis}). Kriterium: r_{ptbis} aufsteigend nach Antwortkategorien.

Item-Infit: weighted mean squares (wMNSQ); Toleranzwert: $0.80 \leq wMNSQ \leq 1.20$. t-Statistik für Abweichung von wMNSQ = 1; Toleranzwert: $|t| > 1.96$ bzw. $p < .05$.

Quelle: Ackermann (2018c).

A2.2 DIF-Kennwerte des WBK-T2

Tabelle A- 8. DIF-Kennwerte des WBK-T2

Item-Nr.	Schwerpunktfach				Biologisches Geschlecht			
	Interaktions-Parameter δ (in Logits) (SE δ)	z-Wert	DIF (in Logits)	DIF Kategorie	Interaktions-Parameter δ (in Logits) (SE δ)	z-Wert	DIF (in Logits)	DIF Kategorie
AHV_1m	-0.254 (0.070)	-3.629	-0.508	B–	-0.223 (0.072)	-3.097	-0.446	B–
AHV_2	-0.172 (0.114)	-1.509	-0.344	A	-0.106 (0.117)	-0.906	-0.212	A
AHV_5	0.031 (0.323)	0.096	0.062	A	-0.619 (0.315)	-1.965	-1.238	C–
AHV_6	0.185 (0.148)	1.250	0.370	A	0.255 (0.153)	1.667	0.510	A
AHV_7	-0.171 (0.110)	-1.555	-0.342	A	-0.228 (0.112)	-2.036	-0.456	B–
AHV_8	0.025 (0.125)	0.200	0.050	A	-0.179 (0.124)	-1.444	-0.358	A
AHV_11	-0.152 (0.077)	-1.974	-0.304	A	-0.102 (0.080)	-1.275	-0.204	A
AHV_12m	-0.049 (0.094)	-0.521	-0.098	A	-0.238 (0.094)	-2.532	-0.476	B–
AHV_13n	-0.328 (0.179)	-1.832	-0.656	A	-0.169 (0.185)	-0.914	-0.338	A
ENE_1	-0.064 (0.128)	-0.500	-0.128	A	-0.036 (0.130)	-0.277	-0.072	A
ENE_2m	-0.110 (0.077)	-1.429	-0.220	A	0.149 (0.082)	1.817	0.298	A
ENE_4m	0.516 (0.123)	4.195	1.032	C+	0.514 (0.126)	4.079	1.028	C+
ENE_8	0.158 (0.118)	1.339	0.316	A	0.217 (0.122)	1.779	0.434	A
ENE_9m	0.112 (0.147)	0.762	0.224	A	0.387 (0.163)	2.374	0.774	C+
ENE_11m	-0.199 (n.a.)	n.a.	-0.398	A	-0.073 (n.a.)	n.a.	-0.146	A
ENE_12n	0.054 (0.068)	0.794	0.108	A	0.280 (0.099)	2.828	0.560	B+
STA_1	0.074 (0.123)	0.602	0.148	A	0.090 (0.126)	0.714	0.180	A
STA_3	-0.328 (0.172)	-1.907	-0.656	A	-0.017 (0.182)	-0.093	-0.034	A
STA_4	0.311 (0.113)	2.752	0.622	B+	-0.256 (0.110)	-2.327	-0.512	B–
STA_5	0.207 (0.109)	1.899	0.414	A	-0.017 (0.110)	-0.155	-0.034	A

Tabelle A- 8. DIF-Kennwerte des WBK-T2

Item-Nr.	Schwerpunktfach				Biologisches Geschlecht			
	Interaktions-Parameter δ (in Logits) (SE δ)	z-Wert	DIF (in Logits)	DIF Kategorie	Interaktions-Parameter δ (in Logits) (SE δ)	z-Wert	DIF (in Logits)	DIF Kategorie
STA_6m	-0.213 (0.110)	-1.936	-0.426	A	-0.065 (0.112)	-0.580	-0.130	A
STA_7	-0.120 (0.077)	-1.558	-0.240	A	-0.217 (0.079)	-2.747	-0.434	B–
STA_8m	-0.059 (0.135)	-0.437	-0.118	A	-0.087 (0.138)	-0.630	-0.174	A
STA_9n	0.091 (0.147)	0.619	0.182	A	0.298 (0.161)	1.851	0.596	A
STA_10n	0.014 (0.136)	0.103	0.028	A	0.044 (0.111)	0.396	0.088	A
MAN_1m	-0.045 (0.160)	-0.281	-0.090	A	0.302 (0.170)	1.776	0.604	A
MAN_2	0.020 (0.120)	0.167	0.040	A	-0.094 (0.120)	-0.783	-0.188	A
MAN_4m	0.304 (0.108)	2.815	0.608	B+	0.130 (0.101)	1.287	0.260	A
MAN_5	-0.212 (0.107)	-1.981	-0.424	A	-0.042 (0.110)	-0.382	-0.084	A
MAN_8	0.337 (0.201)	1.677	0.674	A	-0.089 (0.087)	-1.023	-0.178	A
MAN_9n	0.112 (0.095)	1.179	0.224	A	0.211 (0.195)	1.082	0.422	A
MAN_10n	-0.078 (0.145)	-0.538	-0.156	A	-0.022 (0.148)	-0.149	-0.044	A

Hinweis: Alle Kennwerte basierend auf IRT-Analysen mit dem Partial-Credit-Rasch-Modell und mit ACER ConQuest Version 4 gerechnet (Adams, Wu, & Wilson, 2015).

Schwerpunktfach: Interaktions-Parameter δ = Item*Schwerpunktfach. $\delta > 0$ Logits: DIF zugunsten Fokusgruppe WuR. $\delta < 0$ Logits: DIF zulasten Fokusgruppe WuR.

biologisches Geschlecht: Interaktions-Parameter δ = Item*Geschlecht: $\delta > 0$ = DIF zugunsten Fokusgruppe Jungen. $\delta < 0$ = DIF zulasten Fokusgruppe Jungen.

ETS DIF-Kategorien: A = $|DIF| \leq 0.426$ Logits, vernachlässigbarer Effekt („negligible“). B = $0.426 \text{ Logits} < |DIF| < 0.638$ Logits, moderater Effekt ("medium"),

C = $|DIF| \geq 0.638$ Logits, grosser Effekt ("large").

Quelle: Ackermann (2018c).

A2.3 Kategorien-Kennwerte der Selected-Response-Items

Tabelle A- 9. Kategorien-Kennwerte der Selected-Response-Items (Typus I)

Item-Nr.	Antwort-kategorie	Score	Antworthäufigkeit (in %)	Punkt-biseriale Korrelation (r_{ptbis})	Item-Mittelwert korrigiert: M_{cor} (SD)
AHV_2	1	1	31.5	0.13	0.09 (0.62)
	2	0	57.9	-0.17	
	3	0	1.6	0.03	
	4	0	8.8	0.10	
AHV_5	1	0	1.1	-0.13	0.96 (0.22)
	2	0	0.5	-0.15	
	3	1	97.3	0.18	
	4	0	1.1	-0.05	
AHV_6	1	1	82.1	0.29	0.76 (0.51)
	2	0	11.5	-0.23	
	3	0	5.6	-0.13	
	4	0	0.5	-0.04	
AHV_7	1	0	6.9	-0.15	0.50 (0.65)
	2	0	14.9	-0.12	
	3	0	15.7	0.01	
	4	1	62.4	0.16	
ENE_1	1	0	5.3	-0.09	0.73 (0.54)
	2	0	2.1	-0.14	
	3	1	77.1	0.25	
	4	0	12.5	-0.09	
ENE_4m	1	0	4.5	-0.08	0.56 (0.63)
	2	0	18.9	-0.21	
	3	1	63.5	0.26	
	4	0	8.3	0.08	
ENE_9m	1	1	82.7	0.35	0.82 (0.46)
	2	0	6.9	-0.08	
	3	0	3.7	-0.18	
	4	0	2.4	-0.14	
STA_1	1	0	8.3	-0.08	0.63 (0.60)
	2	1	72.5	0.17	
	3	0	13.1	-0.10	
	4	0	6.1	-0.08	
STA_3	1	0	1.1	-0.09	0.87 (0.40)
	2	1	90.1	0.09	
	3	0	1.9	-0.01	
	4	0	6.9	-0.06	

Tabelle A- 9. Kategorien-Kennwerte der Selected-Response-Items (Typus I)

Item-Nr.	Antwort-kategorie	Score	Antworthäufigkeit (in %)	Punkt-biseriale Korrelation (r_{ptbis})	Item-Mittelwert korrigiert: M_{cor} (SD)
STA_4	1	0	20.5	-0.17	0.44 (0.66)
	2	0	5.9	-0.10	
	3	0	15.2	0.02	
	4	1	57.6	0.17	
STA_6m	1	0	31.5	0.15	0.17 (0.65)
	2	0	14.4	-0.23	
	3	1	37.3	0.03	
	4	0	16.5	0.00	
MAN_2	1	1	70.9	0.18	0.62 (0.60)
	2	0	4.3	0.03	
	3	0	18.9	-0.15	
	4	0	5.3	-0.11	
MAN_9	1	0	56.5	0.14	-0.23 (0.35)
	2	1	7.5	0.02	
	3	0	9.9	0.01	
	4	0	25.9	-0.17	

Quelle: Ackermann (2018c).

Tabelle A- 10. Kategorien-Kennwerte der Selected-Response-Items (Typus VI)

Item-Nr.	Score	Antworthäufigkeit (in %)	Punkt-biserialer Korrelation (r_{ptbis})	Item-Mittelwert korrigiert: M_{cor} (SD)
AHV_13n	1	0.8	0.00	0.94 (0.22)
	2	4.8	-0.10	
	3	2.7	0.01	
	4	91.2	0.07	
ENE_12n	1	10.9	-0.19	0.52 (0.41)
	2	22.4	-0.12	
	3	32.0	0.09	
	4	25.3	0.35	
STA_10n	1	3.2	-0.27	0.74 (0.33)
	2	15.2	-0.08	
	3	34.4	0.00	
	4	45.1	0.32	
MAN_10n	1	4.8	-0.17	0.81 (0.24)
	2	14.1	-0.14	
	3	28.8	-0.04	
	4	51.2	0.22	

Quelle: Ackermann (2018c).

A2.4 Kategorien-Kennwerte der Constructed-Response-Items

Tabelle A- 11. Kategorien-Kennwerte der Short-Constructed-Response-Items (Typus V)

Item-Nr.	Score	Antworthäufigkeit (in %)	Punkt-biseriale Korrelation (r_{ptbis})	Item-Mittelwert korrigiert: M_{cor} (SD)
AHV_1m	0			
	1	17.3	0.00	0.68 (0.42)
	2	58.4	0.16	
ENE_2m	0			
	1	39.5	0.04	0.62 (0.37)
	2	39.5	0.21	
STA_9n	0			
	1	51.7	-0.26	0.70 (0.28)
	2	43.7	0.36	
MAN_1m	0			
	1	19.5	0.23	0.13 (0.25)
	2	2.7	0.14	
MAN_4m	0			
	1	41.1	-0.21	0.71 (0.32)
	2	47.2	0.36	

Quelle: Ackermann (2018c).

Tabelle A- 12. Kategorien-Kennwerte der Extended-Constructed-Response-Items (Typus II, III, IV)

Item-Nr.	Score	Antworthäufigkeit (in %)	Punkt-biseriale Korrelation (r_{ptbis})	Item-Mittelwert adjustiert: M_{adj} (SD)
AHV_8	0	25.6	-0.32	0.75 (0.43)
	1	74.4	0.32	
AHV_11	0	18.4	-0.23	0.71 (0.39)
	1	22.7	-0.10	
	2	58.9	0.26	
AHV_12m	0	18.4	-0.19	0.49 (0.29)
	1	66.1	0.02	
	2	15.5	0.18	
ENE_8	0	32.5	-0.38	0.71 (0.46)
	1	67.5	0.38	
ENE_11m	0	40.5	-0.29	0.46 (0.40)
	1	32.5	0.02	
	2	26.9	0.29	
STA_5	0	51.5	-0.33	0.49 (0.50)
	1	48.5	0.33	
STA_7	0	37.9	-0.10	0.43 (0.37)
	1	42.4	-0.15	
	2	19.7	0.31	
STA_8m	0	47.2	-0.30	0.29 (0.29)
	1	48.8	0.21	
	2	4.0	0.22	
MAN_5	0	47.2	-0.22	0.53 (0.50)
	1	52.8	0.22	
MAN_8	0	14.7	-0.31	0.72 (0.33)
	1	33.9	-0.09	
	2	51.5	0.31	

Quelle: Ackermann (2018c).

A2.5 Personenparameter und Personen-Testwerte

Tabelle A- 13. Personenparameter und Personen-Testwerte der Stichprobe WBKgym im WBK-T2

Testwert (Summenwert)	Lösungsquote (%)	WLE (in Logits)	SE WLE	Lower CI 95 (in Logits)	Upper CI 95 (in Logits)	t-Statistik
0	0.0	-5.113	1.511	-8.075	-2.152	-3.384
1	1.8	-3.897	0.887	-5.635	-2.159	-4.394
2	3.6	-3.296	0.692	-4.653	-1.939	-4.760
3	5.5	-2.890	0.586	-4.039	-1.741	-4.928
4	7.3	-2.585	0.517	-3.598	-1.572	-5.003
5	9.1	-2.342	0.466	-3.256	-1.429	-5.025
6	10.9	-2.142	0.427	-2.980	-1.304	-5.011
7	12.7	-1.972	0.397	-2.750	-1.194	-4.968
8	14.5	-1.826	0.373	-2.556	-1.095	-4.900
9	16.4	-1.696	0.353	-2.388	-1.005	-4.808
10	18.2	-1.580	0.337	-2.240	-0.920	-4.693
11	20.0	-1.475	0.324	-2.109	-0.841	-4.558
12	21.8	-1.377	0.313	-1.991	-0.764	-4.400
13	23.6	-1.286	0.304	-1.882	-0.689	-4.224
14	25.5	-1.199	0.297	-1.782	-0.616	-4.030
15	27.3	-1.115	0.292	-1.688	-0.543	-3.821
16	29.1	-1.035	0.288	-1.598	-0.471	-3.598
17	30.9	-0.956	0.284	-1.513	-0.399	-3.363
18	32.7	-0.878	0.282	-1.430	-0.326	-3.119
19	34.5	-0.802	0.280	-1.350	-0.253	-2.866
20	36.4	-0.726	0.278	-1.271	-0.180	-2.608
21	38.2	-0.650	0.277	-1.193	-0.106	-2.344
22	40.0	-0.575	0.277	-1.117	-0.032	-2.076
23	41.8	-0.499	0.277	-1.041	0.043	-1.805
24	43.6	-0.424	0.277	-0.966	0.118	-1.532
25	45.5	-0.348	0.277	-0.891	0.195	-1.257
26	47.3	-0.272	0.278	-0.817	0.272	-0.981
27	49.1	-0.196	0.279	-0.742	0.350	-0.704
28	50.9	-0.119	0.280	-0.668	0.429	-0.426
29	52.7	-0.042	0.282	-0.594	0.510	-0.148
30	54.5	0.037	0.284	-0.519	0.593	0.130
31	56.4	0.117	0.286	-0.444	0.678	0.408
32	58.2	0.198	0.289	-0.368	0.765	0.686
33	60.0	0.282	0.293	-0.292	0.855	0.963
34	61.8	0.367	0.297	-0.214	0.948	1.238

Tabelle A- 13. Personenparameter und Personen-Testwerte der Stichprobe WBKgym im WBK-T2

Testwert (Summenwert)	Lösungsquote (%)	WLE (in Logits)	SE WLE	Lower CI 95 (in Logits)	Upper CI 95 (in Logits)	t-Statistik
35	63.6	0.455	0.301	-0.135	1.045	1.512
36	65.5	0.546	0.306	-0.054	1.146	1.784
37	67.3	0.640	0.312	0.029	1.251	2.053
38	69.1	0.738	0.318	0.115	1.362	2.320
39	70.9	0.841	0.325	0.203	1.479	2.584
40	72.7	0.948	0.334	0.294	1.602	2.843
41	74.5	1.062	0.343	0.390	1.733	3.098
42	76.4	1.182	0.353	0.490	1.874	3.348
43	78.2	1.310	0.365	0.595	2.024	3.593
44	80.0	1.447	0.378	0.706	2.187	3.830
45	81.8	1.594	0.393	0.825	2.363	4.061
46	83.6	1.754	0.410	0.951	2.557	4.282
47	85.5	1.930	0.430	1.088	2.772	4.492
48	87.3	2.123	0.453	1.236	3.011	4.689
49	89.1	2.340	0.481	1.398	3.282	4.867
50	90.9	2.586	0.515	1.576	3.596	5.017
51	92.7	2.870	0.560	1.772	3.968	5.124
52	94.5	3.210	0.622	1.991	4.430	5.158
53	96.4	3.642	0.719	2.233	5.051	5.066
54	98.2	4.255	0.903	2.485	6.024	4.713
55	100.0	5.463	1.514	2.495	8.430	3.608

Quelle: Ackermann (2018c).

A2.6 Rating-Kennwerte für schwierigkeitsbestimmende Merkmale

Tabelle A- 14. Rating-Kennwerte für schwierigkeitsbestimmende Merkmale im WBK-T2

Item-Nr.	Schwierigkeitsbestimmendes Merkmal			
	Lösungsoffenheit	Wissensart	Informationsvorgabe	Inhaltliche Komplexität
AHV_1m	0.60	1.00	1.00	1.00
AHV_2	1.00	1.00	0.60	0.80
AHV_5	1.00	1.00	1.00	1.00
AHV_6	1.00	1.00	0.60	0.80
AHV_7	1.00	1.00	0.80	1.00
AHV_8	1.00	0.80	1.00	0.60
AHV_11	1.00	0.80	1.00	0.60
AHV_12	1.00	0.80	1.00	0.80
AHV_13n	0.80	0.80	0.80	0.40
ENE_1	1.00	0.80	0.60	1.00
ENE_2m	0.60	1.00	0.60	1.00
ENE_4m	1.00	0.80	0.80	1.00
ENE_8	1.00	0.80	1.00	0.40
ENE_9m	1.00	0.60	0.80	0.60
ENE_11	1.00	0.80	0.80	0.60
ENE_12n	1.00	0.60	0.80	0.80
STA_1	1.00	0.80	0.80	1.00
STA_3	1.00	1.00	1.00	1.00
STA_4	1.00	1.00	0.60	1.00
STA_5	1.00	0.80	0.80	1.00
STA_6m	1.00	0.80	1.00	0.60
STA_7	1.00	0.80	1.00	1.00
STA_8	1.00	0.80	1.00	0.80
STA_9n	1.00	1.00	0.80	1.00
STA_10n	1.00	1.00	0.60	0.80
MAN_1m	1.00	1.00	0.80	0.80
MAN_2	1.00	1.00	0.80	0.80
MAN_4m	1.00	1.00	0.80	1.00
MAN_5	0.60	0.80	0.60	0.80
MAN_8	1.00	0.60	0.80	0.60
MAN_9n	1.00	1.00	0.80	0.60
MAN_10n	1.00	1.00	1.00	0.80

Hinweis: N = 5 Expert*innen, m = 32 Items. Inter-Rater-Agreement nach der Holsti-Formel: rot ≤ 60 %, orange = 80 %.

Quelle: Ackermann (2018c).

A3 Testadministration von WBKgym

A3.1 Anfrage an Schulleiter*innen (Vorlage)



Universität
Zürich

Institut für Erziehungswissenschaft

Universität Zürich
Institut für Erziehungswissenschaft (IFE)
Lehrstuhl für Gymnasialpädagogik
Kantonsschulstr. 3
CH-8001 Zürich

www.ife.uzh.ch

Universität Zürich, Institut für Erziehungswissenschaft (IFE),
Kantonsschulstr. 3, CH-8001 Zürich

Frau/Herr Rektorin/Rektor

XX

Kantonsschule YY

Adresse

PLZ Ort

Nicole Ackermann, lic.oec.publ.,
Lehrdiplom für Maturitätsschulen
Wissenschaftliche Assistentin/Doktorandin
Telefon +41 44 634 66 89
nicole.ackermann@ife.uzh.ch

Zürich, 31. Oktober 2016

Forschungsprojekt WBKgym: Anfrage für Erhebung

Sehr geehrte Frau XX/Sehr geehrter Herr XX

Ich bin wissenschaftliche Assistentin am Lehrstuhl von Prof. Dr. Franz Eberle der Universität Zürich und gelange mit der Bitte an Sie, an Ihrer Schule eine Erhebung für mein Dissertationsprojekt durchführen zu dürfen. Dabei geht es um Folgendes:

Das finale Bildungsziel der gymnasialen Maturitätsschulen in der Schweiz ist gemäss Art. 5 MAR die allgemeine Studierfähigkeit („Hochschulstudium“) und die vertiefte Gesellschaftsreife („anspruchsvolle Aufgaben in der Gesellschaft“) der Schülerinnen und Schüler. Zu den anspruchsvollen gesellschaftlichen Aufgaben gehört u.a. die Bewältigung *komplexer sozio-ökonomischer Anforderungssituationen*, mit denen jeder Mensch als Konsument, Sparer/Anleger, Arbeitnehmer, Steuerzahler, Staatsbürger, etc. im Alltag konfrontiert ist. Dazu benötigt der junge Mensch *wirtschaftsbürgerliche Bildung*, die an den gymnasialen Maturitätsschulen der Schweiz vor allem im Fach „Wirtschaft & Recht“ (WIRE) vermittelt wird. Der Umfang der wirtschaftsbürgerlichen Kompetenz von Gymnasiasten und deren Zusammenhang mit dem WIRE-Unterricht am Gymnasium ist jedoch noch weitgehend unerforscht.

Mein Dissertationsprojekt «Wirtschaftsbürgerliche Kompetenz Deutschschweizer Gymnasiasten (WBKgym)» hat zum Ziel, die wirtschaftsbürgerliche Kompetenz von Deutschschweizer Gymnasiasten und ihre individuellen (z.B. Schulnoten), schulischen (z.B. Unterrichtswahrnehmung) und ausser-schulischen Kontextfaktoren (z.B. Medienkonsum, Abstimmungsverhalten) zu erheben und zu analysieren. Zudem sollen aufgrund der Forschungsergebnisse Problem-orientierte Unterrichtskonzepte entwickelt werden, mit denen die wirtschaftsbürgerliche Kompetenz von Gymnasiasten noch besser gefördert werden kann.

Mit dem Forschungsprojekt WBKgym soll also letztlich auch die Kooperation zwischen empirischer Unterrichtsforschung und Schulpraxis gepflegt werden, um den WIRE-Unterricht am Gymnasium weiterzuentwickeln.

Der Kanton St. Gallen wurde für das Forschungsprojekt WBKgym ausgewählt, weil er in diesem Lernbereich bereits als sehr fortschrittlich eingestuft wird und weil er für die gymnasialen Maturitätsschulen einen einheitlichen kantonalen Lehrplan erlassen hat. Für die Datenerhebung werden insgesamt 20 Gymnasialklassen benötigt, die „Wirtschaft und Recht“ entweder als Schwerpunktfach (SPF), oder als Grundlagenfach (GF) belegen und sich im letzten Unterrichtssemester dieses Fachs befinden (SPF WIRE: 10 Klassen im 12. Schuljahr; GF WIRE: 10 Klassen im 11. Schuljahr).

Wir planen deshalb die Erhebungsphase im Februar/März 2017, eventuell auch schon im Januar 2017. Ein Erhebungstermin dauert pro Klasse ca. 130 Minuten, d.h. es werden drei aufeinander folgende Lektionen benötigt. Bei der Erhebung werden ein Leistungstest zur wirtschaftsbürgerlichen Kompetenz (WBK Test) und ein Fragebogen zu den schulischen und ausser-schulischen Kontextfaktoren (WBK Fragebogen) eingesetzt.

Ich bin mir bewusst, dass an Ihrer Schule neben dem regulären Unterricht noch andere Schulprojekte laufen, in die Schüler- und Lehrerschaft eingebunden sind. Trotzdem wäre ich – und auch mein Doktorvater Prof. Dr. Franze Eberle – Ihnen sehr verbunden, wenn Ihre Schule für das Forschungsprojekt WBKgym *möglichst viele* Klassen im Schwerpunktfach WIRE und etwa gleich viele Klassen mit Grundlagenfach WIRE („Einführung WR“) zur Verfügung stellen würde.

In der Beilage finden Sie ein Antwortformular. Ich bitte Sie höflichst, dieses bis Ende November 2016 ausgefüllt und unterschrieben zu senden an: nicole.ackermann@ife.uzh.ch.

Gerne kann ich Ihnen telefonisch weitere Details zum Forschungsprojekt und zur geplanten Datenerhebung erläutern und Ihre allfälligen Fragen klären.

Ich hoffe sehr auf Ihre Unterstützung meines Dissertationsvorhabens und danke Ihnen für Ihre Bemühungen.

Freundliche Grüsse

Universität Zürich
Institut für Erziehungswissenschaft

Nicole Ackermann
Wissenschaftliche Assistentin/Doktorandin

Beilagen: - Begleitschreiben von Prof. Dr. Franz Eberle
- Antwortformular

Antwortformular: **Kantonsschule YY, Ort**

Wir sind bereit, am Forschungsprojekt WBKgym teilzunehmen und stellen für die Erhebung folgende Klassen zur Verfügung:

	Grundlagenfach WIRE (Einführung Wirtschaft und Recht) im 11. Schuljahr (3. Klasse)	Schwerpunktfach WIRE im 12. Schuljahr (4. Klasse)
Volle Klassen	(Bezeichnung, z.B. 3L, 3S)	(Bezeichnung, z.B. 4W)
Halbe Klassen	(Bezeichnung, z.B. 3LW)	(Bezeichnung, z.B. 4LW, 4SW)
Total Klassen	(Anzahl)	(Anzahl)

Für uns sind grundsätzlich folgende Kalenderwochen für die Erhebung möglich:

☐ KW 02: Mo/Fr 09.-13.01.2017	☐ KW 07: Mo/Fr 13.-17.02.2017
☐ KW 03: Mo/Fr 16.-20.01.2017	☐ KW 08: Mo/Fr 20.-24.02.2017
☐ KW 04: Mo/Fr 23.-27.01.2017	☐ KW 09: Mo/Fr 27.02.-03.03.2017
☐ KW 05: Mo/Fr 30.01.-03.02.2017	☐ KW 10: Mo/Fr 06.-10.03.2017
☐ KW 06: Mo/Fr 06.-10.02.2017	☐ KW 11: Mo/Fr 13.-17.03.2017

Bemerkungen:

Ort, Datum:

Name, Unterschrift:



Frau/Herr Rektorin/Rektor
XX
Kantonsschule YY
Adresse
PLZ Ort

Institut für Erziehungswissenschaft
Abteilung Lehrerinnen- und Lehrerbildung
Maturitätsschulen

Universität Zürich
Institut für Erziehungswissenschaft
Kantonschulstrasse 3
CH-8001 Zürich
Telefon +41 44 634 66 55
Telefax +41 44 634 66 56
www.ife.uzh.ch

Prof. Dr. Franz Eberle
Lehrstuhl für Gymnasialpädagogik
Telefon +41 44 634 66 92
feberle@ife.uzh.ch

Zürich, 31. Oktober 2016

Liebe/r XX

Ich unterstütze das Vorhaben von Nicole Ackermann sehr. Sie ist eine Meisterin, hat ihre Matura an der Kantonsschule Sargans absolviert und ist heute eine vielversprechende Nachwuchsforscherin an meinem Lehrstuhl. Ihr Dissertationsvorhaben ist Teil meines Forschungs- und Doktorandenprogramms, in dem ich verschiedene Facetten des gymnasialen Bildungsziels der Vorbereitung auf anspruchsvolle Aufgaben in der Gesellschaft (vertiefte Gesellschaftsreife) untersuchen möchte. Für mich ist es wichtig, damit nach der Diskussion über „basale Kompetenzen für allgemeine Studierfähigkeit“ vermehrt wieder die ganze Breite der gymnasialen Bildung in den Fokus zu nehmen.

Ich würde mich sehr freuen, wenn du Frau Ackermann in ihrem Vorhaben unterstützen würdest.

Franz Eberle

Abbildung A- 1. Anfrage an Schulleiter*innen für Projekt WBKgym (Vorlage)

A3.2 Erhebungsplan

Projekt WBKgym: Erhebungsplan Kanton St.Gallen, Januar bis März 2017

KW 50 12.-16.12.2016	KW 51 - KW 02 26.12.2016-13.01.2017	KW 03 16.-20.01.2017	KW 04 23.-27.01.2017	KW 05 30.01.-03.02.2017	KW 06 06.-10.02.2017	KW 07 13.-17.02.2017	KW 08 - KW 10 20.02.2017-10.03.2017	KW 11 13.-17.03.2017
Fribourg, 3R (12): Do 15.12, 08:00-09:35, C 301 > nia	Weihnachtsferien	St.Gallen, 4hW (18): Fr 20.1, 07:55-10:30, H54 > nia	St.Gallen, 3pG (23): Mo 23.1, 07:55-10:30, H27 > nia, EW, DK	Winterferien	Wattwil, 3NL (18): Mo 6.2, 09:45-12:10, E6 > nia	Wattwil, 4Wb (18): Mo 13.2, 09:45-12:10, E6 > nia		Heerbrugg, 4Wa (7): Mi 15.3, 8:00-10:30, O220 > nia
Fribourg, 2R (20): Do 15.12, 14:00-15:35, C 003 > nia		St.Gallen, 4iW (25): Fr 20.1, 13:05-15:30, A54 > nia	St.Gallen, 3bLI (23): Mo 23.1, 08:45-11:20, H25 > nia, EW, DK		Wattwil, 3P (19): Mo 6.2, 13:00-15:35, E6 > nia	Wattwil, 4bW (7): Mo 13.2, 09:45-12:10, E6 > nia		Heerbrugg, 4We (9): Mi 15.3, 8:00-10:30, O220 > nia
			St.Gallen, 3oM (23): Mo 23.1, 09:35-12:10, E23 > nia, EW, DK		Wattwil, 3GM (23): Fr 10.2, 09:45-12:10, E6 > nia	Wattwil, 4Wa (15): Mi 15.2, 07:50-10:30, E6 > nia		Sargans, 3GM (19): Mi 15.3, 13:30-16:10, Turnhalle > nia, NY
			St.Gallen, 3cNP (21): Mo 23.1, 13:55-16:20, H51 > nia					Sargans, 3ILS (21): Mi 15.3, 13:30-16:10, Turnhalle > nia, NY
			St.Gallen, 4gW (19): Mi 25.1, 09:35-12:10, H51 > nia					Sargans, 3NP (17): Do 16.3, 09:35-11:55, Turnhalle > nia, (EW)
			St.Gallen, 4fW (20): Fr 27.1, 13:05-15:30, H50 > nia					Sargans, 4bW (9): Do 16.3, 09:35-11:55, Turnhalle > nia, (EW)
			Heerbrugg, 3GM (24): Do 26.1, 09:15-11:50, O218 > nia, DK					Sargans, 4Wa (7): Do 16.3, 13:30-16:10, Turnhalle > nia, (EW)
			Heerbrugg, 3Sa (12): Do 26.1, 09:15-11:50, O218 > nia, DK					Sargans, 4Wb (24): Do 16.3, 13:30-16:10, Turnhalle > nia, (EW)

Legende:

Schule, Klasse (Anzahl SuS): Datum, Lektionen, Zeitfenster, Raum > Leiter

Abbildung A- 2. Erhebungsplan für Projekt WBKgym: Piloterhebung (Dezember 2016) und Haupterhebung (Januar bis März 2017)

A3.3 Erhebungsmanual



Universität Zürich, Institut für Erziehungswissenschaft (IFE)
Forschungsprojekt WBKgyM
Nicole Ackermann, nicole.ackermann@ife.uzh.ch

Forschungsprojekt WBKgyM

Erhebungsmanual

Materialien

WAS	WIE VIEL
Erhebungsmanual	1
Erhebungsprotokoll	1
Klassenliste (Schule, Klasse Name, Vorname, ID)	1
Einverständniserklärung mit Etikette	X (Anzahl SuS)
Testhefte mit Etikette (Code)	X (Anzahl SuS)
Fragebogenhefte mit Etikette (Code)	X (Anzahl SuS)
Schokoriegel	X (Anzahl SuS) + Y (Anzahl LP)
Praliné	1 (Prorektor bzw. verantwortliche LP)
Rollkoffer	1
Uhr	1
Notizpapier	20
Visitenkarte	10
Kugelschreiber, Büroklammern, Gummibänder	10
Testhefte ohne Etikette (Ersatz)	5
Fragebogenhefte ohne Etikette (Ersatz)	5

Instruktionen

- KEINE mobilen Geräte (Handy/Smartphone, Tablet, Notebook) während Erhebung
- Lehrbücher, Unterlagen und mobile Geräte in Schultasche versorgen
- Schultaschen an Zimmerrand (vorne, hinten, seitlich) deponieren
- gemeinsame Pause zwischen Test und Fragebogen, ggf. kleine individuelle Pause (WC)
- EVE und Hefte austeilen anhand Etikette bzw. Klassenliste
- fehlende SuS: im Erhebungsprotokoll vermerken, Etikette auf Deckblatt durchkreuzen
- Hefte spätestens nach Zeitablauf einsammeln, individuell auch früher
- ...



Universität Zürich, Institut für Erziehungswissenschaft (IFE)
Forschungsprojekt WBKgyM
Nicole Ackermann, nicole.ackermann@ife.uzh.ch

Ablauf (total 125 min)

1 Begrüssung/Vorstellung und Einführung (10 min)

- 10 min
- Nicole Ackermann, Universität Zürich; Doktorandin und Gymnasiallehrerin für Wirtschaft & Recht
 - Ziele des Forschungsprojekts WBKgyM: Wissen und Interesse zu wirtschafts- und gesellschaftspolitischen Themen
 - Inhalt und Ablauf der heutigen Untersuchung: 1) WBK-Test, 2) Kontext-Fragebogen
 - Einverständniserklärung: Einwilligung, Anonymisierung
 - Fragen der SuS klären
 - Sitzordnung der SuS überprüfen: einzeln, mind. 1 Platz Abstand

2 Einverständniserklärung (EVE) (10 min)

- 5 min
- EVE austeilen: durchlesen, unterschreiben
- 5 min
- EVE einsammeln: Unterschriften prüfen, Kontaktdaten prüfen

3 WBK-Test (70 min)

- 5 min
- Testhefte austeilen
- 60 min
- Zeit stoppen: 60 min
- 5 min
- Testhefte einsammeln, laufend, spätestens nach 60 min

4 Pause (5 min)

- Befindenden der SuS nachfragen
- Fragen der SuS klären

5 Kontext-Fragebogen (30 min)

- 5 min
- Fragehefte austeilen
- 20 min
- Zeit stoppen: 20 min
- 5 min
- Fragehefte einsammeln, laufend, spätestens nach 20 min

6 Dank/Verabschiedung (5 min)

- 5 min
- Teilnahme verdanken
 - Schokoriegel austeilen
 - Visitenkarte auflegen

Abbildung A- 3. Erhebungsmanual für Projekt WBKgyM

A3.4 Erhebungsprotokoll (Vorlage)



Universität Zürich, Institut für Erziehungswissenschaft (IFE)
Forschungsprojekt WBKgym
Nicole Ackermann, nicole.ackermann@ife.uzh.ch

Forschungsprojekt WBKgym

Erhebungsprotokoll

Schule	
Klasse Bezeichnung (Anzahl SuS)	
Datum, Uhrzeit (von/bis)	
Raum	
Kontaktperson (SL, LP) Name, Tel	

Einverständniserklärung	im KAB: Anzahl und Etiketten geprüft, numerisch sortiert: ☐ in Schule: Anzahl und Unterschriften geprüft, numerisch sortiert: ☐	
Testhefte (60 min)	im KAB: Anzahl geprüft, numerisch sortiert: ☐ in Schule: Anzahl geprüft, numerisch sortiert: ☐	Start: ____ : ____ Ende: ____ : ____
Fragebogenhefte (20 min)	im KAB: Anzahl geprüft, numerisch sortiert: ☐ in Schule: Anzahl geprüft, numerisch sortiert: ☐	Start: ____ : ____ Ende: ____ : ____

Bemerkungen Kooperation mit SL/LP, Atmosphäre in Klasse, Rückfragen der SuS, Verhalten der SuS, fehlende SuS, etc.	
--	--

Administration Name, Unterschrift, Datum	
Notfall-Kontakt	Nicole Ackermann, +41 79 447 76 63

Abbildung A- 4. Erhebungsprotokoll für Projekt WBKgym (Vorlage)

A3.5 Einverständniserklärung für Schüler*innen (Auszug)



Universität Zürich, Institut für Erziehungswissenschaft (IFE)
Forschungsprojekt WBKgyM
Nicole Ackermann, nicole.ackermann@ife.uzh.ch

Kanton	
Schule	
Klasse	
Schüler/-in	
Code	

Forschungsprojekt WBKgyM

Einverständniserklärung

Liebe Schülerin, lieber Schüler

Im Rahmen meines Forschungsprojekts „Wirtschaftsbürgerliche Kompetenz Deutschschweizer Gymnasiastinnen und Gymnasiasten (WBKgyM)“ untersuche ich das Wissen und das Interesse von Schülerinnen und Schülern bezüglich aktueller wirtschafts- und gesellschaftspolitischer Themen. Dazu werden Ihnen an der heutigen Untersuchung zwei Hefte vorgelegt:

- ein **Wissenstest** mit Aufgaben zu aktuellen und relevanten Themen, ca. **60 Minuten**;
- ein **Fragebogen** mit Fragen zu Ihrem schulischen und ausserschulischen Umfeld, ca. **20 Minuten**.

Die Teilnahme an der Untersuchung ist freiwillig. Durch die Teilnahme an der Untersuchung haben Sie jedoch die Gelegenheit, Ihr Wissen zu den vorgelegten Themen des Tests zu erweitern und zu vertiefen. Der Test wird nicht benotet und hat keinen Einfluss auf Ihre Zeugnisnote. Die Schulleitung und Ihre Lehrer/-innen erhalten keinen Einblick in Ihre individuellen Antworten.

Ihre Antworten dienen wissenschaftlichen Zwecken und werden streng vertraulich behandelt. Ihre persönlichen Angaben (Name, Klasse, Schule, etc.) werden nicht im elektronischen Datensatz erfasst und somit nicht für die Datenauswertung verwendet.

Aus Gründen des Persönlichkeits- und Datenschutzes benötige ich für die Untersuchung Ihr Einverständnis. Ich bitte Sie deshalb, die folgende Einverständniserklärung zu lesen und zu unterzeichnen. Sie erhalten eine Kopie dieser Einverständniserklärung.

Ich freue mich auf Ihre Antworten und danke Ihnen herzlich für Ihre Teilnahme am Forschungsprojekt WBKgyM!

Freundliche Grüsse,

Nicole Ackermann
Projektleiterin WBKgyM

Zürich, 30.05.2019

► Bitte wenden.



Universität Zürich, Institut für Erziehungswissenschaft (IFE)
Forschungsprojekt WBKgyM
Nicole Ackermann, nicole.ackermann@ife.uzh.ch

Ich bestätige, dass ich diese Einverständniserklärung gelesen und verstanden habe und, dass ich freiwillig an der Untersuchung des Forschungsprojekts WBKgyM teilnehme.

[Ort, Datum] _____

[Vorname Familienname] _____

[Unterschrift] _____

Falls Sie Anregungen oder Fragen zum Forschungsprojekt WBKgyM haben, kontaktieren Sie bitte:

Nicole Ackermann
Universität Zürich, Institut für Erziehungswissenschaft, Kantonsschulstr. 3, 8001 Zürich

Telefon: +41 (0)44 634 66 89
Email: nicole.ackermann@ife.uzh.ch

Kontakt Daten

Wollen Sie über die Ergebnisse der Studie informiert werden?

- ☐ JA
☐ NEIN

Sind Sie bereit an einer Folgestudie teilzunehmen (z.B. Interview, Fragebogen)?

- ☐ JA
☐ NEIN

Falls JA (bei einer der obigen Fragen): Notieren Sie bitte Ihre Kontaktdaten.

Vorname und Familienname: _____

Strasse und Nummer: _____

Postleitzahl und Wohnort: _____

Email-Adresse: _____

Abbildung A- 5. Einverständniserklärung für Schüler*innen für Teilnahme am Projekt WBKgyM

A3.6 Testheft des WBK-T2 (Auszug)



Universität Zürich, Institut für Erziehungswissenschaft (IfE)
Forschungsprojekt WBKgym, revidierter WBK-Test
Nicole Ackermann, nicole.ackermann@ife.uzh.ch

Code:

Forschungsprojekt WBKgym

Wissenstest zur Wirtschaftsbürgerlichen Kompetenz

Testheft
Erhebung
Januar-März 2017



Universität Zürich, Institut für Erziehungswissenschaft (IfE)
Forschungsprojekt WBKgym, revidierter WBK-Test
Nicole Ackermann, nicole.ackermann@ife.uzh.ch

Hinweise

Liebe Schülerin, lieber Schüler

In diesem **Test** werden Sie zu wirtschafts- und gesellschaftspolitischen Themen geprüft. Der Test beinhaltet 4 Themen (Managervergütung, Altersvorsorge, Staatsverschuldung, Energieversorgung). Jedes Thema beginnt mit einem Einleitungstext, gefolgt von ca. 8 Aufgaben. Einige Aufgaben lassen sich mithilfe des Einleitungstexts lösen, andere nicht.
Das Lösen des Tests dauert **60 Minuten**. Am Schluss dürfen Sie **Ihre Rückmeldung** zum gesamten Test geben.

Bitte beachten Sie beim Lösen des Tests folgende Hinweise:

- Lesen Sie die Einleitungstexte und Aufgabenstellungen sorgfältig.
- Verwenden Sie Leuchttifte zum Markieren und einen Kugelschreiber zum Schreiben.
- Schreiben Sie leserlich.
- Bearbeiten Sie alle Themen bzw. alle Aufgaben.
- Lösen Sie die Aufgaben in der vorgegebenen Reihenfolge.
- Beachten Sie die nachfolgenden Beispiele zum Lösen der Aufgaben.

Ich freue mich auf Ihre Antworten und danke Ihnen herzlich für Ihre Teilnahme am Forschungsprojekt WBKgym!

Nicole Ackermann

Beispiel 1:

Kreuzen Sie bei jeder Frage diejenige Antwort an, die Sie für korrekt halten. Von den vier möglichen Antworten ist jeweils nur eine Antwort zulässig.

Wenn Sie Ihre Antwort korrigieren wollen, dann übermalen Sie die falsche Antwort und kreuzen Sie die korrekte Antwort an. In diesem Beispiel wird nun Antwort 2 bewertet.

Frage ...?	
☐	Antwort 1
☒	Antwort 2
☐	Antwort 3
☐	Antwort 4

Beispiel 2:

Schreiben Sie Ihre Antwort auf die Frage in das freie Feld.

Frage ...?	
Erklärung:	Ihre Antwort ...

Abbildung A- 6. Testheft des WBK-T2 im Projekt WBKgym: Titelblatt und Instruktionen (Auszug)

A3.7 Geheimhaltungs- und Sorgfaltserklärung für Testadministrator*innen und Kodierer*innen (Vorlage)



Universität Zürich, Institut für Erziehungswissenschaft (IFE)
Forschungsprojekt WBKgym
Nicole Ackermann, nicole.ackermann@ife.uzh.ch

Forschungsprojekt WBKgym

Geheimhaltungs- und Sorgfaltserklärung

Liebe Studentin, lieber Student

Bei Ihrer Tätigkeit als Erhebungsleiter/Erhebungsleiterin und/oder Kodierer/Kodiererin (Datenerfassung, Datenbereinigung) für das Forschungsprojekt „Wirtschaftsbürgerliche Kompetenz Deutschschweizer Gymnasiasten (WBKgym)“ kommen Sie mit Informationen, Dokumenten und Daten in Kontakt, die der Geheimhaltung unterliegen.

Um den Datenschutz sicherzustellen, bitte ich Sie folgende Geheimhaltungserklärung zu lesen und zu unterzeichnen.

Besten Dank und freundliche Grüsse,

Nicole Ackermann
Projektleiterin WBKgym

Zürich, 30.05.2019

Ich verpflichte mich, leere und ausgefüllte Hefte sowie leere und erfasste Datentabellen, die ich für die Datenerhebung und Datenerfassung des Forschungsprojekt WBKgym erhalte, weder zu vervielfältigen noch an Dritte zur Ansicht oder zur Vervielfältigung herauszugeben. Über Antworten und Kommentare in den Heften darf ich nicht mit Dritten sprechen.

Ich verpflichte mich, leere und ausgefüllte Hefte sorgfältig zu behandeln und nach der Datenerhebung bzw. Datenerfassung in den Raum KAB-F01 (Universität Zürich, Institut für Erziehungswissenschaft, Kantonsschulstrasse 3, 8001 Zürich) zurückzubringen. Die Hefte darf ich nicht mit eigenen Notizen versehen. Die Ordnung im Schrank KAB-F01 halte ich ein.

Diese Geheimhaltungsverpflichtung bleibt auch nach Beendigung meiner Tätigkeit im Projekt WBKgym bestehen.

[Vorname Familienname] _____

[Unterschrift] _____

[Ort, Datum] _____

Abbildung A- 7. Geheimhaltungs- und Sorgfaltserklärung für Testadministrator*innen und Kodierer*innen im Projekt WBKgym

A3.8 Kodierungsplan

Projekt WBKgym: Kodierungsplan für Erhebung im Kanton St.Gallen, Januar bis März 2017

ID_i	ID_k	SuS #	Zeitraum	WBK-T2 Kodierer K1	Kodierer K2	Konsens	WBK-Q1 Kodierer K3
SG01	SG01-G301	23	ab Mo 23.1	nia	---	---	LZ
SG01	SG01-G302	21	ab Mo 23.1	nia	NY	nia & NY	LZ
SG01	SG01-G303	23	ab Mo 23.1	NY	---	---	LZ
SG01	SG01-G304	23	ab Mo 23.1	NY	---	---	LZ
SG01	SG01-G401	20	ab Mo 30.1	nia	---	---	LZ
SG01	SG01-G402	19	ab Mo 30.1	nia	---	---	LZ
SG01	SG01-G403	18	ab Mo 23.1	NY	nia	nia & NY	LZ
SG01	SG01-G404	25	ab Mo 23.1	NY	---	---	LZ
SG02	SG02-G301	24	ab Fr 27.1	nia	---	---	LZ
SG02	SG02-G302	12	ab Fr 27.1	NY	---	---	LZ
SG02	SG02-G401	7	ab Fr 17.3	NY	---	---	LZ
SG02	SG02-G402	9	ab Fr 17.3	NY	---	---	LZ
SG03	SG03-G301	19	ab Fr 17.3	NY	---	---	LZ
SG03	SG03-G302	21	ab Fr 17.3	NY	---	---	LZ
SG03	SG03-G303	17	ab Fr 17.3	nia	NY	nia & NY	LZ
SG03	SG03-G401	7	ab Fr 17.3	nia	---	---	LZ
SG03	SG03-G402	24	ab Fr 17.3	NY	---	---	LZ
SG03	SG03-G403	9	ab Fr 17.3	nia	NY	nia & NY	LZ
SG04	SG04-G301	23	ab Mo 11.2	nia	---	---	LZ
SG04	SG04-G302	18	ab Di 7.2	NY	nia	nia & NY	LZ
SG04	SG04-G303	19	ab Di 7.2	NY	---	---	LZ
SG04	SG04-G401	15	ab Do 16.2	nia	NY	nia & NY	LZ
SG04	SG04-G402	18	ab Do 16.2	NY	---	---	LZ
SG04	SG04-G403	7	ab Do 16.2	NY	---	---	LZ

4 24 421

Legende:

ID_i = Code Schule

ID_k = Code Klasse

K = Kodierer/in

= Anzahl

Abbildung A- 8. Kodierungsplan für Projekt WBKgym

A4 Fragebogen der Individual- und Kontextmerkmale

A4.1 Skalen- und Item-Spezifikationen des WBK-Q1 (Auszug)

Tabelle A- 15. Skalen- und Item-Spezifikationen des WBK-Q1 (Auszug)

Skala/Item-Krz.	Inhalt	Antwortformat	Antwortkategorien (Score)	Skalenniveau	Item #
Einstellungs- und Interessensmerkmale					
interest	Interesse an der Auseinandersetzung mit sozioökonomischen Problemsituationen	Gebunden: 5-stufige Antwortskala	Polytom (1/2/3/4/5)	Items ordinal > Skala metrisch	3
Leistungsmerkmale					
grade_WR	Schulnote WuR bzw. VWL	Frei: Zahl, mit einer Dezimale (x.x)	Numerisch	Ordinal	1
grade_DE	Schulnote Deutsch	Frei: Zahl, mit einer Dezimale (x.x)	Numerisch	Ordinal	1
grade_MA	Schulnote Mathematik	Frei: Zahl, mit einer Dezimale (x.x)	Numerisch	Ordinal	1
Soziodemografische Merkmale					
gender	biologisches Geschlecht	Gebunden: 2 Antwortoptionen	Dichotom (0/1)	Nominal	1
birthdate	Geburtsdatum	Frei: Zahl (dd.mm.yyyy)	Rekodierung zu „age“: numerisch	Metrisch (age)	1
citizen	Staatsbürgerschaft	Mix: 3 Antwortoptionen, plus Freitext	Polytom (1/2/3)	Nominal	1
Soziokulturelle und sozioökonomische Merkmale					
country	Geburtsland	Mix: 2 Antwortoptionen, plus Freitext	Dichotom (0/1)	Nominal	1
lang_fam	Sprache in Familie	Mix: 3 Antwortoptionen, plus Freitext	Polytom (1/2/3)	Nominal	1

Tabelle A- 15. Skalen- und Item-Spezifikationen des WBK-Q1 (Auszug)

Skala/Item-Krz.	Inhalt	Antwortformat	Antwortkategorien (Score)	Skalenniveau	Item #
edu_fa	Bildungsabschluss Vater	Mix: 9 Antwortoptionen, plus Freitext	Polytom (1/2/3/.../9)	Nominal (edu_fa)	1
			Rekodierung zu „Bildungsstufe“ (1/2/3)	Nominal (edu_fa3)	1
			Rekodierung zu „nicht-akad/akad“ (0/1)	Metrisch (edu_fa1)	1
edu_mo	Bildungsabschluss Mutter	Mix: 9 Antwortoptionen, plus Freitext	Polytom (1/2/3/.../9)	Nominal (edu_mo)	1
			Rekodierung zu „Bildungsstufe“ (1/2/3)	Nominal (edu_mo3)	1
			Rekodierung zu „nicht-akad/akad“ (0/1)	Metrisch (edu_mo1)	1
prof_fa	Beruf Vater	Frei: Begriff	Kodierung zu „ISCO-08“: numerisch	Nominal (ISCO_fa)	1
			Rekodierung zu „ISEI-08“: numerisch	Metrisch (ISEI_fa)	1
prof_mo	Beruf Mutter	Frei: Begriff	Kodierung zu „ISCO-08“: numerisch	Nominal (ISCO_mo)	1
			Rekodierung zu „ISEI-08“: numerisch	Metrisch (ISEI_mo)	1

Quelle: Ackermann (2018b).

A4.2 Skalen- und Item-Kennwerte des WBK-Q1 (Auszug)

Tabelle A- 16. Skalen- und Item-Kennwerte des WBK-Q1 (Auszug)

Skala	Bezeichnung / Formulierung	N total	N miss.	M	SD	Schiefe	Steilheit	Item #	Cronbachs α	r_{item}	Varianz- aufklärung
Einstellungs- und Interessensmerkmale											
interest	Interesse an der Auseinandersetzung mit sozioökonomischen Themen	375	0	3.88	0.77	-0.56	0.04	3	.71	0.58	63.4
<i>Bsp.-Item: Ich finde es interessant, wenn ich mich mit anderen über aktuelle wirtschafts- und gesellschaftspolitische Themen austauschen kann. (int_2)</i>											
Leistungsmerkmale											
grade_DE	Schulnote WuR bzw. VWL	375	0	4.70	0.469	-0.22	-0.01	1	---	---	---
grade_MA	Schulnote Deutsch	375	1	4.60	0.779	-0.13	-0.39	1	---	---	---
grade_WR	Schulnote Mathematik	375	1	4.72	0.480	0.08	-0.01	1	---	---	---

Quelle: Ackermann (2018b).