Berufsbildung, Arbeit und Innovation
Dissertationen/Habilitationen

Christine Caroline Jähnig

28

Die Messung betriebswirtschaftlichen Wissens von Studierenden

Eine quantitativ-empirische Untersuchung situativer Testaufgaben

wbv

Berufsbildung, Arbeit und Innovation
Dissertationen/Habilitationen

Christine Caroline Jähnig

Die Messung betriebswirtschaftlichen Wissens von Studierenden

Eine quantitativ-empirische Untersuchung situativer Testaufgaben

wbv

Bibliografische Information der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.ddb.de> abrufbar.

Reihe Berufsbildung, Arbeit und Innovation –
Dissertationen/Habilitationen, Band 28

Geschäftsführende Herausgeber
Klaus Jenewein, Magdeburg
Marianne Friese, Gießen
Georg Spöttl, Bremen

Wissenschaftlicher Beirat
Rolf Arnold, Kaiserslautern
Ingrid Darmann-Finck, Bremen
Friedhelm Eicker, Rostock
Uwe Faßhauer, Schwäbisch-Gmünd
Martin Fischer, Karlsruhe
Philipp Gonon, Zürich
Richard Huisinga, Siegen
Manuela Niethammer, Dresden
Jörg-Peter Pahl, Dresden
Günther Pätzold, Dortmund
Karin Rebmann, Oldenburg
Tade Tramm, Hamburg
Thomas Vollmer, Hamburg

Die Verantwortung für den Inhalt der Veröffentlichung liegt bei der Autorin. Diese Veröffentlichung lag dem Promotionsausschuss an der Georg-August-Universität Göttingen mit dem Titel „Die Messung betriebswirtschaftlichen Wissens von Studierenden – Eine quantitativ empirische Untersuchung situativer Testaufgaben" vor.

Gutachter: Prof. Dr. Susan Seeber und Prof. Dr. Thomas Kneib
Prüfer: Prof. Dr. Susan Seeber, Prof. Dr. Thomas Kneib und Prof. Dr. Stefan Dierkes
Die Disputation (Dr. rer. pol.) fand am 18.12.2013 statt.

W. Bertelsmann Verlag GmbH & Co. KG, Bielefeld, 2014
Gesamtherstellung: W. Bertelsmann Verlag, Bielefeld
Umschlaggestaltung: FaktorZwo, Günter Pawlak, Bielefeld

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlages unzulässig und strafbar. Insbesondere darf kein Teil dieses Werkes ohne vorherige schriftliche Genehmigung des Verlages in irgendeiner Form (unter Verwendung elektronischer Systeme oder als Ausdruck, Fotokopie oder unter Nutzung eines anderen Vervielfältigungsverfahrens) über den persönlichen Gebrauch hinaus verarbeitet, vervielfältigt oder verbreitet werden.

Für alle in diesem Werk verwendeten Warennamen sowie Firmen- und Markenbezeichnungen können Schutzrechte bestehen, auch wenn diese nicht als solche gekennzeichnet sind. Deren Verwendung in diesem Werk berechtigt nicht zu der Annahme, dass diese frei verfügbar seien.

ISBN 978-3-7639-5394-3
Bestell-Nr. 6004416
Dieses Buch ist auch als E-Book unter der ISBN 978-3-7639-5395-0 erhältlich.

Inhalt

Abbildungsverzeichnis 7

Tabellenverzeichnis 9

Abkürzungsverzeichnis 13

1 Einleitung 17
1.1 Problemstellung und Zielsetzung 17
1.2 Forschungsfragen und forschungsmethodisches Vorgehen 20
1.3 Aufbau und Struktur der Arbeit 21

2 Wissen als Bestandteil professioneller Handlungskompetenz 25
2.1 Die Relevanz universitärer Bildung in der Wissensgesellschaft 25
2.2 Das Verhältnis von Wissen zu Kompetenzen 26
2.3 Die Unterscheidung der Wissensarten 29
2.3.1 Die kognitionspsychologische Perspektive 31
2.3.2 Die didaktische Perspektive 33
2.4 Zusammenführung der zwei Perspektiven auf Wissen 38

3 Stand der Forschung zur Messung des studentischen Wirtschaftswissens 41
3.1 Deutschsprachige Messinstrumente 41
3.1.1 Der Wirtschaftskundliche Bildungs-Test 42
3.1.2 Der Business Administration Knowledge Test 44
3.2 Nationale Forschungsergebnisse 48
3.3 Laufende nationale Forschungsprojekte 53
3.4 Internationale Messinstrumente 54
3.4.1 Der Major Field Achievement Test in Business 54
3.4.2 Assessment of Higher Education Learning Outcomes 56
3.5 Internationale Forschungsergebnisse 58
3.6 Zusammenfassung des Forschungsstands 66

4 Konzepte der Kompetenzmessung in kaufmännischen Handlungsfeldern 69
4.1 Zur Strukturierung der kaufmännischen Domäne 69
4.2 Kompetenzmodellierung in der kaufmännischen Bildung 73

4.2.1 Kompetenzstrukturmodelle 73
4.2.2 Kompetenzniveaumodelle 78
4.3 Zusammenfassung zur kaufmännischen Kompetenzmessung 82

5 Situative Messverfahren und -instrumente 85
5.1 Papierbasierte situative Tests 85
5.2 Videobasierte Tests und Computersimulationen 87
5.3 Besondere Eigenschaften situativer Items 90
5.3.1 Die Konstruktdiskussion 90
5.3.2 Kritische Betrachtung situativer Items 93

6 Entwicklung eines situativen betriebswirtschaftlichen Wissenstests 97
6.1 Methodische Grundlagen der Testentwicklung 97
6.1.1 Modelle der Testentwicklung 97
6.1.2 Definition des Zielkonstrukts 100
6.2 Anforderungen des Lernens und Arbeitens 101
6.2.1 Universitäre Anforderungen des Lernens 102
6.2.2 Anforderungen des Lernens und Arbeitens in betriebswirtschaftlichen Handlungsfeldern 108
6.3 Rahmenmodell der Itementwicklung 111
6.3.1 Implikationen der bisherigen Forschung für die Testentwicklung 111
6.3.2 Spezifikation der kognitiven Anforderungen 113
6.3.3 Itementwicklung und resultierende Testcharakteristika 114
6.4 Zusammenfassung 120

7 Standards pädagogisch-psychologischer Diagnostik 121
7.1 Modelle zur Auswertung von Tests 121
7.1.1 Die klassische Testtheorie 122
7.1.2 Probabilistische Testtheorien 123
7.2 Kriterien der Beurteilung der Güte von Tests 131
7.2.1 Objektivität 131
7.2.2 Reliabilität 133
7.2.3 Validität 134
7.2.4 Testfairness und Testökonomie 138

8 Testpilotierung 141
8.1 Zielsetzung der Pilotierung 141
8.2 Testdurchführung, Testmaterial und Stichprobenbeschreibung 142

8.3 Ergebnisse der Pilotierung 144
8.3.1 Itemselektion und Auswertung nach Rasch-Modell. 144
8.3.2 Erste Indikatoren der Testvalidität. 148
8.4 Konsequenzen aus der Pilotierung für die Haupterhebung 149

9 Testvalidierung und weiterführende Analysen 153
9.1 Ableitung der Hypothesen 153
9.1.1 Hypothesen zur Struktur betriebswirtschaftlichen Wissens 154
9.1.2 Hypothesen zu den Determinanten betriebswirtschaftlichen Wissens 156
9.2 Testdurchführung, Testmaterial und Stichprobenbeschreibung. 159
9.3 Empirische Ergebnisse 162
9.3.1 Auswertungen nach Rasch-Modell und Prüfung der Dimensionalität 162
9.3.2 Konstruktvalidierung 167
9.3.3 Kriteriumsvalidierung 178
9.3.4 Analyse der Testfairness und Testakzeptanz 184

10 Diskussion 193
10.1 Zusammenfassung und Interpretation der empirischen Ergebnisse 193
10.2 Limitation der empirischen Ergebnisse und Forschungsbedarfe. 200
10.3 Fazit und Ausblick 211

Literaturverzeichnis 213

Anhang A Synoptische Darstellung der Modulbeschreibungen 235

Anhang B Pilotierungsergebnisse 239

Anhang C Instruktion und Begleitfragebogen zum Test 245

Anhang D Skalendokumentationen 257

Anhang E Situative Items und Itemcharakteristika 263

Anhang F Berechnungen zur Haupterhebung. 275

Abbildungsverzeichnis

Abb. 1 Strukturierung des Untersuchungsgegenstandes „professionelles Wissen" als Facette professioneller Handlungskompetenz 29

Abb. 2 Beispiel für eine typische Produktionsregel (Quelle: Anderson, 2001, S. 525) 32

Abb. 3 Beispielaufgabe für prozedurales Wissen auf Stufe B 38

Abb. 4 Domänen und Subdomänen des Wirtschaftswissenstests (BAKT-L) 45

Abb. 5 Lernziele des Economics Assessment Frameworks der AHELO Studie 57

Abb. 6 Der Instrumentenentwicklungszyklus entlang der „Four Building Blocks" 99

Abb. 7 Darstellung der Verankerung von Iteminhalten 115

Abb. 8 Gegenstandsbereich pädagogisch-psychologischer Testtheorien 121

Abb. 9 Grafische Darstellung Kategorienfunktionen dichotomer Items 126

Abb. 10 Darstellung der Pilotierungsergebnisse in Form einer Wright-Map 147

Abb. 11 Darstellung der Ergebnisse der Haupterhebung in Form einer Wright-Map 163

Abb. 12 Unterschiede in der Testleistung in Abhängigkeit des Studiengangs 171

Abb. 13 Schwierigkeitsindizes der Testaufgaben nach kognitiven Anforderungen 174

Abb. 14 Schwierigkeitsindizes der Testaufgaben nach Komplexität 175

Abb. 15 Schwierigkeitsindizes der Testaufgaben nach mathematischer Modellierungsanforderung 176

Abb. 16 Prozentuale Lösungshäufigkeit der einzelnen Testaufgaben nach Aufgabentyp 177

Abb. 17 Prozentsatz gelöster Aufgaben nach Aufgabentyp und absolvierter kaufmännischer Ausbildung 178

Abb. 18 Geschlechtsspezifische Item Characteristic Curves für das Item Nummer 19 186

Abb. 19 Vergleichende Darstellung der Einschätzung situativer und nicht-situativer Items hinsichtlich unterschiedlicher Facetten der Testakzeptanz (Kersting, 2008) sowie der curricularen Validität 189

Abb. 20 Vergleichende Darstellung der selbstberichteten Anstrengungsbereitschaft 191

Abb. B-1 Grafische Darstellung des Prozesses der Itementwicklung 239

Abb. F-1 Grafischer Modelltest für männliche und weibliche Subgruppen in der Stichprobe mit 95 %-Konfidenzregion um jedes Item 277

Abb. F-2 Unterschiede in der Testleistung in Abhängigkeit des Studiengangs 278

Abb. F-3 Darstellung der Einschätzung der Augenscheinvalidität in Abhängigkeit der durch Berufsausbildung, Praktika oder Nebentätigkeiten erworbenen betriebswirtschaftlichen Praxiserfahrung 280

Tabellenverzeichnis

Tab. 1 Die überarbeitete Taxonomie der Wissensformen 34

Tab. 2 Das Faktenwissen und seine Unterformen 36

Tab. 3 Das prozedurale Wissen und seine Unterformen 37

Tab. 4 Lineare Regression auf die Summe der Rohpunkte des WBTs 51

Tab. 5 Zusammenfassung der zentralen Prädiktoren für Studienleistung in betriebswirtschaftlichen Fächern 61

Tab. 6 Überblick über Regressionsanalysen zur Identifikation relevanter Prädiktoren für die Leistung im MFAT-B aus ausgewählten Studien im englischen Sprachraum 65

Tab. 7 Lernziele für Studierende der Wirtschaftswissenschaften 106

Tab. 8 Die zehn häufigsten Funktionsbereiche von Absolventen der Betriebswirtschaftslehre 109

Tab. 9 Die zehn am häufigsten im Onlinestellenmarkt ausgeschriebenen Berufsfelder für BWL-Absolventen im April 2012 110

Tab. 10 Tabellarische Darstellung der Testzusammensetzung und intendierte Itemeigenschaften 117

Tab. 11 Interpretationshilfe für den Kennwert Cronbachs Alpha 133

Tab. 12 Beschreibung der Pilotierungsstichprobe nach Studiengang und Studienabschnitt (N = 154) 143

Tab. 13 Darstellung der Stichprobenzusammensetzung für die gesamte Haupterhebung nach Studiengang und Studienabschnitt (N = 421) . . . 160

Tab. 14 Beschreibung des Aufbaus der beiden Studien der Haupterhebung 161

Tab. 15 Vergleich der Passung des eindimensionalen Modells mit zwei alternativen Modellen (N = 351) 165

Tab. 16 Ergebnisse der explorativen Faktoranalyse (Hauptkomponentenanalyse) über die Modulnoten in den testrelevanten Modulen für metrische und ordinale Daten 167

Tab. 17 Vergleichende Darstellung der Modellpassung unterschiedlicher Modelle über die Items des BAKT (Bothe, 2003) und des situativen Tests (N = 351) 169

Tab. 18 Interkorrelationsmatrix potenzieller Prädiktoren zur Vorhersage der Testleistung 181

Tab. 19 Multiple OLS-Regression zur Vorhersage der geschätzten Personenparameter im situativen betriebswirtschaftlichen Wissenstest im Vergleich zu der Vorhersage der Leistung in Items des BAKT (Bothe, 2003) von N = 272 Studierenden 182

Tab. 20 Interpretation der Ergebnisse der Regressionsanalyse mit Bezug auf die aufgestellten Hypothesen 183

Tab. A-1 Ziele und Inhalte der Pflichtmodule „Finanz- und Rechnungswesen" . . . 235

Tab. A-2 Ziele und Inhalte des Pflichtmoduls „Unternehmensführung" 236

Tab. A-3 Ziele und Inhalte des Pflichtmoduls „Marketing" 237

Tab. A-4 Ziele und Inhalte des Pflichtmoduls „Produktion" 238

Tab. B-1 Zusammenfassende Darstellung der Pilotierungsergebnisse 240

Tab. B-2 Multiple OLS-Regression zur Vorhersage der geschätzten Personenparameter im situativen betriebswirtschaftlichen Wissenstest von N = 126 Studierenden der Universität Göttingen 243

Tab. C-1 Aufbau des Testhefts und Quellen der Skalen für die Haupterhebung . . 245

Tab. D-1 Skalendokumentation zur Skala „Studieninteresse“ aus der Kurzform des FSI 257

Tab. D-2 Skalendokumentation zur Skala „Leistungsmotivation“ 257

Tab. D-3 Skalendokumentation zur Skala „Wettbewerbsmotivation“ 258

Tab. D-4 Skalendokumentation zur Skala „akademisches Selbstkonzept“ 258

Tab. D-5 Skalendokumentation zur Skala „mathematisches Selbstkonzept“ 259

Tab. D-6 Skalendokumentation zur Skala „Lernstrategien: Zusammenhangslernen“ 259

Tab. D-7 Skalendokumentation zur Skala „Lernstrategien: kritisches Hinterfragen“ 260

Tab. D-8 Skalendokumentation zur Skala „Lernstrategien: Wiederholungsstrategien“ 260

Tab. D-9 Skalendokumentation zur Skala „Lernstrategien: Anstrengungsbereitschaft“ 261

Tab. E-1 Itemkennwerte der Haupterhebung (N = 351) und Kennwerte der verlängerten Testversion (N = 35) 271

Tab. E-2 Bezüge der Items zum Curriculum sowie zu Arbeitsanforderungen und Itemlösungen 272

Tab. F-1 Interkorrelationsmatrix (Pearson Korrelation, zweiseitiges Signifikanzniveau) von Lernstrategien und Testleistung 275

Tab. F-2 Ergebnisse des Wald-Tests zur Identifikation von Differential Item Functioning auf Itemebene 275

Tab. F-3 OLS-Regression zur Kriterumsvalidierung nach Reduktion des Tests um „unfaire“ Items 278

Tab. F-4 Unterschiede zwischen nicht-situativen und situativen Items bezüglich der Bewertung der Facetten der Testakzeptanz sowie der curricularen Validität 279

Tab. F-5 Skalendokumentation für die Dimension „curriculare Validität" 279

Tab. F-6 Skalendokumentation für die Skala „Anstrengungsbereitschaft" 280

Abkürzungsverzeichnis

ACT	American College Testing
ACT-R	Adaptive Control of Thought Rational
AERA	American Educational Research Association
AHELO	Assessment of Higher Education Learning Outcomes
ANOVA	Analysis of Variance (Varianzanalyse)
APA	American Psychological Association
BA	Bundesagentur für Arbeit
BAKT	Business Administration Knowledge Test
BLK	Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung
BMBF	Bundesministerium für Bildung und Forschung
BWL	Betriebswirtschaftslehre
CFI	Comparative Fit Index
COACTIV	Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung von mathematischer Kompetenz
DIF	Differential Item Functioning
ECD	Evidence-Centered Design
ETS	Educational Testing Service
F-DUP-K	Fragebogen zur Diagnose unternehmerischer Potenziale
FSI	Fragebogen zum Studieninteresse
ICC	Item Characteristic Curve
ILLEV	Innovativer Lehr-Lernortverbund in der akademischen Hochschulausbildung
IRT	Item-Response-Theorie
IST 2000 R	Intelligenzstrukturtest 2000 R
KMU	Kleine und mittlere Unternehmen

KoMeWP	Modellierung und Erfassung fachwissenschaftlicher und fachdidaktischer Kompetenzen im wirtschaftspädagogischen Studium
KTT	Klassische Testtheorie
LCE	Leaving Certificate
LR	Likelihood Ratio
MFAT-B	Major Field Achievement Test in Business
ML	Maximum-Likelihood
n. s.	Nicht signifikant
NCME	National Council for Measurement in Education
OECD	Organisation for Economic Co-operation and Development
OEKOMA	Ökonomische Kompetenzen von Maturandinnen und Maturanden
OLS	Ordinary Least Squares Regression
PISA	Programme for International Student Assessment
RMSEA	Root Mean Square Error of Approximation
SAT	Scholastic Aptitude Test
SE	Standard Error
SJT	Situational Judgement Test
SPSS	Statistical Package for the Social Sciences
TEL	Test of Economic Literacy
TIE	Typisches intellektuelles Engagement
TIMSS	Third International Mathematics and Science Study
TUCE	Test of Understanding College Economic
ULME	Untersuchungen von Leistungen, Motivation und Einstellungen der Schülerinnen und Schüler in den Abschlussklassen der Berufsschulen
VWL	Volkswirtschaftslehre
WBT	Wirtschaftskundlicher Bildungs-Test

Wipäd	Wirtschaftspädagogik
WiwiKom	Modellierung und Messung wirtschaftswissenschaftlicher Fachkompetenz bei Studierenden bzw. Hochschulabsolventen
WLE	Weighted Likelihood Estimation
wLSMV	Weighted Least Squares Means and Variance adjusted
wMNSQ	Weighted Mean Square
WWT	Wirtschaftswissenstest
ZFA	Zahnmedizinische Fachangestellte

1 Einleitung

„Kompetenzorientierung", „Employability" und „Outcomeorientierung" sind drei zentrale Begriffe, die durch den Bologna-Prozesse (Bologna Declaration, 1999) Einzug in die deutsche Hochschullandschaft gehalten haben. Die Frage, was an deutschen Universitäten gelehrt und gelernt wird, rückt insbesondere wirtschaftswissenschaftliche Studiengänge immer mehr in das Interesse der Öffentlichkeit. Im Gegensatz zum öffentlichen Interesse liegen aus wissenschaftlicher Perspektive zu Fragen des Wissenserwerbs und der Wissensentwicklung im tertiären Bildungssektor nur wenige belastbare Ergebnisse vor. Universitäre Lehr-Lernprozesse und deren Ergebnisse gelten in der deutschen Bildungsforschung als „Black Box", mit deren Durchleuchtung erst jüngst begonnen wurde (Bülow-Schramm & Braun, 2013). Die Zurückhaltung bei der Erforschung von Bildungsergebnissen im Hochschulsektor ist nicht zuletzt einem Mangel an validen Instrumenten und einem damit einhergehenden Mangel an grundlegenden Konzepten der Testentwicklung und -auswertung in diesem Bereich geschuldet (Blömeke, 2013).

Die vorliegende Arbeit fokussiert theoretische sowie methodische Herausforderungen der Erfassung von Lernergebnissen im Hochschulsektor am Beispiel der Entwicklung und Validierung eines betriebswirtschaftlichen Wissenstests. Die Problemlage und die daraus resultierende Zielsetzung der Arbeit werden im folgenden Abschnitt beschrieben. In Abschnitt 1.2 erfolgt die Darstellung der Forschungsfragen und des forschungsmethodischen Vorgehens. Das Kapitel schließt mit einem Überblick über den Aufbau der Arbeit in Abschnitt 1.3.

1.1 Problemstellung und Zielsetzung

Die Pilotstudie der OECD zur „Erfassung von Lernergebnissen im Hochschulsektor" (AHELO) (OECD, 2011) unterstreicht, dass das internationale Forschungsinteresse an hochschulischen Lehr-Lernprozessen und deren Ergebnissen steigt. Doch während das Lehren und Lernen im deutschen Schulsystem seit rund 10 Jahren systematisch erforscht und national sowie international verglichen wird, wurde die Hochschule erst jüngst von der deutschen empirischen Bildungsforschung als Forschungsfeld erkannt (Blömeke, 2013). Entsprechend limitiert ist in diesem Bereich der Zugang zu wissenschaftlichen Erhebungsinstrumenten und grundlegenden Forschungsarbeiten, sodass sich eine Forschungslücke auftut, zu deren Schließung die vorliegende Arbeit beitragen soll. Doch welcher Nutzen ist von einer wissenschaftlichen Erfassung der Lernergebnisse an Universitäten zu erwarten? Hartig und Jude (2007) arbeiteten vier Ebenen heraus, auf denen von der wissenschaftlichen Erfassung von Lern-

ergebnissen profitiert werden kann: (1) die individuelle Ebene, (2) die institutionelle Ebene, (3) die Systemebene und (4) die Ebene der Forschung. Im Folgenden wird in Anlehnung an Seeber et al. (2010) beschrieben, wie die wissenschaftliche Erfassung von Lernergebnissen auf den jeweiligen Ebenen wirksam werden kann:

(1) Auf *individueller Ebene* ist die Erfassung von Lernergebnissen ein Baustein zur Aufdeckung von Lernprozessen, um diese gezielt zu steuern und zu unterstützen. Durch die Identifikation von Lernproblemen oder Begabungen können z. B. individuell angepasste Fördermaßnahmen angeboten werden (Seeber et al., 2010).

(2) Auf *institutioneller Ebene* dient die Erfassung von Lernergebnissen in erster Linie der Verbesserung institutioneller Lehr-Lern-Arrangements und der Weiterentwicklung der didaktischen Kultur (Seeber et al., 2010). Zum Beispiel können spezielle Programme für identifizierte Risikogruppen angeboten und evaluiert werden.

(3) Auf der *Ebene des Bildungssystems* können systematisch erfasste Lernergebnissen Auskunft über die Effektivität von Bildungsprogrammen und als Informationsquelle für die Systemsteuerung herangezogen werden (Seeber et al., 2010).

(4) Aus *Forschungsperspektive* bietet die Erfassung von Lernergebnissen die Möglichkeit, Forschungslücken zu schließen (Seeber et al., 2010) und Erkenntnisse zu generieren, die über die empirische Bildungsforschung hinaus nutzbar gemacht werden können. Dazu zählen Weiterentwicklungen im Bereich der Testentwicklung und -auswertung, die in weiten Teilen der quantitativ-empirischen Sozialforschung Verwendung finden.

Das deutsche Hochschulsystem weist verglichen mit der beruflichen und schulischen Bildung viele Besonderheiten auf, die es bei der Erfassung von Lernergebnissen zu berücksichtigen gilt. Diese Besonderheiten werden in der vorliegenden Arbeit identifiziert und bezogen auf die Entwicklung, Implementation, Auswertung und Interpretation von Assessments im Hochschulbereich diskutiert. Die Fokussierung auf die Erfassung betriebswirtschaftlichen Wissens von Studierenden ist unter anderem deshalb relevant, da Betriebswirtschaftslehre konstant der Studiengang mit den höchsten Immatrikulationszahlen in Deutschland ist (Statistisches Bundesamt, 2013). Im Jahr 2011 waren 199.477 Studierende in betriebswirtschaftlichen Studiengängen immatrikuliert. Das waren 7.9 % mehr Studierende als im Vorjahr und entsprach 8.38 % aller Studierenden im Jahr 2011 (Statistisches Bundesamt, 2013). Diese Zahlen umfassen noch nicht die Vielzahl von Studierenden, die betriebswirtschaftliche Module im Nebenfach oder mit Lehramtsperspektive belegen (z. B. Studierende der

Wirtschaftspädagogik). Darüber hinaus ist ökonomische Bildung von großer Bedeutung für die individuelle, wirtschaftliche sowie gesellschaftliche Teilhabe und ist in zahlreichen beruflichen Zusammenhängen relevant (Schumann, Oepke & Eberle, 2011). Betriebswirtschaftliches Wissen kann jedoch noch nicht als eine Entität betrachtet werden, sondern ist ein komplexes Konstrukt, das mindestens in zwei Wissensarten unterteilt werden kann. In zahlreichen Forschungsarbeiten hat sich eine Zweiteilung in deklaratives und prozedurales Wissen etabliert (z. B. Anderson, 1996). Diese Zweiteilung wird auch dieser Arbeit zugrunde gelegt. Es wird davon ausgegangen, dass deklaratives Wissen das in Schemata abgespeicherte Faktenwissen einer Person umschreibt. Deklaratives Wissen reicht in der Regel jedoch nicht aus, um in komplexen Situationen kompetent zu handeln (Gruber, Mandl & Renkl, 2000). Hierfür wird das Zusammenspiel von deklarativem und prozeduralem Wissen benötigt (Gruber, 1999). Prozedurales Wissen ist handlungsnah und beschreibt das Wissen über Methoden und Prozeduren und deren adäquate Anwendung (Anderson & Krathwohl, 2001).

Im Sinne der im Bologna Prozess (Bologna Declaration 1999) und im Qualifikationsrahmen für Deutsche Hochschulabschlüsse (HRK, KMK & BMBF, 2005) verankerten Kompetenz-und Employabilityorientierung des tertiären Bildungssektors sollten bei einer Erfassung von Lernergebnissen an Universitäten und Hochschulen handlungsnahe Wissensarten berücksichtigt werden. Zur Überprüfung des betriebswirtschaftlichen Wissenstandes von Studierenden liegt mit dem Business Administration Knowledge Test (BAKT) von Bothe (2003) im deutschen Sprachraum bisher jedoch nur ein wissenschaftliches Testinstrument vor, das entwickelt wurde, um deklaratives betriebswirtschaftliches Wissen zu erfassen.

Das Ziel der vorliegenden Forschungsarbeit ist, einen wissenschaftlich fundierten Test zu entwickeln und zu validieren, der über deklaratives Wissen hinaus auch anwendungsorientierte prozedurale Wissensbestände auf Bachelorniveau erfasst. Dafür werden in Anlehnung an Arbeiten aus dem Bereich der Berufs- und Wirtschaftspädagogik und der Personalauswahl situative Aufgaben entwickelt und unter Bezugnahme auf item-responsetheoretische Modelle quantitativ-empirisch analysiert und ausgewertet. Im Rahmen der Testauswertung sollen Erkenntnisse über die Struktur und die Determinanten betriebswirtschaftlichen Wissens gewonnen werden. Ein weiteres Ziel der Arbeit ist es, methodische Besonderheiten der Erfassung von Lernergebnissen im Hochschulsektor im Allgemeinen und mittels situativer Aufgaben im Speziellen herauszuarbeiten und somit die Grundlage für weitere Forschung in diesem Bereich auszubauen.

1.2 Forschungsfragen und forschungsmethodisches Vorgehen

Mit der Entwicklung eines Wissenstests gehen zahlreiche methodische Fragestellungen einher. Bevor ein Test zur Beantwortung inhaltlicher Fragen herangezogen werden kann, muss belegt werden, dass der Test den wissenschaftlichen Gütekriterien der pädagogisch-psychologischen Diagnostik entspricht. Die zentrale Frage bei der Bewertung eines neu entwickelten Tests lautet dementsprechend:

Ist der Test objektiv, reliabel und valide?

Das Vorgehen bei der Erstellung der Items orientiert sich am Ideal der deduktiven Testentwicklung (Hartig & Jude, 2007). Die theoretische Basis der Testentwicklung wird in weiten Teilen aus Erträgen der berufs- und wirtschaftspädagogischen Forschung sowie der empirischen Bildungsforschung abgeleitet. Der Abgleich zwischen den theoretisch intendierten und den tatsächlichen Testeigenschaften erfolgt quantitativ-empirisch. Dafür werden die Daten von ca. 500 Studierenden erhoben und nach aktuellen psychometrischen Standards unter Nutzung des Rasch-Modells ausgewertet.

Die aus dem Modell geschätzten Personenparameter dienen als Fähigkeitsindikatoren der Testteilnehmer und werden zur Prüfung der folgenden Forschungsfragen herangezogen:

1. **Welche Binnenstruktur weisen die situativen Items auf und inwiefern lassen sie sich empirisch von nicht-situativen Items abgrenzen?**
2. **Inwiefern unterscheiden sich Studierendengruppen bezüglich ihrer Testleistung?**
3. **Welche individuellen Determinanten sind für die Testleistung bedeutsam?**
4. **Ist der Test fair und wie wird der Test von den Testteilnehmern hinsichtlich verschiedener Akzeptanzkriterien wahrgenommen?**

Die Beantwortung der Forschungsfragen steht in erster Linie im Dienste der Absicherung der Validität der neu entwickelten Items. Darüber hinaus liefern sie erste Hinweise, welche Wissensarten sich Studierende an der Universität aneignen und welche Studierendengruppen mit welchen Eigenschaften günstige Voraussetzungen für den Erwerb von unterschiedlichen betriebswirtschaftlichen Wissensarten an Hochschulen aufweisen.

1.3 Aufbau und Struktur der Arbeit

Um die Forschungsfragen dieser Arbeit zu beantworten, ist die Arbeit in zehn Kapitel unterteilt. Die Kapitel können vier übergeordneten thematischen Schwerpunkten zugeordnet werden: (1) theoretische Fundierung und Darstellung des Forschungsstands, (2) Beschreibung des methodischen Vorgehens, (3) Darstellung der empirischen Ergebnisse und (4) Diskussion der Ergebnisse.

Die theoretische Basis der Arbeit wird in den *Kapiteln 2 bis 4* beschrieben.

In *Kapitel 2* wird die Verortung von Wissen als Teil professioneller beruflicher Handlungskompetenz vorgenommen. Das Wissenskonstrukt wird aus kognitionspsychologischer und didaktischer Perspektive in seine Teilbereiche untergliedert und deren Bedeutung für kompetentes berufliches Handeln dargelegt.

In *Kapitel 3* wird der Stand der Forschung zur Messung des wirtschaftswissenschaftlichen Wissens von Studierenden nachgezeichnet. Der Fokus liegt dabei auf nationalen sowie internationalen Ergebnissen zu Struktur und zu Determinanten betriebswirtschaftlichen Wissens. In diesen Zusammenhang werden aktuelle nationale und internationale Erhebungsinstrumente zur Erfassung betriebswirtschaftlichen Wissens vorgestellt. In *Kapitel* 3 wird aufgezeigt, dass speziell im deutschsprachigen Raum Forschungsdefizite im Bereich der Messung betriebswirtschaftlichen Wissens bestehen. Für die theoretische Fundierung der Testentwicklung wird deshalb auf Konzepte der Kompetenzmessung in kaufmännischen Handlungsfeldern zurückgegriffen, die wiederum in *Kapitel 4* beschrieben werden.

In *Kapitel 4* werden die Grundzüge der kaufmännischen Kompetenzmessung dargestellt. Dabei liegt der Fokus des Kapitels darauf, zu prüfen, welche Konzepte für die Messung betriebswirtschaftlichen Wissens adaptiert werden können und welche Konzepte unter kritischer Betrachtung nicht für diesen Zweck oder die Zielgruppe geeignet sind. Das besondere Augenmerk liegt einerseits auf der Strukturierung der betriebswirtschaftlichen Domäne und andererseits auf der Modellierung von Aufgabenschwierigkeiten durch kognitive Anforderungsstrukturen.

Die methodischen Grundüberlegungen zur Testentwicklung beginnen in *Kapitel 5* mit der Betrachtung situativer Messverfahren und -instrumente und gehen in *Kapitel 6* in die Beschreibung des Vorgehens bei der Testentwicklung über. Sie schließen mit der Darstellung der Standards pädagogisch-psychologischer Diagnostik in *Kapitel 7*.

In *Kapitel 5* werden die speziellen Eigenschaften situative Aufgaben diskutiert und begründet, warum ein papierbasierter situierter Ansatz anderen, zum Beispiel computerbasierten Verfahren vorgezogen wird.

In *Kapitel 6* werden Modelle der Testentwicklung vorgestellt. Anschließend werden die Besonderheiten der Anforderungen des Lernens und Arbeitens an Studierende und Bachelorabsolventen der Betriebswirtschaftslehre herausgearbeitet und in Form einer vereinfachten Domänenstruktur als Basis der Testentwicklung herangezogen. Das Rahmenmodell der Testentwicklung gibt vor, wie die im theoretischen Teil der Arbeit zusammengetragenen Kenntnisse im Test umgesetzt werden.

Kapitel 7 dient insbesondere der Vorbereitung auf den empirischen Teil der Arbeit. Hier werden alle grundlegenden Konzepte der empirischen Testentwicklung und Testauswertung beschrieben, die in der Arbeit genutzt werden. In Anbetracht der übergeordneten Fragestellung nach der Güte des Tests und dem Anspruch, den Test nach neusten psychometrischen Standards zu entwickeln, stehen die klassischen Gütekriterien sowie Überlegungen zu den probabilistischen Testtheorien im Vordergrund.

Die Darstellung der empirischen Auswertungen erfolgt in den *Kapiteln 8 und 9*.

In *Kapitel 8* werden als wichtiger Bestandteil der Arbeit das Vorgehen bei der Pilotierung sowie die Pilotierungsergebnisse beschrieben. Das Kernstück der vorliegenden Arbeit, die Darstellung der Validierung des Tests, erfolgt in *Kapitel 9*.

Das *Kapitel 9* beginnt mit der Ableitung der Hypothesen zu den oben aufgeführten zentralen Forschungsfragen und der Beschreibung der Datenerhebung sowie der Stichprobe. Im Anschluss folgt die Ergebnisdarstellung. Im ersten Schritt der Auswertungen wird geprüft, ob die Daten, die über den entwickelten Fragebogen erhoben wurden, mit Hilfe des Rasch-Modells ausgewertet werden können und welche dimensionale Struktur den Items zugrunde liegt. In den weiteren Schritten werden die Konstruktvalidität und die Kriteriumsvalidität anhand unterschiedlicher Berechnungen überprüft. Abschließend wird geprüft, inwiefern der Test im statistischen Sinne als fair bezeichnet werden kann und wie die am Test teilnehmenden Personen die situativen Items im Vergleich zu den nicht-situativen Items einschätzen.

In *Kapitel 10* werden zunächst die Ergebnisse der Arbeit zusammenfassend dargestellt und weiterführend interpretiert. Vor dem Hintergrund, dass die vorliegende Arbeit richtungweisend für weiterführende Arbeiten sein soll, widmet sich ein großer Teil der Diskussion der Darstellung von Limitationen hinsichtlich

der theoretischen Basis und der methodischen Umsetzung der Arbeit. Aus diesen Limitationen werden Verbesserungsvorschläge für zukünftige Arbeiten im Forschungsfeld der Erfassung von betriebswirtschaftlichem Wissen herausgearbeitet sowie zentrale Ergebnisse dieser Arbeit als Basis für weitere Forschung herausgestellt. Die Arbeit schließt mit einer kritischen Betrachtung der Funktion von Assessments im Hochschulsektor und einem integrativen Fazit sowie einem Ausblick auf zukünftige Forschungsfragen.

2 Wissen als Bestandteil professioneller Handlungskompetenz

Hochschulen und Universitäten kommt eine tragende Rolle als Wissensvermittler zu. Vor dem Hintergrund der aktuellen Relevanz des wirtschaftlichen Wissens von Studierenden ist dieses Kapitel (Kapitel 2) den theoretischen und empirischen Grundlagen der wissenschaftlichen Betrachtung von Wissen gewidmet. Die Relevanz der universitären Bildung in der Wissensgesellschaft wird in Abschnitt 2.1 herausgearbeitet. Das Verhältnis von Wissen zu Kompetenzen wird in Abschnitt 2.2 aufgegriffen. In Abschnitt 2.3 wird zwischen verschiedenen Wissensarten unterschieden. Darauf aufbauend werden in Kapitel 3 der Stand der Forschung zur Erfassung wirtschaftlichen Wissens vorgestellt und nationale sowie internationale Projekte und Instrumente in diesem Bereich näher erläutert.

2.1 Die Relevanz universitärer Bildung in der Wissensgesellschaft

Deutschland hat sich in den letzten Jahren stetig in Richtung einer Wissensgesellschaft entwickelt. Das heißt, im Vergleich zu vorausgegangenen Gesellschaften hat die Bedeutung von Wissen stetig zugenommen (Zöllösi-Janze, 2004). Dieser Strukturwandel wirkt sich auf die Tätigkeits- und Qualifikationsstruktur des Arbeitsmarktes aus (Weishaupt, 2010). Es zeichnet sich ein Trend zu immer anspruchsvolleren Tätigkeiten, mit der Folge stetig steigender Qualifikationsanforderungen an Berufstätige (Dostal & Reinberg, 1999), ab. Diese Entwicklung führt zu einem steigenden Bedarf an Hoch- und Fachhochschulabsolventen (Dostal & Reinberg, 1999; Weishaupt, 2010). Wissen beeinflusst nicht nur die internationale Konkurrenzfähigkeit des Landes, sondern auch den persönlichen Lebensweg des Einzelnen (Mandl & Krause, 2001). Hochschulen haben die Funktion, den Lernenden auf die Anforderungen in der Wissensgesellschaft vorzubereiten und ihm somit eine aktive Teilhabe am gesellschaftlichen Leben zu ermöglichen (Mandl & Krause, 2001). Insbesondere der Bildung in wirtschaftlichen Belangen wird dabei verstärkt Bedeutung zugesprochen (Wiepcke, 2009). Das Verständnis komplexer ökonomischer Sachverhalte wird sowohl im Arbeitsleben als auch in vielen privaten Bereichen vorausgesetzt. Gut ausgebildete Wirtschaftswissenschaftler sind für Unternehmen aller Größen gleichermaßen ein Schlüssel zum unternehmerischen Erfolg (Größler, Wilhelm, Wittmann & Milling, 2002). Nicht nur für angestellte Mitarbeiter, sondern auch für den Erfolg von Unternehmensgründungen, die wiederum

volkswirtschaftlich und arbeitsmarktpolitisch von großer Bedeutung sind, ist wirtschaftliches und speziell betriebswirtschaftliches Wissen eine wesentliche Voraussetzung (Kailer et al., 2013).

Von Hochschulen und anderen Bildungsinstitutionen fordert der beschriebene Wandel der Qualifikationsstrukturen am Arbeitsmarkt die Auseinandersetzung mit den eigenen Konzepten und Zielen der Lehre und die Bereitschaft, sich veränderten Gegebenheiten anzupassen (Teichler, 2005). Mit der Implementierung des Bologna-Prozesses (1999) im tertiären Bildungssektor wurde ein Wandel von Prozessorientierung hin zu Output-und Outcome-Orientierung angestoßen. Als wichtige Lernziele der universitären Bildung werden in diesem Rahmen erworbene Kompetenzen geltend gemacht (Schaeper & Wolter, 2008). Dabei werden sowohl der Aufbau fachspezifischer als auch fachübergreifender Kompetenzen als Zielvorgabe formuliert (Bologna Working Group, 2005). Der Qualifikationsrahmen für deutsche Studienabschlüsse schließt sich dieser Forderung an und systematisiert die Lernziele von Studiengängen entlang der Kategorien „Wissen und Verstehen", „Können" und „Formale Aspekte" (HRK, KMK & BMBF, 2005, S. 2). Mit der Anforderung an Universitäten, gezielt den Aufbau von Kompetenzen zu fördern, geht konsequenterweise die Frage einher, wie man diese erfassen kann und welche Bedingungen den Aufbau dieser Kompetenzen begünstigen oder erschweren. Darüber hinaus gilt es zu klären, welche Vorstellungen hinter dem Kompetenzbegriff stehen und inwiefern sich dieser vom Wissensbegriff abgrenzt.

2.2 Das Verhältnis von Wissen zu Kompetenzen

Der Begriff der Kompetenz ist in der deutschen Bildungsforschung allgegenwärtig (Klieme & Hartig, 2007). Im allgemeinen Sprachgebrauch werden Kompetenzen oft als Fähigkeit, Begabung, Talent, Qualifikation und Leistungsvermögen verstanden. Aus wissenschaftlicher Perspektive ist eine Konsensbildung zum Kompetenzbegriff noch nicht abgeschlossen. Zwar wird dieser disziplinübergreifend kontrovers diskutiert, bisher jedoch ohne das Resultat einer allgemeingültigen Definition (Klieme & Hartig, 2007). Die vorliegende Arbeit fokussiert als Untersuchungsobjekt gezielt das Wissen und nicht die Kompetenzen von Studierenden, jedoch wird Fachwissen als elementarer Bestandteil beruflich kompetenten Handelns verstanden. Entsprechend lehnen sich die Testentwicklung und Testauswertung an Errungenschaften und Konzepte der kognitiv orientierten Kompetenzmessung, insbesondere aus dem beruflichen Bereich, an (vgl. Kapitel 4).

Die Eingrenzung des interessierenden Konstrukts ist die Grundvoraussetzung für dessen gezielte, wissenschaftliche Erfassung (Klieme & Hartig, 2007). Ziel

dieses Abschnitts ist es, den Kompetenzbegriff einzugrenzen und Wissen als Teil der beruflichen Handlungskompetenz zu verorten. Der Begriff der Kompetenz spielt in den Bereichen eine Rolle, in denen menschliche Leistungsfähigkeit unter einer anwendungsorientierten Perspektive betrachtet wird (Klieme & Hartig, 2007, S. 128). Trotz der Anforderung an das deutsche Bildungssystem, Bildungsziele in Form von Kompetenzen zu charakterisieren (Klieme & Hartig, 2007), ist die Bestimmung des Kompetenzbegriffs keineswegs eindeutig.

Eine historisch wichtige erste Konzeptualisierung von Kompetenzen erfolgte durch Roth (1971), der eine systematische Einteilung der Handlungskompetenz in Selbst-, Sach-, und Sozialkompetenz vornahm. Selbstkompetenz umschreibt die Fähigkeit des selbstverantwortlichen Handelns, Sachkompetenz beschreibt die Urteils- und Handlungsfähigkeit bezogen auf bestimmte Handlungsbereiche und Sozialkompetenz die Handlungsfähigkeit in sozial-, gesellschaftlich und politisch relevanten Bereichen (Klieme & Hartig, 2007). Diese Kompetenztrias, insbesondere die Ausdifferenzierung der Sachkompetenz in Fach- und Methodenkompetenz, hat den Kompetenzdiskurs der Erziehungswissenschaften nachhaltig geprägt (Klieme & Hartig, 2007). Neben den anthropologisch verankerten Kompetenzüberlegungen von Roth (1971), haben auch funktional-psychologische Sichtweisen auf Kompetenzen Eingang in das heutige Kompetenzverständnis der empirischen Bildungsforschung gefunden. Diese grenzt sich von der traditionellen dekontextualisierten Intelligenzdiagnostik (z. B. Sternberg, 1982) ab, in dem die Fähigkeit einer Person, situative Anforderungen des „realen Lebens" zu bewältigen, in den Mittelpunkt des Interesses gestellt wird (Klieme & Hartig, 2007). Es wird davon ausgegangen, dass Kompetenzen im Bereich einer bestimmten Domäne wirksam werden. Der Domänenbegriff umschreibt fachspezifische Leistungsbereiche, die sich anhand selektierter Anforderungssituationen charakterisieren lassen (Winther, 2010). Mit der Definition von Kompetenzen als „kontextspezifische kognitive Leistungsdispositionen, die sich funktional auf Situationen und Anforderungen in bestimmten Domänen beziehen", betonen Klieme und Leutner (2006, S. 879) sowohl die Domänenspezifizität als auch den funktionalen Charakter von Kompetenzen. Der Domänenbegriff und dessen Spezifikation im kaufmännischen und betriebswirtschaftlichen Bereich wird in Abschnitt 4.1 erneut aufgegriffen und näher erläutert. Diese Bezugnahme auf spezifische Situationen und Anforderungen legt nahe, dass Kompetenzen durch die Auseinandersetzung mit den entsprechenden Situationen prinzipiell erlernbar sind (Klieme & Hartig, 2007). Eine weitergefasste Kompetenzdefinition erfolgt durch Weinert (2002). Dieser betrachtet Kompetenzen, aus einer kognitionspsychologischen Perspektive, als

> *„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, 2002, S. 27 f.).*

Neben der Erlernbarkeit von Kompetenzen und deren zielgerichteten, auf Situationen bezogenen, Funktionalität betont Weinert (2001) motivationale, volitionale und soziale Bereitschaften als Bestandteil von Kompetenzen. Der so verstandene Kompetenzbegriff umschreibt in seiner Gesamtheit nicht nur das Können, sondern auch das „Können wollen". Über das Vorhandensein von kognitiven und motorischen Fähigkeiten hinaus, zeigt sich Kompetenz in der tatsächlichen Manifestation dieser Fähigkeiten in zielgerichteten Handlungen. Der Begriff der beruflichen Handlungskompetenz umschließt die genannten Kompetenzcharakteristika und bezieht sich jeweils auf die Anforderungen und Aufgaben eines beruflichen Handlungsfeldes (Klieme, 2004).

Dieser Arbeit liegt ein Kompetenzverständnis zugrunde, das zum einen auf den Annahmen der Erlernbarkeit und der Domänenspezifizität beruht, zum anderen dem Kompetenzverständnis der Definition von Weinert (2001) folgt. Kognitive Fähigkeiten und Fertigkeiten werden nur in Kombination mit motivationalen, volitionalen und sozialen Bereitschaften als Kompetenz verstanden. Domänen- oder professionsspezifisches Wissen ist dementsprechend ein zentraler Bestandteil beruflicher Handlungskompetenz (Baumert & Kunter, 2006). Die vorliegende Arbeit betrachtet das Wissen von Studierenden als elementaren Bestandteil der beruflichen Handlungskompetenz. Überzeugungen und Werthaltung sowie motivationale Orientierung und selbstregulative Fähigkeiten werden in dieser Arbeit nicht, oder nur randständig, betrachtet.

Eine Darstellung des Verhältnisses von domänenspezifischem Wissen zu der übergeordneten beruflichen Handlungskompetenz erfolgt in Anlehnung an die Studie „Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung von mathematischer Kompetenz" (COACTIV) (Baumert & Kunter, 2006). Das Modell wurde ursprünglich für Kompetenzen im Lehrerberuf entwickelt und wird in Abbildung 1 in adaptierter Form dargestellt. Der Gegenstandsbereich der vorliegenden Arbeit ist durch den mit einer gestrichelten Linie umrahmten Bereich kenntlich gemacht.

Nachdem verdeutlicht wurde, dass domänenspezifisches Wissen im Rahmen der aktuellen Kompetenzorientierung eine zentrale Rolle spielt, wird im Folgenden der Wissensbegriff näher beleuchtet.

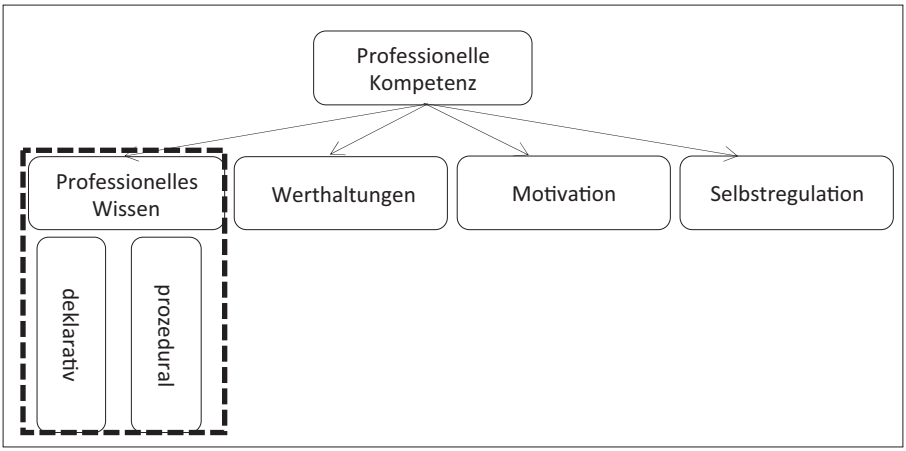


Abb. 1: Strukturierung des Untersuchungsgegenstandes „professionelles Wissen" als Facette professioneller Handlungskompetenz in Anlehnung an Baumert und Kunter (2006, S. 482), ergänzt durch die Unterteilung von Wissen in ein deklaratives und ein prozedurales System (z. B. Anderson, 2001)

Wissen kann auf unterschiedliche Weise im Gehirn abgespeichert werden und ist somit auch in verschiedenen Anforderungssituationen unterschiedlich schnell abrufbar und nutzbar (Anderson, 2001). In der Regel wird in diesem Zusammenhang zwischen verschiedenen Wissensarten unterschieden. Eine differenzierte Betrachtung der Wissensarten ist im besonderen Maße für das Konstruktverständnis und die Operationalisierung wirtschaftswissenschaftlichen Wissens in der vorliegenden Arbeit wichtig und erfolgt im folgenden Abschnitt.

2.3 Die Unterscheidung der Wissensarten

Definitionen des Wissensbegriffs erfolgen in der Regel disziplinspezifisch. Vor dem Hintergrund unterschiedlicher Fragestellungen variieren dabei die Schwerpunktsetzungen. Die unstrittige Relevanz von Wissen für die unterschiedlichsten Forschungsdisziplinen erschwert jedoch eine allgemeingültige Begriffsbestimmung und verleiht dem Wissensbegriff die Charakteristika eines „Komplexbegriffs" (Gottschalk-Mazouz, 2007). Wissen im alltagssprachlichen Sinne bezieht sich vornehmlich auf die Kenntnisse einer Person in einem bestimmten Bereich (Renkl, 2009). Auch wird Wissen üblicherweise als personenbezogenes Merkmal verstanden (Gruber, 1999). Für die Genese kompetenten Handelns ist in der Regel ein Zusammenspiel unterschiedlicher Wissensarten und Wissensorganisationsformen nötig (Gruber, 1999). Aus dieser Annahme

resultiert die Anforderung, den Wissensbegriff nicht als geschlossene Entität zu betrachten, sondern genauer zu analysieren und zu kategorisieren. Eine Unterteilung in systematisch unterscheidbare Wissensarten legt den Grundbaustein für die differenzierte Betrachtung des Wissens.

Die grundlegende Idee der Unterscheidung von Wissensarten geht auf den Philosophen Gilbert Ryle (1949) zurück. In seinem vielzitierten Werk „The concept of Mind" differenzierte er im Zusammenhang mit Problemlöseleistungen erstmals zwischen dem anwendungsorientierten „knowing how" und dem praxisfernen „knowing that". Ryle (1949) kritisierte die zur damaligen Zeit vorherrschende Tendenz zur Überbewertung des intellektuellen „knowing that" und die Vernachlässigung des handlungsnahen „knowing how". Einem intelligenten Menschen schreibt er den Besitz beider Wissensarten zu: „To be intelligent is not merely to satisfy criteria, but to apply them […]"(Ryle 1949, S. 28 ff.).

Seither wurde die Trennung der Wissensarten von vielen unterschiedlichen Forschern und Forschungsrichtungen aufgegriffen und modifiziert. Diese vielschichtige Rezeption des Konzepts der Trennung von Wissensarten begünstigte eine Vermischung unterschiedlicher Wissensbegriffe, die die Abgrenzung der Begrifflichkeiten erschwert. Raju, Lonial und Mangold (1995) unterschieden zwischen subjektivem Wissen, objektivem Wissen und Gebrauchserfahrung, während sich de Jong und Ferguson-Hessler (1996) bemühten, eine Strukturierung der Wissensarten nach Wissenstyp und Funktion vorzunehmen. De Jong und Ferguson-Hessler (1996) nennen konzeptuelles, situatives, prozedurales und strategisches Wissen als zentrale Wissenstypen und unterscheiden innerhalb der Wissenstypen zwischen Ebene, Struktur, Grad der Automatisierung, Modalität und Generalisierung. Diese Systematisierung wurde jedoch weder von der kognitionspsychologischen Forschung noch von der didaktischen Forschung vollständig übernommen. Zahlreiche weitere Wissensbezeichnungen, wie zum Beispiel Metawissen (Anderson & Krathwohl, 2001), explizites und implizites Wissen (Anderson, 2001), Experten- und Alltagswissen (Edelmann, 1996) finden in der Forschungspraxis Anwendung. Wie verwoben die Nutzung unterschiedlicher Wissensbegriffe ist, zeigt Schneider (2005) eindrucksvoll in einer Zitationsanalyse der synonymen Verwendung von Namen für Wissensarten. Schneider (2005) konzentrierte sich bei seiner Analyse auf das Begriffspaar „konzeptuelles" und „prozedurales" Wissen und dessen direkt und indirekt verwendeten Synonyme.

Zahlreiche Theorien der Wissenskategorisierung stellen auf die Existenz zweier Wissenssysteme ab. Theorien, die dieser Zweiteilung des Wissens folgen, werden als dual component oder dual process Theorien der Kognition bezeichnet (Schneider, 2005). Die bekannteste unter diesen Theorien wird von dem Kognitionswissenschaftler John R. Anderson (2001) vertreten. Die Grundlagen ei-

ner elaborierten Differenzierung der Wissensarten mit einem besonderen Fokus auf deklarativem und prozeduralem Wissen legte Anderson in der ACT-Theorie (Adaptive Control of Thought) (Anderson, 1983).

Im Folgenden werden zwei Perspektiven auf die differenzierte Beschreibung von Wissen eingenommen, die für die vorliegende Arbeit relevant sind. In Abschnitt 2.3.1 wird die kognitionspsychologische Perspektive aufgearbeitet und in Abschnitt 2.3.2 durch eine didaktische Perspektive ergänzt.

2.3.1 Die kognitionspsychologische Perspektive

Die Kognitionspsychologie erforscht menschliche kognitive Prozesse. Zu Kognition zählen unter anderem Prozesse des Wahrnehmens, Erkennens, Vorstellens, Urteilens, Lernens, Denkens und Gedächtnisleistungen (Wagenknecht, 1980). Zusammengefasst handelt es sich beim Untersuchungsgegenstand der Kognitionspsychologie um menschliche Prozesse der Informationsverarbeitung (Anderson, 2001). Die kognitionspsychologische Perspektive ist für die vorliegende Arbeit zentral, weil durch Kognitionen Wissen erworben wird (Edelmann, 1996). Eine der prominentesten und vielfach beschriebenen Wissenskategorisierungen ist die, bereits beschriebene, an Ryle (1948) angelegte Dyade deklaratives Wissen versus prozedurales Wissen (Anderson, 1976; Gruber, 1999). Deklaratives Wissen wird gemeinhin als Faktenwissen bezeichnet, welches im Gedächtnis in Form von Schemata vorliegt und als „Wissen was" beschrieben wird. Prozedurales Wissen wird hingegen mit dem handlungsnahen, in kognitiven Operationen verinnerlichten „Wissen wie" umschrieben (Wittmann, Süß & Oberauer, 1996). Prozedurales Wissen ist dadurch gekennzeichnet, dass es situations- und ablauforientiert organisiert ist (König & Blömeke, 2009). Die Nähe prozeduralen Wissens zu situativen Anforderungen legt den Schluss nahe, dass prozedurales Wissen ein zentraler Bestandteil beruflicher Handlungskompetenz ist (vgl. Abschnitt 2.2). Jedoch reichen prozedurale Wissensbestände alleine nicht aus, um in komplexen (beruflichen) Situationen kompetent zu handeln (Gruber, 1999). Vielmehr ist davon auszugehen, dass ein Zusammenspiel verschiedener Wissensarten erforderlich ist, um kompetentes Handeln zu generieren (Gruber, 1999).

Die Interaktion von deklarativem und prozeduralem Wissen bei komplexen kognitiven Prozessen ist ein zentraler Bestandteil der ACT-Theorie von dem Kognitionspsychologen J.R. Anderson (Anderson, 2001). ACT steht für Adaptive Control of Thought und bildet ein Modell kognitiver Funktionsweisen, das im Jahr 1974 seine erste Ausarbeitung fand und im Jahr 1983 erstmals publiziert wurde (Anderson, 1983). Seither existieren zahlreiche Weiterentwicklungen und Anwendungen, sowohl durch den Autor selbst (z. Anderson, 1996) als auch durch andere Forscher (z. B. Lovett, Daily & Reder, 2000). Die bekann-

teste Weiterentwicklung der ACT-Theorie ist die ACT-R Theorie, in der das R für „rational" steht (Anderson & Lebiere, 1998). Die Grundzüge der ACT-R Theorie und deren Bezüge zu der vorliegenden Arbeit werden im Folgenden dargestellt.

Die ACT-R Theorie ist, wie ihr Vorgängerwerk, eine kognitive Architektur. Das heißt, sie ist eine Theorie darüber, wie Kognition funktioniert und wie diese modellhaft nachgebaut werden kann. Der modellhafte Charakter der ACT-R erlaubt die experimentelle Testung einzelner Modellannahmen. Eine Grundannahme der ACT-R Theorie ist, dass es zwei verschiedene Gedächtnissysteme gibt. Das deklarative Gedächtnissystem ist in Form von sogenannten „Chunks" gespeichert, die eine Einheit der Wissensrepräsentation bezeichnen (Anderson, 2001). Prozedurales Wissen ist hingegen in Form von Produktionen repräsentiert. Produktionen sind Regeln, die zur Aktivierung der in Chunks organisierten Wissensbestände herangezogen werden (Anderson, 2001).

Laut Anderson (2001) liegt der Ursprung des prozeduralen Wissens in den Prozessen des Problemlösens. Dabei werden Problemsituationen bei der Lösung in Teilziele zerlegt. Für diese Teilziele besitzt der Problemlösende sogenannte Operatoren. Problemlöseoperatoren helfen dem Problemlösenden, von einer gegebenen Ausgangssituation zu der gewünschten Zielsituation zu gelangen. Die formale Darstellung dieser Problemlöseoperatoren erfolgt in Produktion oder ganzen Produktionssystemen. Ein typisches Beispiel für eine Problemlöseproduktion besteht aus einem Ziel, mehreren Überprüfungen zur Anwendbarkeit der Regel und einer Aktion (Anderson, 2001). In Abbildung 2 ist eine einfache Form einer Produktionsregel dargestellt.

Wenn	das Ziel darin besteht, ein Auto mit Schaltgetriebe zu fahren,
	und der erste Gang eingelegt ist
	und das Auto schneller als 20 Kilometer die Stunde fährt,
Dann	lege den zweiten Gang ein.

Abb. 2: Beispiel für eine typische Produktionsregel (Quelle: Anderson, 2001, S. 525)

Die Produktion enthält einen „Wenn-Teil" (Quelle: Anderson, 2001, S. 525) mit Bedingungscharakter und einen „Dann-Teil" als Aktion. In der Bedingung ist notwendigerweise das Ziel der Aktion definiert sowie eine variable Anzahl an situativen Prüfungen, ob die Regel anwendbar ist. Fällt diese Prüfung positiv aus, kann die Aktion durchgeführt werden. Wesentlich ist, dass es für Problemlöseoperatoren einen Lösungsraum gibt, das heißt mehrere Lösungen können dazu führen, dem gewünschten Zielzustand näherzukommen. Prozedura-

les Wissen spiegelt sich darin wider, aus der Menge an Operatoren denjenigen auszuwählen, der für die Zielerreichung am angemessensten ist. Die Produktionsregeln enkodieren somit die Problemlöseoperatoren als Bedingungs-Aktions-Regeln (Anderson, 2001). Problemlöseoperatoren werden formal in Produktionssystemen dargestellt. Der Erwerb von Operatoren kann durch Entdecken, durch Analogie zur Lösung eines Beispielproblems oder durch direkte Instruktion erfolgen (Anderson, 2001).

2.3.2 Die didaktische Perspektive

Während sich Andersons ACT-Theorien aus der kognitionspsychologischen Perspektive in erster Linie mit der grundlegenden Verankerung unterschiedlicher Wissensarten im Gedächtnis und dessen Strukturen befassen, haben Anderson und Krathwohl (2001) eine Taxonomie der Wissensarten erstellt, die Aufschluss darüber gibt, wie unterschiedliche Verarbeitungsprozesse beim Lernenden durch Aufgaben angeregt werden können. Somit ist die Taxonomie der Wissensarten von Anderson und Krathwohl (2001) auf eine didaktische Nutzung ausgerichtet und nicht auf strenge Modellbildung und dessen empirische Überprüfbarkeit. Dementsprechend liegen bisher keine umfassenden empirischen Prüfungen der Gültigkeit der gesamten Taxonomie vor. Jedoch hat diese sich für die Kategorisierung und Strukturierung von Testaufgaben in Studien zur Kompetenz- und Wissenserfassung als nützlich erwiesen (Hofmeister, 2005; Lehmann & Seeber, 2003; Riese & Reinhold, 2012).

Historisch betrachtet handelt es sich bei der Wissenstaxonomie von Anderson und Krathwohl (2001) um eine Weiterentwicklung einer Taxonomie von Lernzielen im kognitiven Bereich nach Bloom (1972). Die Taxonomie nach Anderson und Krathwohl (2001) benennt zum einen die kognitiven Prozesse und zum anderen die Leistungsdimensionen, die mit bestimmten Anforderungssituationen einhergehen. Es wird versucht, diese beiden Dimensionen und ihre Ausprägungen in einer Matrix auszudifferenzieren und zu beschreiben. Die Trennung der Wissensdimension von der Leistungsdimension ist eine zentrale Weiterentwicklung im Vergleich zu der Taxonomie von Bloom (1972). Die Leistungsdimension wird konsequent durch Verbformen beschrieben, während die Wissensdimension in vier Dimensionen mit jeweils unterschiedlichen Subkategorien unterteilt wird. Anderson und Krathwohl (2001) unterscheiden auf der Ebene der Wissensdimensionen Faktenwissen, Konzeptwissen, prozedurales Wissen und Meta-kognitives Wissen. Auf der Ebene der kognitiven Prozesse wird zwischen Erinnern, Verstehen, Anwenden, Analysieren, Evaluieren und Kreieren unterschieden. Die überarbeitete Taxonomie ist in Tabelle 1 dargestellt.

Tab. 1: Die überarbeitete Taxonomie der Wissensformen

Die Wissens-dimension	Die kognitive Prozessdimension					
	Erinnern	**Ver-stehen**	**An-wenden**	**Ana-lysieren**	**Eva-luieren**	**Kreieren**
Faktenwissen						
Konzeptwissen						
Prozedurales Wissen						
Metakogniti-ves Wissen						

(Quelle: Anderson & Krathwohl, 2001, S. 28, frei übersetzt)

Aufgrund der didaktischen Zweckbestimmung der Taxonomie kann nicht davon ausgegangen werden, dass die beiden Achsen der Taxonomie im statistischen Sinne unabhängig voneinander modelliert werden können. Für die Kategorisierung von Testaufgaben ist die Taxonomie in ihrer ursprünglichen Form aufgrund der fehlenden Trennschärfe zwischen den Kategorien nur bedingt geeignet (Müller, Fürstenau & Witt, 2007). Für die Hamburger Studien zur Untersuchung von Leistungen, Motivation und Einstellungen der Schülerinnen und Schüler in den Abschlussklassen der Berufsschulen (ULME I-III) (Lehmann & Seeber, 2007) wurde eine komprimierte Taxonomie entwickelt. Die reduzierte Klassifikationsmatrix (Hofmeister, 2005) diente dazu, bei der Testentwicklung ein Analysetool zur Hand zu haben, das bei der systematischen Itemerstellung hilfreich ist. Das mit den Aufgaben anvisierte Leistungsniveau, bezogen auf das aus den entsprechenden Inhaltsbereichen ausgewählte und in den Aufgaben repräsentierte Wissen, soll somit festgestellt werden (Hofmeister, 2005).

Die reduzierte Klassifikationsmatrix nach ULME (Hofmeister, 2005) verzichtet auf den Einbezug metakognitiven Wissens und fasst die sechsstufige Prozessdimension in die drei Stufen Reproduzieren, Anwenden/Verstehen und Kritisieren/Reflektieren zusammen.

Beim Reproduzieren wird lediglich das Wiedererkennen und Wiedergeben von Informationen abverlangt (Hofmeister, 2005). Anzuwenden umschreibt die Fähigkeit, Gelerntes sinngemäß abzubilden. Der Lernende soll zeigen, dass er die Lerninhalte verstanden hat. Sprachlich und inhaltlich neue Aufgaben können so vom Lernenden gelöst werden. Vorhandene Handlungsschemata werden unverändert auf neue Situationen angewendet. Typische Formulierungen bei Anwendungsaufgaben sind z. B. Beschreiben, Erklären, Interpretieren, Begründen, Verstehen (Hofmeister, 2005).

Kritisches Reflektieren umschreibt die Fähigkeit, einen Sachverhalt umfassend, anhand von Kriterien, zu untersuchen. Die Lösung der Aufgabenstellung kann nur unter Hinzunahme neuer Kriterien generiert werden. Es findet ein Perspektivwechsel statt, der zum Aufbau neuer Wissensstrukturen führt. Zu diesem Prozess gehört, dass Kriterien selbständig gewichtet werden und eine eigenständige Einschätzung getroffen wird. Werden in einer Aufgabe Kriterien als auch deren Gewichtung genannt, handelt es sich um den kognitiven Prozess des Anwendens (Hofmeister, 2005).

Obwohl sich die Klassifikation von Aufgaben entlang des kognitiven Anforderungsniveaus bei der Testentwicklung als hilfreich erwiesen hat, ist deren empirische Gültigkeit umstritten. So zeigten Lehmann und Seeber (2007) bei Analysen zum Zusammenhang von Aufgabenschwierigkeit und a priori Aufgaben-Klassifikation, dass das kognitive Anforderungsniveau keinen bedeutsamen Anteil an der Aufklärung der empirischen Schwierigkeit im Fachleistungstest Wirtschaft für den Ausbildungsberuf „Kaufmann/Kauffrau im Einzelhandel" beiträgt. Vergleichbare Ergebnisse wurden für angehende Bürokaufleute gefunden (Seeber, 2008). Hier zeigte sich, wie schon bei den Kaufleuten im Einzelhandel, dass die Wissensart einen bedeutenden Einfluss auf die Aufgabenschwierigkeit hat. So waren als prozedural eingestufte Items deutlich schwerer als Aufgaben, die Kozept- oder Faktenwissen erfassen sollten. Auch bei Analysen der Aufgabenschwierigkeiten im WBT (Beck, Krumm & Dubs 1998) entsprechen die empirischen Ergebnisse nicht den theoretischen Annahmen der Bloomschen Taxonomie (Bloom, 1972). Aufgaben aus der Kategorie „Anwenden" waren leichter als Aufgaben aus der Kategorie „Verstehen" ebenso wie Items aus der Kategorie „Evaluation" höhere Schwierigkeitsindizes erzielten als Items aus der Kategorie „Analyse" (Witt, 2006). In einer Studie zu Kompetenzen von Kfz-Mechatronikern konnten hingegen im Rahmen einer post-hoc Schwierigkeitsanalyse 45 % der Varianz der Aufgabenschwierigkeit durch die Bloomsche Taxonomie (Bloom, 1972) aufgeklärt werden (Nickolaus, Gschwendtner & Geißel, 2008). Die sich widersprechenden empirischen Ergebnisse unterstreichen die Notwendigkeit über das kognitive Anforderungsniveau hinaus weitere schwierigkeitsbestimmende Aufgabenmerkmale zu identifizieren und deren Eignung für die Konstruktion von Wissenstests zu prüfen. Das Konzept der schwierigkeitsbestimmenden Aufgabenmerkmale und dessen Verwendung in der vorliegenden Arbeit werden in Abschnitt 4.2.2 vertieft.

Im Folgenden werden anhand der Arbeit von Anderson und Krathwohl (2001) deklaratives und prozedurales Wissen fokussiert.

Deklaratives Wissen

Deklaratives Wissen wird oft mit dem Begriff Faktenwissen umschrieben (Götz, Frenzel & Pekrun, 2009). Faktenwissen in einem bestimmten Fachgebiet bildet die Grundlage für gelungene Kommunikation in diesem Bereich. Ohne Faktenwissen währen die Ausführung vieler beruflicher Tätigkeiten oder der Austausch über diese undenkbar. Jedes Fachgebiet erfordert ein grundlegendes Maß an Faktenwissen, um Probleme in diesem Gebiet erfolgreich lösen zu können. Anderson und Krathwohl (2001, S. 42) weisen darauf hin, dass Lernende bei einer starken Fokussierung auf Faktenwissen Gefahr laufen, sich „träges Wissen" anzueignen. Wissen wird als „träge" bezeichnet, wenn es in Anwendungssituationen nicht nutzbar gemacht werden kann (Gruber, 1999). In der Regel wird davon ausgegangen, dass sich Faktenwissen auf einem relativ niedrigen Abstraktionsgrad befindet. Wie in Tabelle 2 dargestellt, bezieht sich das Faktenwissen nach Anderson und Krathwohl (2001) weitgehend auf Wissen über Terminologien und spezifische Details von Elementen.

Tab. 2: Das Faktenwissen und seine Unterformen

Wissensart und ihre Unterformen	**Beispiele**
Faktenwissen	Grundlegende Elemente, die Lernende wissen müssen, um mit einer Disziplin vertraut zu sein oder Probleme innerhalb der Disziplin zu lösen
A. Wissen über Terminologien	Technische Vokabel, musikalische Symbole
B. Wissen über spezifische Details und Elemente	Zentrale natürliche Ressourcen, verlässliche Informationsquellen

(Quelle: Anderson & Krathwohl, 2001, S. 45–46, frei übersetzt)

Das Konzeptwissen nach Anderson und Krathwohl (2001) ist hingegen auf einer höheren Abstraktionsstufe angesiedelt als das reine Faktenwissen. Es beschreibt das Wissen über die Beziehungen und Funktionsweisen einzelner Elemente innerhalb eines übergeordneten Systems (Anderson & Krathwohl, S. 45). Jedoch entspricht seine kognitive Struktur derjenigen von statischem Faktenwissen und wird dementsprechend häufig unter der Bezeichnung deklaratives Wissen subsummiert (z. B. de Jong & Ferguson-Hessler, 1996) oder mit diesem gleichgesetzt (Schneider, 2005). In der vorliegenden Arbeit liegt der Fokus auf der Unterscheidung zwischen deklarativem und prozeduralem Wissen. Die Abgrenzung wird im Folgenden vorgenommen.

Prozedurales Wissen

Während beide bisher beschriebenen Wissensarten das „Wissen was" repräsentieren, bezeichnet das prozedurale Wissen das „Wissen wie". „Wissen wie"

reicht von der simplen Routine bis hin zur Lösung unbekannter Probleme (Hofmeister, 2005). Anderson und Krathwohl (2001) gehen von einer dreischrittigen Unterteilung innerhalb des prozeduralen Wissens aus (vgl. Tabelle 3).

Tab. 3: Das prozedurale Wissen und seine Unterformen

Wissensart und ihre Unterformen	**Beispiele**
Prozedurales Wissen	Wie man etwas tut, Methoden des Vorgehens und Kriterien der Nutzung von Fähigkeiten von Algorithmen, Techniken und Methoden
A. Wissen über fachspezifische Fähigkeiten und Algorithmen	Fähigkeiten, mit Wasserfarbe zu malen, Algorithmen und ganze Zahlen zu dividieren
B. Wissen über fachspezifische Techniken und Methoden	Interview-Techniken, wissenschaftliche Methoden
C. Wissen über Kriterien, wann welche Prozedur angewendet werden soll	Kriterien, die indizieren, wann Newtons zweites Gesetz angewendet werden soll, Kriterien für die Beurteilung der Durchführbarkeit bestimmter Methoden der Unternehmenskostenrechnung

(Quelle: Anderson & Krathwohl, 2001, S. 45–46, frei übersetzt)

Die einfachste Form des prozeduralen Wissens (Stufe A) bezieht sich auf typische Situationen, die es in jeder Profession gibt und die der Handelnde mit standardisierten Fertigkeiten und Verfahren bewältigen kann. Es handelt sich dabei nicht im strengen Sinne um Problemlösen (Dörner, Schaub & Strohschneider, 1999), da die Aufgaben mit bereits gelernten Methoden gelöst werden können. In diesen Bereich gehören zum Beispiel diejenigen Testaufgaben, bei denen ein bekannter Rechenalgorithmus angewendet werden muss (Hofmeister, 2005). Diese Art der prozeduralen Wissensanwendung wird häufig automatisiert und ohne bewussten Abruf deklarativen Wissens vollzogen (Süß, 1996).

Prozeduren auf den Ebenen B und C beziehen sich auf themenspezifische Techniken und Methoden und das Wissen darüber, wann deren Einsatz angemessen ist. Als Beispiel für eine Aufgabe, die prozedurales Wissen bei angehenden zahnmedizinischen Fachangestellten (ZFA) prüft, gibt Hofmeister (2005) die in Abbildung 3 aufgeführte Aufgabe an. Zur Lösung der Beispielaufgabe müssen die angehenden ZFA wissen, welche Techniken im Rahmen der Verbesserung der Mundhygiene angewandt werden können. Darüber hinaus muss sie die situative Befindlichkeit des Patienten berücksichtigen und ihre Maßnahmenentscheidung entsprechend anpassen (Hofmeister, 2005).

Prozedurales Wissen kann sich auf allen kognitiven Anforderungsstufen befinden. Es reicht von der automatisierten Durchführung von Routinen bis hin zur

kritischen Reflexion darüber, wann welche Prozeduren zur Lösung einer Aufgabe angemessen sind (Anderson & Krathwohl, 2001).

Ein Patient zeigt nach der ersten Hygieneeinweisung einen nur leicht verbesserten Mundhygienestatus. Welche Reaktion der ZFA halten Sie für gut, um den Patienten verstärkt zu motivieren?

- ☐ Die ZFA macht dem Patienten Vorwürfe, dass er trotz genauer Einweisung und besseren Wissens immer noch nicht richtig die Zähne putzt.
- ☐ Die ZFA versucht mit dem Patienten herauszubekommen, woran die nach wie vor nicht ausreichende Mundhygiene liegen könnte und kontrolliert noch einmal die Putztechnik des Patienten.
- ☐ Die ZFA macht den Patienten darauf aufmerksam, dass er bei mangelnder Mitarbeit mit einem vorzeitigen Zahnverlust rechnen muss.
- ☐ Die ZFA zeigt am Modell noch einmal die Putztechnik, verzichtet aber auf weitere Motivationsversuche, da eine anhaltende Motivation nicht von außen erzeugt werden kann.

Abb. 3: Beispielaufgabe für prozedurales Wissen auf Stufe B nach Anderson und Krathwohl (2001), Aufgabe 79 aus dem Fachtest für zahnmedizinische Fachangestellte in ULME III

(Quelle: Hofmeister, 2005, S. 14)

2.4 Zusammenführung der zwei Perspektiven auf Wissen

Die Ausführungen zur Unterscheidung der Wissensarten machen deutlich, dass es unterschiedliche Arten der Wissensspeicherung und des Wissensabrufs gibt. Aus den ACT-Theorien (Anderson, 2001) lässt sich ableiten, dass Wissen in eine deklarative und eine prozedurale Komponente zerlegt werden kann. Zudem gibt die Darstellung von prozeduralem Wissen in Form von situationsspezifischen Problemlöseoperatoren erste Hinweise auf die Gestaltung von Aufgaben, die prozedurales Wissen erfassen sollen. Testaufgaben, die entwickelt werden, um Wissensbestände aus dem prozeduralen Gedächtnissystem abzurufen, sollten derart gestaltet sein, dass sie Problemlöseoperatoren vor dem Hintergrund einer konkreten Problemsituation aktivieren. Die Taxonomie von Anderson und Krathwohl (2001) spezifiziert die Überlegungen zur Gestaltung von Aufgaben, die unterschiedliche Wissensarten erfassen sollen. Sie gibt Aufschluss darüber, wie Aufgaben gestaltet werden sollten, um unterschiedliche Wissenssysteme zu aktivieren, indem Beispiele für jede einzelne Wissensart ge-

nannt werden. So kann prozedurales Wissen sowohl über die einfache Abfrage von Routinen als auch über die Abfrage von Wissen über Methoden und deren adäquate Anwendung erfasst werden. Eine Besonderheit der Wissenstaxonomie von Anderson und Krathwohl (2001) ist, dass sie neben der Wissensart auch zwischen verschiedenen kognitiven Anforderungsniveaus unterscheidet, die als wichtige Orientierungshilfe bei der Aufgabenentwicklung herangezogen werden können. Im Speziellen kann die Taxonomie genutzt werden, um Aufgabenschwierigkeiten im Vorfeld der Testung abzuschätzen und Aufgaben entsprechend zusammenzustellen (Hofmeister, 2005). Bothe (2003, S. 9 f.) weist unter in Bezugnahme auf Süß (1996) darauf hin, dass häufig eine unsystematische Vermischung zwischen den Wissensinhalten und der Wissensstruktur vorliegt, wenn zwischen deklarativem und prozeduralen Wissen unterschieden wird. Süß (1996) argumentiert, dass Sachwissen und Handlungswissen sowohl prozedural als auch deklarativ gespeichert sein können. Dieser Sichtweise wird in der Konstruktoperationalisierung in Kapitel 6 Rechnung getragen.

Obwohl die beiden dargestellten Perspektiven unterschiedliche Zielsetzungen verfolgen, lassen sie sich doch sinnvoll integrieren, um die Entwicklung eines Wissenstests für wirtschaftswissenschaftliches Wissen bei Studierenden theoretisch zu fundieren. Insbesondere wird deutlich, dass Wissen nicht als Entität betrachtet werden kann, sondern mindestens zwei Wissensarten existieren, die durch gezielte Formulierung der Anforderungen in Form von Testaufgaben gemessen werden können. Welche Wissensart die andere bedingt, ist nicht abschließend geklärt (Schneider, 2005). Am wahrscheinlichsten ist es, dass ein interaktives Zusammenspiel der Wissensarten für die erfolgreiche Lösung (beruflicher) Anforderungssituationen benötigt wird (Gruber, 1999; Schneider, 2005). Für die gleiche Bedeutsamkeit deklarativen und prozeduralen Wissens sprechen auch die Ergebnisse von McCloy, Campbell und Cudeck (1994), die deklaratives und prozedurales Wissen sowie Motivation als Determinanten der vierfaktoriell gemessenen Performanz von U.S. Soldaten in einem Selektions- und Platzierungsverfahren identifizierten. Keine der beiden Wissensarten war der anderen in dieser Studie in ihrer Vorhersagekraft für die berufliche Performanz überlegen. Diese Befunde sprechen dafür, bei der Erfassung von betriebswirtschaftlichem Wissen von Studierenden beide Wissensarten in Betracht zu ziehen.

Die bisherigen Ausführungen bezogen sich auf die allgemeine Verortung von Wissen im Rahmen beruflicher Kompetenzen und die Unterscheidung der Wissensarten. Im folgenden Kapitel wird der Stand der Forschung zur Erfassung von Wirtschaftswissen von Studierenden dargelegt.

3 Stand der Forschung zur Messung des studentischen Wirtschaftswissens

Die systematische Erfassung und Bewertung von Wirtschaftswissen von Studierenden hat erst im Verlauf der vergangenen zwei Jahrzehnte Eingang in die deutsche Forschungslandschaft gefunden. Erste Arbeiten zur wirtschaftswissenschaftlichen Problemlösefähigkeit wurden von Renkl et al. (1994) durchgeführt und dokumentiert. Die Autoren prägen in diesem Zusammenhang den Begriff des „trägen Wissens", dessen Ausdifferenzierung und Begründung in Abschnitt 3.2 erfolgt. Aktuelle Befunde zum Wirtschaftswissen von Studierenden deutscher Hochschulen stammen zum großen Teil aus dem Projekt „Innovativer Lehr-Lernortverbund in der akademischen Hochschulausbildung" (ILLEV) der Universität Mainz (z. B. Förster, Happ & Zlatkin-Troitschanskaia, 2012; Zlatkin-Troitschanskaia, Förster & Happ, 2012). Ein Schwerpunkt des Projekts lag in dem Vergleich wirtschaftswissenschaftlicher Kompetenzen zwischen Studierenden des alten Diplom-Studienmodells und Studierenden im neuen Bachelor- und Master-Studienmodell. Darüber hinaus liefert die ILLEV-Studie weitere interessante Hinweise auf individuelle und institutionelle Determinanten des wirtschaftswissenschaftlichen Wissens. Diese Studie und ihre Ergebnisse werden in Abschnitt 3.2 beschrieben.

Die Forschung zu wirtschaftswissenschaftlichem Wissen von Studierenden ist vornehmlich im englischsprachigen Ausland deutlich weiter vorangeschritten als in Deutschland. Aus diesem Grund werden im letzten Abschnitt dieses Kapitels zentrale Befunde internationaler Forschungsarbeiten vorgestellt. Dabei wird sowohl ein Überblick über die Determinanten des Studienerfolgs in betriebswirtschaftlichen Fächern als auch über die Determinanten der Leistung in standardisierten Tests gegeben. Darüber hinaus werden die Determinanten des betriebswirtschaftlichen Wissens und erste Befunde zur Struktur betriebswirtschaftlichen Wissens vorgestellt.

3.1 Deutschsprachige Messinstrumente

Die Verfügbarkeit von wissenschaftlich fundierten Testinstrumenten zur Erfassung von Wirtschaftswissen auf Hochschulniveau ist in Deutschland sehr begrenzt (Zlatkin-Troitschanskaia & Kuhn, 2010). Die im Rahmen dieser Forschungsarbeit relevanten Veröffentlichungen werden im Folgenden dargestellt. Das Kapitel beginnt mit der Vorstellung von zwei empirisch bewährten Tests aus dem deutschen Sprachraum. Zuerst wird der in Deutschland vielfach eingesetzte „Wirtschaftskundliche Bildungs-Test" (WBT) (Beck, Krumm & Dubs,

1998) vorgestellt, im Anschluss daran der „Business Administration Knowledge Test" (BAKT) (Bothe, 2003). Der WBT wurde ursprünglich für den Einsatz auf dem Niveau der Sekundarstufe zwei aus dem Englischen übersetzt, gilt jedoch für den Hochschulbereich als weitgehend validiert (Förster et al., 2012). Der BAKT ist derzeit der einzige für wissenschaftliche Zwecke zugängliche deutschsprachige Test, der gezielt und wissenschaftlich entwickelt wurde, um betriebswirtschaftliches Wissen von Studierenden zu erfassen. Ihm kommt in der vorliegenden Arbeit im Rahmen der Konstruktvalidierung eine zentrale Bedeutung zu. Trotz der Nutzung des Tests im Rahmen von Forschungsprojekten (z. B. im Projekt ILLEV, siehe Abschnitt 3.2), liegen zum jetzigen Zeitpunkt nur sehr begrenzte veröffentlichte Informationen zu den Items und ihren Eigenschaften vor. Dementsprechend wird bei der Beschreibung des BAKT in weiten Teilen auf unveröffentlichte Quellen (Bothe, 2003 und Bothe, Wilhelm & Beck, 2005) zurückgegriffen, deren zentrale Ergebnisse im Abschnitt 3.1.2 zusammengetragen werden.

3.1.1 Der Wirtschaftskundliche Bildungs-Test

Der aus dem englischen Sprachraum adaptierte „Wirtschaftskundliche Bildungs-Test" (WBT) (Beck et al., 1996) ist der einzige allgemein zugängliche, objektive Test zur Erfassung beruflicher Fachkundigkeit im kaufmännischen Bereich, der seit mehr als 10 Jahren in der empirischen Bildungsforschung regelmäßig eingesetzt wird (z. B. Förster et al., 2012; Süß, 1999). Der WBT hat das Ziel, ökonomische Grundbildung objektiv, reliabel und valide für unterschiedliche Zielgruppen zu erfassen. Die Wurzeln des WBT liegen im Test of Economic Literacy (TEL) (Soper, 1979), dessen Entwicklung in den Vereinigten Staaten von Amerika durch das gestiegene Interesse an der Erfassung ökonomischer Grundbildung angestoßen wurde. Eine Grundüberlegung der Entwicklung standardisierter Tests war der Wunsch, Bildungsergebnisse national zu vergleichen. Der deutschsprachige WBT baut auf der zweiten überarbeiteten Version des TEL auf (Beck et al. 1998). Bei der Übersetzung des TEL wurde darauf geachtet, so nah wie möglich an den Formulierungen der Originalitems zu bleiben, um eine möglichst hohe internationale Vergleichbarkeit zu gewährleisten. Es wurden lediglich marginale kulturspezifische Veränderungen an den Items vorgenommen.

Der WBT liegt in zwei Parallelformen mit jeweils 46 geschlossenen Multiple-Choice-Items vor. Pro Aufgabe sind vier Antwortalternativen vorgegeben, von denen jeweils eine als richtige Antwort mit einem Punkt bewertet wird. Jede Parallelform enthält Items zu den Inhaltsbereichen: Grundlagen der VWL, Mikroökonomie, Makroökonomie und internationale Beziehungen. Zudem sind die Items entlang der taxonomischen Systematik nach Bloom (1972) (Erinnern, Verstehen, Anwenden, Analyse, Synthese) kategorisiert (vgl. Abschnitt 2.2.2).

Die Taxonomiestufe „Evaluation" wurde nicht berücksichtigt. Die hierarchische Struktur der kognitiven Anforderungen hält jedoch einer empirischen Prüfung nicht stand (z. B. Müller et al., 2007). Wie bereits kritisch in Abschnitt 2.3.2 vermerkt, entsprechen die empirischen Aufgabenschwierigkeiten im WBT nicht den durch die Kategorisierung vorhergesagten Aufgabenschwierigkeiten (Witt, 2006). Diese Problematik wird im Verlauf der Arbeit und insbesondere in Abschnitt 4.3 erneut aufgegriffen und diskutiert.

Die Autoren schreiben dem WBT einen umfassenden Einsatzbereich zu (Beck et al., 1998, S. 8): Schulische Allokationsdiagnostik, schulische Evaluation, Berufsberatung, betriebliche Personalauslese und Personalentwicklung. Zudem wird die hochschulische Eingangs- und Erfolgsmessung für wirtschaftswissenschaftliche Haupt- und Nebenfachstudiengänge als Anwendungsfeld genannt. Darüber hinaus soll der Test mehr als Faktenwissen erfassen. Testteilnehmenden werden die Kenntnis und Anwendung von ökonomischen Konzepten wie Angebot und Nachfrage, Präferenzen, Kosten und Nutzen und Wachstum abverlangt (Bothe et al., 2005).

Für die Auswertung auf der Ebene des Individuums werden die Rohpunkte der Einzelpersonen erfasst. Für jede korrekt gelöste Aufgabe wird ein Punkt vergeben. Punktabzug für falsch gelöste Aufgaben ist nicht vorgesehen. Da die Parallelformen nicht den gleichen Schwierigkeitsgrad aufweisen, stellen die Autoren eine Tabelle zum Umrechnen der Rohpunktäquivalente zwischen beiden Testformen zur Verfügung (Beck et al., 1998). Außerdem liegen dem Testmaterial Tabellen für die Angabe von Prozenträngen bei. Die Auswertung der Testergebnisse mittels Prozenträngen lässt einen Rückschluss über die relative Position des Probanden im Verhältnis zu einer Vergleichsgruppe (Normstichprobe) zu. Die Normstichproben beziehen sich auf Probanden aus verschiedenen kaufmännischen Bildungsgängen, Fachwirtschaftsgymnasien, allgemeinbildenden Gymnasien, Realschulen und Studierenden der Wirtschaftswissenschaften aus Deutschland. Die Normstichprobe der Studierenden ist mit (N = ca. 180) allerdings zu klein und für den Vergleich mit aktuellen Testdaten veraltet (Amelang & Schmidt-Atzert, 2006), sodass eine Angabe von Prozenträngen auf dieser Grundlage kritisch zu bewerten ist.

Der WBT zeichnet sich durch einen hohen Grad an Standardisierung aus. Das Handbuch zum WBT enthält präzise Anweisungen bezüglich der Testinstruktion, -durchführung und dessen Auswertung. Die Testzeit orientiert sich mit 45 Minuten an einer regulären Schulunterrichtsstunde und darf im Regelfall von den Testteilnehmenden nicht überschritten werden.

Für die vorliegende Forschungsarbeit stellt sich die Frage, inwiefern der WBT geeignet ist, um als Diagnostikum von betriebswirtschaftlichem Wissen an

Hochschulen eingesetzt zu werden. Für die Messung betriebswirtschaftlichen Wissens an der Hochschule sind unter anderen zwei Aspekten des WBTs problematisch. Zum einen sind die Items veraltet und beziehen sich auf Begrifflichkeiten, die heute nicht mehr gelehrt werden (Bank & Retzmann, 2011). Zum anderen bescheinigt Förster et al. (2012) dem WBT zwar eine für den Einsatz an Hochschulen ausreichende Validität, jedoch klammert die starke Fokussierung auf volkswirtschaftliche Inhalte wichtige betriebswirtschaftliche Bereiche, wie den Bereich des Rechnungswesens, aus (Schumann et al., 2010). Es ist umstritten, inwieweit das eingeschränkte Themenfeld des WBT es erlaubt, im umfassenden Sinne von der Erfassung ökonomischer Bildung oder Fachkundigkeit zu sprechen (Tramm & Seeber, 2006). Aspekte der Betriebswirtschaftslehre (BWL) als auch des privaten Haushaltens werden in dem Test vernachlässigt (Tramm & Seeber, 2006). Die ausgeprägte Fokussierung des WBT auf volkswirtschaftliche Inhalte begründet zum einen die Notwendigkeit, einen weiteren deutschsprachigen Test zur Erfassung des Wirtschaftswissens im Bereich der BWL zu entwickeln und schließt den Test für die Verwendung im vorliegenden Forschungsvorhaben zugleich aus. Alternativ wird im folgenden Kapitel ein weiterer deutschsprachiger Test vorgestellt, der speziell für die Erfassung betriebswirtschaftlichen Wissens an Universitäten entwickelt wurde.

3.1.2 Der Business Administration Knowledge Test

Der Ursprung der Entwicklung des BAKT liegt in einem Forschungsprojekt des Instituts für Mittelstandsforschung der Universität Mannheim zur Erfassung von Wirtschaftswissen für die Personalauswahl in kleinen und mittleren Unternehmen (KMU) (Größler, Wilhelm, Wittman, & Milling, 2002). Das Projekt hatte zum Ziel, der Nicht-Existenz von effektiven standardisierten Wissenstests für betriebswirtschaftliche Inhalte im deutschen Sprachraum zu begegnen (Größler et al., 2002). In diesem Rahmen wurde ein Test entwickelt, der betriebswirtschaftliches Wissen unter besonderer Berücksichtigung der Anforderungen an Absolventen betriebswirtschaftlicher Studiengänge in KMU erfassen soll. In den ersten Veröffentlichungen zum Test wird dieser als Wirtschaftswissenstest (WWT) bezeichnet (Bothe, 2003; Größler et al., 2002). In Anlehnung an ein englischsprachiges Manuskript von Bothe et al. (2005) wird der Test in allen weiteren Forschungsarbeiten als Business Administration Knowledge Test (BAKT) bezeichnet (Förster et al., 2012). Itementwicklung und Ergebnisse erster Validierungsstudien wurden sowohl in (Bothe, 2003) als auch in Bothe et al. (2005) ausführlich beschrieben. Die folgende Beschreibung des Tests und der Validierungsergebnisse beziehen sich auf die Ausführungen in Bothe (2003), sofern nicht anders gekennzeichnet.

Die Testentwicklung des BAKTs folgte einem curricular verankerten Domänen-Verständnis. Im ersten Schritt der Testentwicklung erfolgte die Klassifikation

von betriebswirtschaftlichem Wissen anhand von Lehrbüchern sowie universitären und schulischen Curricula. Auf diese Weise wurden neun Hauptdomänen der Betriebswirtschaft identifiziert: General Management, Production & Operations, Finance, Marketing & Sales, Financial Accounting & Reporting, Cost Accounting, Strategy, Tax und Human Resources. Diese neun inhaltlichen Hauptdomänen wurden jeweils in weitere Subdomänen unterteilt (vgl. Abbildung 4).

Bilanzierung	Bilanz	Erfolgsrechnung	Rechnungslegung im Konzern	Bilanzauffassung		
Kostenrechnung	Betriebsabrechnung	Kostenträgerrechnung	Deckungsbeitragsrechnung	Plankostenrechnung		
Finanzierung	Investitionsplanung	Unternehmensbewertung	Grundlagen Finanzplanung	Außenfinanzierung	Innenfinanzierung	Portfolio Management
Absatz	Marktforschung	Produktpolitik	Preispolitik	Kommunikationspolitik	Distributionspolitik	
Oparation Management	Produktions- und Kostentheorie	Produktionsplanung	Materialbedarfsplanung	Integration der Produktionsplanung und Planungssteuerung	Logistik	
Allgemeines Wirtschaftswissen	Rechtsformen der Unternehmen	Wirtschaftspolitik	Themen aus dem Gebiet Recht	Börse	Marktformen	Außenwirtschaft/EU
Strategisches Management	Normative Entscheidungstheorie	Deskriptive Entscheidungstheorie	Organisation	Planung	Strategische Unternehmensführung	
Steuern	Einkommenssteuer	Körperschafts- und Gewerbesteuer				
Human Resources	Führung	Mitbestimmung/Ziele	Personalplanung/-beschaffung	Personalentwicklung/-entgelt		

Abb. 4: Domänen und Subdomänen des Wirtschaftswissenstests (BAKT-L) und seiner Kurzform (grau unterlegt)

(Quelle: Bothe, 2003, S. 94)

Die Autoren geben an, dass Experten und Expertinnen der Betriebswirtschaft die Unterteilung als weitgehend sinnvoll bewertet haben, weisen aber darauf hin, dass eine Klassifikation der relevanten betriebswirtschaftlichen Wissensdomänen im strengen Sinne nur anhand eines definierten (Berufs-)Ziels vorge-

nommen werden kann (Größler et al., 2002). Damit weisen die Autoren auf die ausgeprägte Heterogenität der beruflichen Anforderungen von Absolventen betriebswirtschaftlicher Studiengänge hin, die in Abschnitt 6.2 dieser Arbeit näher erläutert wird.

Das in Abbildung 4 dargestellte Raster wurde in allen weiteren Bearbeitungsschritten der Testentwicklung des BAKT und dessen Auswertung als Grundlage herangezogen. In Anlehnung an die zuvor definierten Domänen wurde eine Aufgabendatenbank mit 284 Aufgaben erstellt. Dabei wurde darauf geachtet, jede Subdomäne mit mindestens 5 Aufgaben zu repräsentieren. Mittels Expertendelphi wurde pro Subdimension ein Item ausgewählt und somit eine Kurzversion des BAKT mit 22 Items erstellt (BAKT-S) (Größler et al., 2002) (vgl. grau unterlegte Kästchen in Abbildung 4). Anschließend wurde der Test in Kombination mit unterschiedlichen Begleitfragbögen online administriert. 784 Teilnehmerdaten wurden für weitere Berechnungen herangezogen. Davon erhielten 235 Teilnehmer Zusatzfragen aus dem „Fragebogen zur Diagnose unternehmerischer Potenziale“ (F-DUP-K) (Müller, 2003), 449 Teilnehmer eine Kurzform des WBT (Beck et al., 1998) und wiederum eine Teilstichprobe davon einen biografischen Fragebogen zu Führungsverhalten. Das Durchschnittsalter der Testteilnehmer betrug 28.6 Jahre und 35.5 % der Testteilnehmer waren weiblich. Die Testteilnehmer kamen aus unterschiedlichen Berufsgruppen, wobei die größten beiden Gruppen von Studierenden und Angestellten gebildet wurden.

Nach dem Ausschluss von drei Items aufgrund zu niedriger Schwierigkeit oder niedriger Diskriminanz wurde mittels konfirmatorischer Faktoranalyse eine Prüfung unterschiedlicher dimensionaler Modelle vorgenommen (Bothe, 2003). Es wurden folgende vier Modelle gegeneinander getestet: (1) Ein Generalfaktormodell zur Messung einer übergeordneten latenten betriebswirtschaftlichen Wissensdimension, (2) ein higher-order Modell, in dem die zweite Ebene die Korrelationen zwischen den neun Inhaltsdomänen der BWL beschreibt, (3) ein korreliertes Gruppenfaktormodell, in dem jede Inhaltsdomäne einen Faktor darstellt und die Faktoren korrelieren dürfen und (4) ein hierarchisches Modell im Sinne von Modell 2 mit zusätzlicher Erklärung der Residualkovarianzen zwischen Items einer Inhaltsdomäne. Der Vergleich der vier Modelle identifizierte das eindimensionale Modell als dasjenige mit der besten Passung. Das bedeutet, dass Wissen in einer inhaltlichen Domäne der BWL in der Regel mit Wissen in allen anderen Domänen einhergeht und vice versa. Entsprechend dieses Ergebnisses wurden alle weiteren Analysen des BAKT auf der Grundlage eines eindimensionalen Modells durchgeführt. Erste Ansätze der Testvalidierung durch korrelative Beziehungen zu theoretisch als relevant identifizierten Konstrukten zeigten sich als weitgehend hypothesenkonform. Die Skala des F-

DUP-K zu „Unsicherheitstoleranz" korrelierte signifikant positiv mit den Ergebnissen des BAKT-S, ebenso wie „Risikoneigung". Eine signifikante Korrelation wurde im Zusammenhang mit der Kurzversion des WBT festgestellt ($r = .87$). Der Vergleich zwischen einer gemeinsamen eindimensionalen Modellierung von WBT und BAKT und einer zweidimensionalen Modellierung mit interkorrelierten Faktoren fiel erwartungskonform zu Gunsten des zweidimensionalen Modells aus. Die Autoren begründen die hohe Interkorrelation der Faktoren zum einen durch die Überlappung der Testinhalte im Bereich des allgemeinen Wirtschaftswissens und zum anderen durch die enge Verknüpfung betriebswirtschaftlicher und volkswirtschaftlicher Inhalte, sowohl in der universitären Lehre als auch im öffentlichen Leben (z. B. in den Medien). Die Reliabilität des BAKT-S ist mit einem Cronbachs Alpha von .63 bei einer Länge von 19 Items jedoch als kritisch einzustufen (Cronbach, 1951).

In einer zweiten Studie wurde die Dimensionalität des Tests anhand einer 42-Item-Version erneut geprüft. Zusätzlich wurde das typische intellektuelle Engagement (TIE) als Außenkriterium für die Validierung aufgenommen. Diese zweite Studie wurde als papierbasierte Studie mit 405 Studierenden der Betriebswirtschaft, Volkswirtschaft und Wirtschaftspädagogik durchgeführt. Das Durchschnittsalter der Testteilnehmer betrug 22.1 Jahre, die Studierenden hatten mehrheitlich zwei bis drei Studiensemester abgeschlossen. Die Ergebnisse bestätigten die einfaktorielle Struktur der BAKT-S sowie die hohe Interkorrelation mit den Ergebnissen des WBT. Auch der BAKT-L wird am besten durch ein eindimensionales Modell beschrieben. Lediglich Accounting und Cost-Accounting wiesen einen eigenen Faktor aus, der jedoch hoch mit den anderen Items des BAKT-L korrelierte. Der Autor schlussfolgert, dass der BAKT-L, ebenso wie der BAKT-S, am besten durch ein eindimensionales Modell erklärt wird. Alle drei Skalen des typischen intellektuellen Engagements (TIE) waren positiv mit den Testwerten des BAKT korreliert, was dahingehend interpretiert wurde, dass TIE eine große Rolle bei der Wissensakquisition spielt. Mit einer Reliabilität von Cronbachs Alpha = .78 konnte die Messgenauigkeit der 29 Items des BAKT-L deutlich verbessert werden (Cronbach, 1951).

Faktoranalytische Untersuchungen des BAKT (Bothe et al., 2005) legen vorübergehend eine eindimensionale Struktur des betriebswirtschaftlichen Grundlagenwissens nahe. Es kann jedoch nicht ausgeschlossen werden, dass fachspezifische Subdimensionen existieren, die sich aufgrund der begrenzten Itemzahl pro Fach nicht abbilden lassen. Vertiefende Analysen weisen darauf hin, dass insbesondere der Bereich Rechnungswesen möglicherweise eine eigene Dimension abbildet (Bothe et al., 2005).

Die Studien zum BAKT sind für die vorliegende Arbeit unter zweierlei Gesichtspunkten wichtig. Zum einen geben sie erste Anhaltspunkte für die Aufteilung

betriebswirtschaftlicher Studieninhalte in Hauptdomänen und Subdomänen. Zum anderen bilden sie eine wichtige Grundlage hypothetischer Überlegungen zur Testdimensionalität. Folgt man den bisherigen empirischen Ergebnissen zur Dimensionalität des BAKT, so ist davon auszugehen, dass Wissen über Betriebswirtschaft, das die Breite der Inhalte des Bachelorstudiums abdeckt, durch ein eindimensionales Modell abgebildet werden kann. Es handelt sich um ein heterogenes, eindimensionales Konstrukt, was zu Schwierigkeiten bei der Bestimmung der Reliabilität durch Cronbachs Alpha führen kann, ebenso wie bei der Auswertung durch Item-Response-theoretische Methoden (Bothe, 2003) (siehe Diskussion dieser Problematik in Abschnitt 10.2). Insbesondere letzteres ist ein Kritikpunkt am BAKT, der unter der Annahme der klassischen Testtheorie entwickelt wurde und dementsprechend nicht davon ausgegangen werden kann, dass die Items anhand probabilistischer Modelle ausgewertet werden können. Der Versuch einer probabilistischen Auswertung der Items des BAKT bei Bothe (2003) ergab, dass ein Dreiparametermodell (vgl. 7.1.2) die vergleichsweise beste Passung aufwies, einzelne Items aber signifikant vom Modell abwichen und die Stichprobe zu klein war um weitere Modelltests zu rechnen. Die probabilitische Auswertung wurde nicht weiter verfolgt. Eine probabilistische Auswertung wird jedoch als State of the Art der empirischen Bildungsforschung angesehen (Rost, 2004) und in der vorliegenden Arbeit angestrebt. Probabilistische Auswertungsmodelle bieten den großen Vorteil, dass sie eine von der Stichprobe unabhängige Betrachtung der Aufgabenschwierigkeiten ermöglichen und für jedes Item mehr Informationen zur Verfügung stehen, die für die Itemselektion genutzt werden können als unter Betrachtung klassischer Itemkennwerte (vgl. Abschnitt 7.1.1). Probabilitische Testmodelle und deren Vorteile gegenüber klassischen Modellen werden im Abschnitt 7.1 dieser Arbeit beschrieben.

Kritisch zum BAKT anzumerken ist, dass er vor dem Hintergrund des alten Diplomstudienmodells entwickelt wurde (Größler et al., 2002). Es gilt zu prüfen, ob dadurch seine Übertragbarkeit auf die heutigen Bachelor- und Masterstudiengänge eingeschränkt.

Trotz der sehr begrenzten Verfügbarkeit von aktuellen und wissenschaftlich fundierten Testinstrumenten zur Erfassung betriebswirtschaftlichen Wissens werden zentrale Ergebnisse der nationalen Forschung zu Wirtschaftswissen von Studierenden im Folgenden vorgestellt.

3.2 Nationale Forschungsergebnisse

Renkl, Gruber, Mandl und Hinkhofer (1994) veröffentlichten eine der ersten deutschsprachigen empirischen Arbeiten zum betriebswirtschaftlichen Wissen

und Können von Hochschulstudierenden. Für ihre Untersuchung nutzten sie keinen der vorgestellten standardisierten Tests, sondern eine Unternehmenssimulation von Preiss und Klauser (1992). Dabei griffen sie die aufkeimende Kritik, es werde an deutschen Bildungsinstituten zu viel „träges Wissen" produziert, auf und führten einen Leistungsvergleich im betriebswirtschaftlichen Problemlösen zwischen Münchener Studierenden der Wirtschaftswissenschaften und Studierenden der Psychologie und Pädagogik durch. Unter „trägem Wissen" wird theoretisches Wissen verstanden, das für die Lösung komplexer, realitätsnaher Probleme nicht genutzt werden kann. Renkl et al. (1994) kritisierten in diesem Zusammenhang die Kluft zwischen Wissen und Handeln an deutschen Hochschulen. Ganz ähnlich dem heutigen, durch den Bologna-Prozess gestärkten (Bologna Declaration, 1999), Kompetenzverständnis gingen Renkl et al. (1994) davon aus, dass die universitäre Bildung den Studierenden ermöglichen sollte, nützliches Wissen zu erwerben, welches auch außerhalb des instruktionalen Kontexts der Universität gebraucht wird und die Studierenden zum Lösen komplexer Probleme in Beruf und Alltag befähigt. Als Maß für diese Anwendung theoretischen (deklarativen) Wissens wurde das betriebswirtschaftliche Planspiel „Jeansfabrik" (Preiss & Klauser, 1992) eingesetzt. Ziel des Planspiels ist es, Jeanshosen auf einem Duopolmarkt mit Gewinn abzusetzen. Dabei müssen die Teilnehmer in jedem Planspielmonat Entscheidungen über Verkaufspreis und Produktionsmenge treffen, um Gewinne zu erzielen.

Das erstaunliche Ergebnis dieser Studie war, dass die Studierenden der Wirtschaftswissenschaften deutlich weniger Gewinne erzielten als die Vergleichsgruppe der Studierenden der Psychologie und Pädagogik. Auch die mentalen Modelle der eigentlichen Experten (Studierende der Wirtschaftswissenschaften) unterschieden sich erst nach der Beendigung des Planspiels in ihrer Qualität von denen der Wirtschaftsnovizen (Studierende der Psychologie und Pädagogik). Renkl et al. (1994) schlussfolgerten, dass das im wirtschaftswissenschaftlichen Studium erworbene Wissen den Studierenden nicht, wie erwartet, die Steuerung eines komplexen ökonomischen Systems erleichtert, sondern im Gegenteil der erfolgreichen Umsetzung des Gelernten im Wege steht. Eine weitere Schlussfolgerung der Autoren bestand darin, dass Studierende mehr an komplexen Problemstellungen lernen sollten. In Angesicht des heutigen Kompetenzverständnisses muss einschränkend hinzugefügt werden, dass die Studie von Renkl et al. (1994) sich weniger mit Kompetenzen oder Wissen von Studierenden auseinandergesetzt hat, sondern mit komplexer Problemlösefähigkeit. Trotzdem unterstreicht die Forschungsarbeit von Renkl et al. (1994) die Relevanz der Auseinandersetzung mit der Entwicklung unterschiedlicher Arten wirtschaftswissenschaftlichen Wissens an Hochschulen.

Ein Forschungsprojekt in Deutschland, in dem sich jüngst mit der Messung von volkswirtschaftlichem und betriebswirtschaftlichem Wissen bei Studierenden der Wirtschaftswissenschaften und der Betriebswirtschaft befasst wurde, ist das Projekt innovativer Lehr-Lernortverbund (ILLEV) der Universität Mainz (z. B. Förster et al., 2012; Zlatkin-Troitschanskaia et al., 2012). Im Projekt ILLEV ging es um die Klärung der Frage, inwieweit die akademische Hochschulbildung die Entwicklung der professionellen Handlungskompetenz von Studierenden beeinflusst und welche Rolle individuelle Personencharakteristika dabei spielen. In diesem Zusammenhang wurden unter anderem Testleistungen von Studierenden auslaufender Diplomstudiengänge mit den Testleistungen Studierender innerhalb des Bachelor-Master-Modells verglichen. Die Befunde dieses Projekts werden im Folgenden vorgestellt. Alle das Projekt betreffende Aussagen beziehen sich, sofern nicht anders gekennzeichnet, auf eine Publikation zum Projekt von Zlatkin-Troitschanskaia et al. (2012).

Um Aussagen über die Entwicklung der wirtschaftswissenschaftlichen Fachkompetenz über die Studienzeit treffen zu können, wurde im ILLEV Projekt ein als Längsschnitt angelegtes Forschungsdesign mit vier Erhebungszeitpunkten umgesetzt. Als Erhebungsinstrument wurden der Wirtschaftskundliche Bildungstest (WBT) von Beck et al. (1998) für volkswirtschaftliches Wissen und der Business Administration Knowledge Test (BAKT) von Bothe (2003) für betriebswirtschaftliches Wissen eingesetzt. Beide Testinstrumente wurden im vorangegangenen Abschnitt 3.1 detailliert vorgestellt. Darüber hinaus wurden im Projekt ILLEV die motivationalen Orientierungen der Studierenden in Form von extrinsischer berufsbezogener Motivation sowie von intrinsischer Motivation nach Schiefele, Krapp, Wild und Winteler (1993) erfasst. Zudem wurden epistemologische Überzeugungen zur „Objektivität des Wissens" durch einen Fragebogen von Schiefele und Moschner (1997) (zit. nach Zlatkin-Troitschanskaia et al., 2012) erfasst. Als zu kontrollierende Variablen zählten zum einen die kognitiven Voraussetzungen und zum anderen soziodemografische Aspekte der Studierenden. Kognitive Voraussetzungen wurden in ILLEV über die Note der Hochschulzugangsberechtigung und über die zwei Skalen „Analogien" und „Zahlenreihen" aus dem Intelligenzstrukturtest 2000 R (IST 2000 R) (Liepmann, Beauducel, Brocke & Amthauer, 2007) operationalisiert. Im Rahmen der soziodemografischen Fragen wurden Alter, Geschlecht und eine eventuell absolvierte Berufsausbildung erfragt.

Der Vergleich zwischen Studierenden nach altem und neuem Studienmodell bezüglich der kognitiven Leistungsvoraussetzungen und intrinsischer Motivation zeigte keine signifikanten Unterschiede. Die extrinsische Motivation war bei Studierenden des neuen Studienmodells signifikant stärker ausgeprägt. Das Studienmodell hatte aber unter Kontrolle der Semesterzahl keinen Einfluss auf

Leistungen im volkswirtschaftlichen Fachwissenstest (WBT). Insgesamt konnten strukturelle Merkmale, wie Anzahl der besuchten Lehrveranstaltungen, zwar als positive Einflussfaktoren auf die Testleistung identifiziert werden, im Vergleich zu dem Einfluss individueller Faktoren ist die Bedeutung struktureller Merkmale jedoch (überraschend) gering. So lässt sich die Varianz in der Ausprägung des wirtschaftswissenschaftlichen Fachwissens hauptsächlich auf die Faktoren „Deutsch als Muttersprache", Geschlecht, Note der Hochschulzugangsberechtigung und Intelligenz erklären. Konkret fallen die Testwerte für weibliche Teilnehmer, Teilnehmer mit schwachen Abiturnoten und Personen mit fremdsprachlichem Hintergrund schlechter aus als für die übrigen Personen. In einer ebenfalls aus dem ILLEV-Projekt entstandenen Studie zur Validierung des WBT (Förster et al., 2012) wurde zudem der Einfluss einer absolvierten Berufsausbildung vor dem Studium in die Auswertungen aufgenommen. Es zeigte sich, dass eine absolvierte Berufsausbildung, auch unter Kontrolle weiterer Faktoren, stark positiv mit den Testwerten des WBTs zusammenhängt. Die Ergebnisse werden in Tabelle 4 in Form einer Regressionsanalyse zusammenfassend dargestellt.

Tab. 4: Lineare Regression auf die Summe der Rohpunkte des WBTs

Modell **R^2 = 0.257** **Korr. R^2 = 0.253**	**Regressions-koeffiz. *ß****	**Standard-fehler (SE)**	***t***	***p***
(Konstante)	5.735	.590	9.719	.000
weibliches Geschlecht	–.865	.176	–4.918	.000
andere als deutsche Muttersprache	–.703	.254	–2.770	.006
absolvierte Berufsausbildung	.520	.201	2.580	.011
Analogiescore	.292	.036	8.209	.000
Zahlenreihenscore	.081	.023	3.562	.001
Anzahl besuchter VWL-Veranstaltungen	.673	.070	9.571	.000
***** nicht-standardisiert				

(Quelle: Förster et al., 2012, S. 11)

Für die Testergebnisse des BAKT wurden entsprechende Berechnungen noch nicht veröffentlicht. Jedoch liegen korrelative Analysen zwischen dem BAKT und dem WBT vor (Förster et al., 2012). Mittels einer einfachen bivariaten Pearson-Korrelation zwischen der Summe der Rohpunkte des BAKTs und des WBTs wurde ein Zusammenhang von $r = .34$ ermittelt. Unter Kontrolle der Intelligenzmaße „Analogien" und „Zahlenreihen" sinkt dieser Zusammenhang

leicht auf $r = .27$. Die Ergebnisse unterstützen die Hypothese der Autoren und Autorinnen, dass die beiden Tests zwar miteinander korrelieren, aber doch deutlich voneinander abgrenzbar sind. Intelligenz wird als Prädiktor für den Aufbau fachspezifischen Wissens gesehen, der WBT bildet jedoch nachweislich der Regressionsanalyse in Tabelle 4 deutlich mehr als reine Intelligenzleistungen ab (Förster et al., 2012).

Der Vergleich zwischen den Studiengängen Wirtschaftspädagogik (Wipäd) und Wirtschaftswissenschaften bezüglich der Testleistung fiel zum ersten Erhebungszeitpunkt zu Gunsten der Studierenden mit Lehramtsperspektive (Wirtschaftspädagogik) aus. Die Testwerte für Bachelorstudierende der Wirtschaftspädagogik waren signifikant höher als für Bachelorstudierende der Wirtschaftswissenschaften (Förster & Zlatkin-Troitschanskaia, 2010). Dieser Effekt bestätigte sich nicht für Studierende in den Diplomstudiengängen und wurde unter Kontrolle relevanter individueller und institutioneller Einflussfaktoren nicht mehr signifikant[1] (Zlatkin-Troitschanskaia, 2013). Stattdessen zeigte sich sowohl für betriebswirtschaftliches als auch für volkswirtschaftliches Wissen ein positiver Effekt zu Gunsten von Studierenden mit einer absolvierten Berufsausbildung und zu Ungunsten weiblicher Testteilnehmer (Förster & Zlatkin-Troitschanskaia, 2010). Erste Analysen der Entwicklung betriebswirtschaftlichen Wissens, über die vier Messzeitpunkte hinweg, sprechen für einen umgekehrt U-förmigen Verlauf der Wissensentwicklung, in der das Wissen in den ersten Semestern ansteigt und am Ende des Studiums leicht abfällt (Happ, 2013). Die Analysen des Längsschnitts befinden sich allerdings noch in der explorativen Phase und müssen weiter ausgebaut werden, bevor verbindliche Aussagen über den Verlauf der Wissensentwicklung über den Studienverlauf getroffen werden können (Happ, 2013).

Über die Struktur wirtschaftswissenschaftlichen Wissens ist auf Grundlage deutscher Studien noch wenig bekannt. Aufgrund mittlerer Korrelationen zwischen der Summe der Rohwerte des BAKT und des WBT wird davon ausgegangen, dass betriebswirtschaftliches und volkswirtschaftliches Wissen zwei hinreichend unterscheidbare Wissensdimensionen abbilden (Förster & Zlatkin-Troitschanskaia, 2010). Hinsichtlich der Binnenstruktur des betriebswirtschaftlichen Wissens weisen die Ergebnisse auf ein eindimensionales Konstrukt hin (Bothe, 2003; Bothe et al. 2005).

Die nationalen veröffentlichten Forschungsbefunde beziehen sich zum großen Teil auf Ergebnisse, die mit Leistungstests generiert wurden, die einen starken

1 Eine Differenzierung zwischen Studierenden der Wirtschaftpädagogik mit betriebswirtschaftlichem Zweitfach (Wipäd I) und Studierenden der Wirtschaftspädagogik mit nicht-betriebswirtschaftlichem Zweitfach (Wipäd II) wurde nicht vorgenommen.

volkswirtschaftlichen Bezug haben. Diese Tatsache ist vor allem dem Mangel an deutschsprachigen Instrumenten im betriebswirtschaftlichen Bereich geschuldet. Zum Stand der nationalen Forschung ist anzumerken, dass für den Zeitraum von vier Jahren (2011–2015) 23 Projekte zum Thema „Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor" im Rahmen des gleichnamigen Förderprogrammes des Bundesministeriums für Bildung und Forschung (BMBF) gefördert werden. Zwei Projekte mit dem Schwerpunkt im Bereich der Erfassung wirtschaftswissenschaftlicher Fähigkeiten werden im Folgenden skizziert.

3.3 Laufende nationale Forschungsprojekte

Die vom BMBF geförderten Projekte decken unterschiedliche Fachbereiche ab. Zwei davon befassen sich mit der Erfassung wirtschaftswissenschaftlichen Fachwissens. Die Koordination der Projekte obliegt der Humboldt-Universität zu Berlin, in Kooperation mit der Johannes Gutenberg-Universität Mainz, die eine Webseite zu allen Projekten administriert (Gutenberg-Universität Mainz, 2013). Auf diese Webseite beziehen sich alle weiteren Angaben zu den Projekten. Das Projekt „Modellierung und Erfassung fachwissenschaftlicher und fachdidaktischer Kompetenzen im wirtschaftspädagogischen Studium" (KoMeWP) unter der Leitung von Prof. Dr. Jürgen Seifried hat zum Ziel, Tests zur fachwissenschaftlichen und fachdidaktischen Kompetenz angehender Lehrkräfte an kaufmännischen Schulen für die Domäne des Rechnungswesens zu entwickeln. Die Kompetenzmodellierung soll im Hinblick auf folgende drei zentrale Kompetenzfacetten stattfinden: (1) Unterstützung bei der Bearbeitung von Aufgaben, (2) Anbieten verschiedener Lösungswege und (3) Antizipation von Schülerfehlern und Lernschwierigkeiten. Für das Jahr 2013 ist eine Erhebung mit insgesamt ca. 1.000 Masterstudierenden im Bereich Wirtschaftspädagogik an 26 deutschen Standorten und einem österreichischen Standort geplant. Ein weiteres Projekt zur „Modellierung und Messung wirtschaftswissenschaftlicher Fachkompetenz bei Studierenden bzw. Hochschulabsolventen mittels Adaptation und Weiterentwicklung vorliegender amerikanischer und spanischsprachiger Messinstrumente" (WiWiKom) unter der Leitung von Prof. Dr. Olga Zlatkin-Troitschanskaia hat das Ziel, ein Kompetenzmodell und ein valides Messinstrument zur Erfassung wirtschaftswissenschaftlicher Fachkompetenz von Studierenden und Hochschulabsolventen im deutschsprachigen Raum durch Adaptation internationaler Tests zu entwickeln. Als Grundlage für die Testentwicklung werden der spanischsprachige „EGEL[2]" aus Mexiko für betriebswirtschaftliches Wissen und der amerikanische Test „Test of Unterstan-

2 Exámenes Generales de Egreso dela Licenciatura

ding in College Economics (TUCE)" (Saunders, 1991) für volkswirtschaftliches Wissen möglichst originalgetreu adaptiert. Da die Projekte derzeit noch nicht abgeschlossen sind, kann in dieser Arbeit nicht oder nur sehr eingeschränkt auf Konzepte und Ergebnisse dieser Forschungsvorhaben zurückgegriffen werden. Sie untermauern jedoch das aktuelle Interesse der deutschen Bildungsforschung und Bildungspolitik an der Messung von Lernergebnissen im Hochschulsektor.

Internationale Befunde zu betriebswirtschaftlichem Wissen sind deutlich umfassender publiziert als deutschsprachige und liefern über den deutschsprachigen Forschungsstand hinaus wichtige Hinweise auf die Determinanten und die Struktur betriebswirtschaftlichen Wissens an Hochschulen. Entsprechend werden im folgenden Kapitel internationale Messinstrumente und Forschungsergebnisse fokussiert.

3.4 Internationale Messinstrumente

Im englischsprachigen Bereich, speziell in den Vereinigten Staaten von Amerika, ist die Verwendung von standardisierten Tests deutlich etablierter als in Deutschland. Ein prominentes Beispiel für die weitläufige Akzeptanz von standardisierten Testverfahren ist der Scholastic Aptitude Test (SAT) (Donlon & Angoff, 1971). Der SAT misst verbale und mathematische sowie fachspezifische Fähigkeiten und fungiert an einer Vielzahl von Universitäten als Zulassungstest (Cleary, 1968). Der größte Anbieter standardisierter Tests ist der Educational Testing Service (ETS), der laut eigenen Angaben jährlich mehr als 50 Millionen Tests in 180 Ländern administriert und auswertet (vgl. www.ets.org, 14. März 2013). Ein weiterer Test des ETS ist der Major Field (Achievement) Test in Business (MFAT-B), der im Rahmen von Akkreditierungsbemühungen zahlreicher universitärer Studienprogramme im betriebswirtschaftlichen Bereich genutzt wird. Die Items des MFAT-B sind, mit Ausnahme einiger Beispielitems, nicht veröffentlicht. Jedoch liegen gut publizierte Forschungsbefunde zur Dimensionalität des Tests und zu den Zusammenhängen mit verschiedenen Außenkriterien vor. Der MFAT-B wird in Abschnitt 3.4.1 vorgestellt. Im darauffolgenden Abschnitt 3.4.2 wird ein Projekt der Organisation for Economic Co-operation and Development (OECD) vorgestellt, das in erster Linie praktische Hinweise zur Erfassung von Wissen an Hochschulen liefert.

3.4.1 Der Major Field Achievement Test in Business

Der Major Field Achievement Test in Business (MFAT-B) wurde 1990 das erste Mal eingesetzt. Über seine Entstehung ist wenig bekannt, da er nicht primär zu Forschungszwecken entwickelt wurde, sondern im Rahmen von Akkreditierun-

gen genutzt und kommerziell vertrieben wird (vgl. www.ets.org). Der ETS (2011) gibt an, dass erfahrene Universitätsangestellte unterschiedlicher Fachrichtungen an der Entwicklung der Aufgaben beteiligt waren und der Test alle fünf Jahre unter Mithilfe dieser nicht näher beschriebenen Expertengruppe überarbeitet wird. Dabei wird darauf geachtet, möglichst allgemeingültige Aufgaben in den Test zu integrieren Beispielitems können auf der Webseite des Anbieters heruntergeladen werden (vgl. www.ets.org). Jährlich werden die Testergebnisse aller Testteilnehmer und Testteilnehmerinnen in ihrem letzten Studienjahr in Form von Mittelwert, Streuung und Perzentilen als Vergleichsbasis öffentlich zur Verfügung gestellt (ETS, 2012). Weitere Vergleichsdaten können beim Testanbieter angefordert werden.

Der MFAT-B setzt sich aus 120 Multiple-Choice-Items zusammen. Es werden jeweils vier Antwortoptionen vorgegeben, von denen eine als richtig gewertet wird. Für die Bearbeitungszeit werden zwei Zeitstunden kalkuliert. Ziel des Tests ist es, zum einen das Fachwissen und zum anderen die Anwendung dieses Wissens auf Praxisfälle zu erfassen. Die Items sind textbasiert und enthalten Grafiken, Diagramme und Tabellen. Der Test deckt sieben Subdomänen der Betriebswirtschaftslehre ab. Dazu zählen: (a) S1 – Accounting, (b) S2 – Economics, (c) S3 – Management, (d) S4 – Quantitative Business Analysis and Information Systems, (e) S5 – Finance, (f) S6 – Marketing und (g) S7 – Legal and Social Environment. 12 bis 21 Items repräsentieren dabei je eine Subdomäne (Ling, 2012). Trotz der Gruppierung der Items in fachinhaltliche Subdomänen gehen die Testentwickler von einer eindimensionalen Wissensstruktur aus. Diese Annahmen wurden anhand einer Stichprobe von 155.235 amerikanischen Testteilnehmenden unter Verwendung verschiedener statistischer Methoden empirisch bestätigt (Ling, 2012). Im Rahmen einer Faktoranalyse wies Ling (2012) gute Passungsstatistiken für das eindimensionale Modell nach (CFI = .903 und RMSEA = .015, für weiterführende Informationen zur Interpretation von Modellpassungsstatistiken wird auf Abschnitt 9.3 verwiesen). Darüber hinaus wurde deutlich, dass trotz ausreichender Itemzahl keine der Subdimensionen, eine zufriedenstellende Reliabilität aufwies (Cronbachs Alpha lag zwischen .43 und .65, siehe Tabelle 11 als Interpretationshilfe). Die eindimensionale Version des Tests erzielte hingegen ein Cronbachs Alpha von .89. Die Annahme einer eindimensionalen Struktur wurde dadurch unterstrichen, dass die Vorhersage der wahren Testleistung der Stichprobe durch subdomänenspezifische Schätzer schlechter ausfiel als durch die Gesamtpunktzahl (Ling, 2013, S. 11). Weitere Ergebnisse, die unter Anwendung des MFAT-B entstanden sind, werden in Abschnitt 3.5 dargestellt.

Neben dem amerikanischen Testanbieter ETS gibt es von Seiten der Organisation for Economic Co-operation and Development (OECD) Bestrebungen, wirt-

schaftswissenschaftliches Wissen von Studierenden zu erfassen und international vergleichend zu bewerten. Das jüngste Projekt zu diesem Vorhaben wird im Folgenden vorgestellt.

3.4.2 Assessment of Higher Education Learning Outcomes

Dass es sich bei der Messung von Outputs im Hochschulsektor um ein nationales und internationales Desiderat mit Vorrang handelt, wird unter Betrachtung aktueller Forschungsprojekte deutlich. Seit 2008 plant die OECD das Projekt "Assessment of Higher Education Learning Outcomes" (AHELO), das die Durchführbarkeit einer international vergleichenden Leistungsmessung an Hochschulen erforscht. Im Januar 2010 begann, nach langen Vorbereitungen, die eigentliche Laufzeit des Projekts. Um die Komplexität des Unterfangens vorerst begrenzt zu halten, wurden die zu untersuchenden Inhaltsbereiche auf einige wenige beschränkt. Die ausgewählten Bereiche waren Generic Skills, Economics und Engineering (OECD, 2011). Die Teilnahme an der Studie war freiwillig und wurde von 17 Ländern wahrgenommen. Deutschland beteiligte sich nicht an der Studie.

Das Projekt wurde so konzipiert, dass die Erfassung des Kompetenzniveaus von Bachelorstudierenden im letzten Studienjahr ermöglicht werden sollte. Eine Herausforderung war dabei, die sprachliche und kulturelle Validität der Testinstrumente sicherzustellen. Der Test soll an verschiedenen internationalen Institutionalisierungsformen hochschulischer Bildung zum Einsatz kommen. Ziel ist es, ein indikatorengestütztes Bildungsmonitoring aufzubauen, was schlussendlich der Verbesserung der universitären Bildung dienen soll (Bülow-Schramm & Braun, 2013).

Für die Entwicklung des Ökonomie-Tests hat sich eine Gruppe von nationalen Experten (Group of National Experts on the AHELO Feasibility Study, 2012) im Rahmen des Economic Assessment Frameworks auf fünf übergeordnete Lernziele geeinigt, die in Abbildung 5 dargestellt sind.

Die Lernziele beschreiben die wichtigsten Fähigkeiten, über die ein Bachelorabsolvent am Ende eines wirtschaftswissenschaftlichen Studiengangs verfügen soll und wurden als Orientierungspunkte für die Testentwicklung im AHELO Projekt herangezogen. Sie reichen vom einfachen Verstehen der Fachinhalte, über das Lösen realer Probleme in der Domäne, bis hin zu spezifischen quantitativen und kommunikativen Fähigkeiten innerhalb der Domäne. Darüber hinaus werden selbstregulative Fähigkeiten als Lernziele hervorgehoben. Die in Abbildung 5 dargestellten Lernziele werden im Rahmen der Beschreibung der Testentwicklung in Kapitel 6 erneut aufgegriffen und differenziert dargestellt (vgl. Tabelle 7).

The learning outcomes of the Economics Assessment framework is based on the following five learning outcomes, all of which specify outcomes which students should be able to achieve by the end of their bachelor's degrees:

(i) Demonstrate subject knowledge and understanding;

(ii) Demonstrate subject knowledge and its application to real world problems;

(iii) Demonstrate the ability to make effective use of relevant data and quantitative methods;

(iv) Demonstrate the ability to communicate to specialists and non -specialists; and

(v) Demonstrate the ability to acquire independent learning skills.

Abb. 5: Lernziele des Economics Assessment Frameworks der AHELO Studie

(Quelle: Group of National Experts on the AHELO Feasibility Study, 2012, S. 6)

Der im AHELO Projekt entwickelte Economics Test soll unterschiedliche Schwierigkeitsgrade abbilden und ist auf 90 Minuten Testzeit ausgelegt (Group of National Experts on the AHELO Feasibility Study, 2012). Die erste Hälfte des Tests besteht aus Aufgaben mit einem offenen Antwortformat, mit dem die Fähigkeit, effektiv über Wirtschaftswissenschaften kommunizieren zu können, erfasst werden soll. Aus technischen Gründen konnte jedoch lediglich die schriftliche Kommunikationsfähigkeit erfasst werden (Group of National Experts on the AHELO Feasibility Study, 2012). Der zweite Teil des Tests besteht aus Multiple-Choice-Items, deren Beantwortung 60 Minuten der Testzeit in Anspruch nimmt. Die Definition des zu erfassenden Merkmals wird von der Expertengruppe wie folgt vorgenommen:

> *„The AHELO Economics Assessment does not focus on the recall of factual knowledge, but rather focuses on 'above content' skills including application of concepts, use of appropriate statistical and non statistical tools, drawing conclusions, recommending policy, and being conversant with the 'language of Economics'." (Group of National Experts on the AHELO Feasibility Study, 2012, S. 6)*

Der Test fokussiert volkswirtschaftliche Inhalte und klammert betriebswirtschaftliche Fragestellungen aus.

Die Ergebnisse des AHELO-Projekts wurden in zwei Berichten veröffentlicht. Der erste Bericht beschreibt das Studiendesign und die Umsetzung (Tremblay,

Lalancette & Roseveare, 2012). Der zweite Bericht geht auf die Auswertung und Erfahrungen auf nationaler Ebene ein (AHELO Consortium, 2013). Die Ergebnisse der Studie liefern weniger wissenschaftliche Erkenntnisse als vielmehr forschungspraktische Hinweise zur Entwicklung und Umsetzung von Forschungsvorhaben an Hochschulen mit wirtschaftswissenschaftlichem Fokus. Ein Punkt, der als problematisch herausgestellt wurde, war die geringe Beteiligung der Studierenden an der freiwilligen Testung (Tremblay, Lalancette & Roseveare, 2012). Dieser Punkt wird in der vorliegenden Arbeit aufgegriffen, indem die Testakzeptanz der Studierenden explorativ analysiert wird (vgl. Abschnitt 9.3.4).

3.5 Internationale Forschungsergebnisse

Die internationale Forschung zur Erfassung wirtschaftswissenschaftlichen Wissens ist sehr breit gefächert. Der überwiegende Teil der Studien fokussiert volkswirtschaftliches Wissen und erfasst dieses über die Studienleistung (gemessen durch Modulnoten) in volkswirtschaftlichen Modulen an Universitäten (z. B. Anderson, Benjamin & Fuss, 1994). In diesem Rahmen wurde beispielsweise ein positiver Effekt von Anwesenheit in Lehrveranstaltungen auf Studienleistungen in volkswirtschaftlichen Modulen empirisch bestätigt (z. B. Cohn & Johnson, 2006; Garey & Ellis, 1995). Zudem manifestiert sich in der Mehrzahl der Untersuchungen ein geschlechtsspezifischer Effekt zu Ungunsten weiblicher Studierender (Williams, Waldauer & Duggal, 1992). Studien, die standardisierte Tests volkswirtschaftlichen Wissens, wie den Test of Understanding College Economics (TUCE) als Leistungsindikator heranziehen, sind hingegen seltener (z. B. Agarwal & Day, 1998).

Um dem Fokus dieser Arbeit auf betriebswirtschaftliches Wissen gerecht zu werden, wird im Folgenden nur auf solche internationalen Forschungsarbeiten Bezug genommen, die explizit betriebswirtschaftliches Wissen an Universitäten als Untersuchungsgegenstand fokussieren. In einem ersten Schritt werden Studien vorgestellt, die sich mit den Determinanten der Studienleistung (in Form von Abschlussnoten) befassen. Abschließend werden Studien vorgestellt, in denen Ergebnisse aus standardisierten Leistungstests wie der bereits beschriebene MFAT-B (vgl. Abschnitt 3.4.1) als abhängige Variable fungieren.

Determinanten der betriebswirtschaftlichen Studienleistung

Ein Großteil der Studien zu Lernergebnissen in betriebswirtschaftlichen Studiengängen bezieht sich auf Studiennoten als abhängige Variable (z. B. Eskew & Faley, 1988; Rigney, 2002). So gingen Eskew und Faley (1988) der

Frage nach, mit welchen Determinanten die Leistung der Studierenden unterschiedlicher Hauptfächer im Kurs „Financial Accounting" der Purdue Universität in den Vereinigten Staaten von Amerika vorhergesagt werden kann. Als abhängige Variable fungierte die Gesamtpunktzahl in den vier Leistungskontrollen des Kurses. Als mögliche Prädiktoren wurden die Ergebnisse des Scholastic Aptitude Test (SAT) aus dem sprachlichen und mathematischen Teil und die Abschlussnoten der High School in Mathematik und Englisch als Maß für akademische Leistungsfähigkeit definiert. Zudem wurde die aktuelle Durchschnittsnote im College als Maß der universitätsbezogenen Leistungsfähigkeit herangezogen. Als Maß der kursbezogenen Motivation wurde die Anzahl freiwillig absolvierter Leistungsüberprüfungen betrachtet. Zudem wurde über eine künstlich dichotomisierte (dummy) Variable abgebildet, ob die Studierenden über Vorbildung im Bereich Accounting und verwandten Fächern wie Mathematik oder Statistik verfügten. Als Die Studienerfahrung wurde über die zum Zeitpunkt der Erhebung absolvierten Semesterwochenstunden in die Berechnungen aufgenommen. Mit diesem Modell konnten die Autoren 54 % der Varianz in den Klausurergebnissen des Financial Accounting-Kurses erklären. Den größten Erklärungsanteil hatten die Punkte im SAT und die Teilnahme an freiwilligen Leistungsüberprüfungen. Zudem sagen die Noten der High School und die bisherigen Studienleistungen einen signifikanten Teil der Abschlussnote vorher. Vorerfahrungen in Accounting und verwandten Fächern wirkten sich ebenfalls positiv auf die Leistung aus. Kein signifikanter Zusammenhang konnte hingegen zwischen absolvierten Semesterwochenstunden und Klausurleistung gefunden werden. Allgemeine akademische Leistungsfähigkeit und kursspezifische Motivation sind laut der Autoren die Schlüssel für ein gutes Abschneiden im Financial Accounting-Kurs. Zusammenfassend betonen die Autoren der Studie die positive Rolle von Vorerfahrung und die möglichen Implikationen für die Gestaltung weiterer Collegekurse unter der Berücksichtigung von unterschiedlichen Ausprägungen der Vorbildung.

Aktuellere Studien zur Vorhersage von Studienleistung in betriebswirtschaftlichen Fächern liefern ähnliche Ergebnisse. In einer mit der Studie von Eskew und Faley (1988) vergleichbaren irischen Studie von Rigney (2002) wurde der Frage nachgegangen, ob und wie sich wirtschaftliche und nicht-wirtschaftliche Vorbildung, erworben durch Belegung spezieller Kurse im Leaving Certificate (LCE[3]), vor dem Studium positiv oder negativ auf den späteren Studienerfolg in diesem Bereich auswirkt. Dabei wurden technische Hochschulen in Irland in die Analyse mit einbezogen. Für die Untersuchung wurden die Daten von 200 Studierenden im Bachelor Business Studies (Betriebswirtschaftslehre) aus den 1990er Jahren deskriptiv ausgewertet. Als Maß für den Studienerfolg wurde

3 LCE ist vergleichbar mit der deutschen Hochschulreife.

die erreichte Punktzahl in den Klausuren zum Jahresende herangezogen. Ein zentrales Ergebnis, das mit den Ergebnissen der ILLEV Studie (Förster & Zlatkin-Troitschanskaia, 2010) und der Studie von Eskew und Faley (1988) übereinstimmt, ist der positive Zusammenhang zwischen der allgemeinen Note im LCE und der Gesamtnote im betriebswirtschaftlichen Studienfach. Die Leistungskurse Deutsch und Englisch stehen insbesondere im ersten Studienjahr in einem positiven Zusammenhang mit der Leistung im Bachelor Betriebswirtschaftslehre. Von den betriebswirtschaftlichen Kursen im LCE zeigte sich nur für Kurse im Accounting ein positiver Zusammenhang mit der Studienleistung, der ebenfalls im Zeitverlauf schwächer wird.

In einer Studie mit 455 Teilnehmern eines einführenden Accounting-Moduls in Hong Kong identifizierten Gul und Fong (1993) Selbsteinschätzung der Klausurleistung, Abiturnoten in Englisch und Mathematik, introvertierter Persönlichkeitstyp, vorhandenes Vorwissen über Accounting sowie die Absicht, das Hauptfach Business weiter zu verfolgen als signifikante Prädiktoren der Abschlussleistung. Zudem zeigte sich eine englischsprachige Schulbildung als vorteilhaft, was die Autoren darauf zurückführen, dass ein Großteil der Kursmaterialen auf Englisch zur Verfügung stand (Gul & Fong, 1993).

In einer in Großbritannien durchgeführten Studie fand Halpern (2007) zwar einen Zusammenhang zwischen der Anwesenheit während der Lehrveranstaltungen und Studienleistung im Modul „Business Management", weist aber darauf hin, dass die Anwesenheit bedeutsam von individuellen Charakteristika der Studierenden, z. B. vom Alter und vom kulturellen Hintergrund abhängt.

Einen erweiterten Ansatz der Erklärung betriebswirtschaftlicher Studienleistungen wählte Duff (2004). In einer in Schottland durchgeführten Studie untersuchte er die Zusammenhänge zwischen Kontextvariablen wie Alter, Geschlecht, Studienfach und akademischer Vorleistung und der Studienleistung im ersten Studienjahr sowie die Zusammenhänge zwischen Kontextfaktoren, Lernstrategien und Studienleistung. In dieser Studie mit 60 Studierenden im ersten Semester der Bachelorstudiengänge „Accounting" und „Business Economics" waren schulische Vorleistungen (operationalisiert über die Abschlussnote der Hochschulzugangsberechtigung) der beste Prädiktor für die Studienleistung im ersten Studienjahr (operationalisiert über die Durchschnittsnote in den vier Pflichtmodulen Rechnungswesen, Volkswirtschaft, Recht und Management) und die Vorhersage eines Studienabbruchs. Darüber hinaus konnte der Autor zwei Typen von Lernern identifizieren. Die Gruppe der „ineffektiven Lerner" griff auf oberflächliche Lernstrategien wie Auswendiglernen zurück und wies wenig zielgerichtete Vorbereitung auf. „Effektive Lerner" griffen auf Tiefenlernstrategien wie das Herstellen von Zusammenhängen zurück und verfügten über mehr metakognitive Wahrnehmung und akademisches Selbstbe-

wusstsein (Duff, 2004). Diese beiden Cluster konnten einen signifikanten Anteil an der Varianz des Studienfortschritts nach einem Studienjahr vorhersagen.

Die Studie ist insofern interessant, als dass neben den klassischen Kontextvariablen auch individuelle Lernstrategien berücksichtigt werden. Das hat Implikationen für Instruktionen an Hochschulen, da Lernstrategien prinzipiell durch Interventionen veränderbar sind, während die Mehrzahl der anderen identifizierten Einflussfaktoren einer willentlichen Veränderung weniger oder nicht zugänglich sind (z. B. das Geschlecht der Studierenden). Ein Nachteil der Studie von Duff (2004) ist die geringe Probandenzahl und die Verwendung der mittleren Modulabschlussnoten als Leistungsindikator. Da der Aufbau von Lernstrategien zu den von der OECD anvisierten Lernzielen der Hochschulbildung gehören (vgl. Abbildung 5), werden Lernstrategien als Prädiktor von Testleistungen in der vorliegenden Arbeit mit in Betracht gezogen (vgl. Absatz 9.1).

Zusammenfassend kann gesagt werden, dass die Note der Hochschulzugangsberechtigung häufig einen großen Erklärungswert in Bezug auf die Studienleistung in betriebswirtschaftlichen Fächern aufweist. Im Speziellen die Mathematiknote und die Note in der Sprache des Studiengangs scheinen hohe Vorhersagekraft zu haben. Maße der allgemeinen kognitiven Leistungsfähigkeit hängen ebenfalls positiv mit Leistungen im betriebswirtschaftlichen Studium zusammen. Bis auf wenige Ausnahmen, z. B. bei Bartlett, Peel und Pendlebury (1993) hat sich fachspezifisches Vorwissen als günstig für die Studienleistung, insbesondere zu Beginn des Studiums, erwiesen. Diese Ergebnisse sind konform mit den Untersuchungen zur Auswirkung einer kaufmännischen Berufsausbildung auf die Testergebnisse im WBT (Förster et al., 2012). Motivationale Aspekte scheinen ebenso wie eine extrovertierte Persönlichkeit, der Lerntyp und die Einschätzung der eigenen akademischen Leistung mit der Studienleistung in Verbindung zu stehen. Forschung zu diesen Zusammenhängen ist jedoch noch ausbaufähig. Tabelle 6 gibt einen Überblick über die wichtigsten Prädiktoren betriebswirtschaftlicher Studienleistung und die Studien, in denen der Prädiktor untersucht wurde.

Tab. 5: Zusammenfassung der zentralen Prädiktoren für Studienleistung in betriebswirtschaftlichen Fächern

Prädiktor	Autoren und Jahr der Veröffentlichung
SAT mathematischer und sprachlicher Teil	Eskew und Faley (1988)
Leistungen beim Erwerb der Hochschulzugangsberechtigung	Eskew und Faley (1988) Gul und Fong (1993) Rigney (2002)
Fachspezifische Vorbildung	Eskew und Faley (1988) Gul und Fong (1993)

(Fortsetzung Tab. 5)

Prädiktor	Autoren und Jahr der Veröffentlichung
Motivation	Eskew & Faley (1988) Gul und Fong (1993)
Selbsteinschätzung der akademischen Leistung	Gul und Fong (1993)
Extrovertierter Persönlichkeit und Lerntyp mit überwiegend Tiefenstrategien	Duff (2004) Gul und Fong (1993)
Anwesenheit in Lehrveranstaltungen	Halpern (2007)

Kritisch zu betrachten ist, dass die jeweiligen Modulnoten als abhängige Variable nur unzureichend standardisiert sind und somit eine Verallgemeinerung der Befunde nicht zulässig ist. Im Einzelfall bleibt unklar, aus welchen Elementen die Prüfungsleistungen bestanden, welches Konstrukt gemessen wurde und welche Messgenauigkeit die Modulnoten aufwiesen. Duff (2004) verwies darauf, dass eine differenziertere Bestimmung der Fähigkeit der Studierenden nur durch spezifische Assessments möglich sei.

Determinanten der Leistung in standardisierten betriebswirtschaftlichen Tests

Modulnoten sind nur mit Einschränkungen als Maß des studentischen Wirtschaftswissens interpretierbar (Mirchandani, Lynch & Hamilton, 2001). Jede Universität oder teilweise jede Lehrperson bestimmt, im Rahmen der Vorgaben der Modulkataloge, in der Regel eigenständig über Form, Inhalte und Auswertung der Leistungsprüfung. Damit bleibt im Einzelfall unklar, was durch die Modulnoten gemessen wird (Validitätsaspekt) und ob das, was gemessen werden soll, mit hinreichender Genauigkeit erfasst wird (Reliabilitätsaspekt) (Black & Duhon, 2003). Eine Lösung dieses Dilemmas bieten standardisierte Tests. Diese sind unabdingbar, wenn es um die Erforschung von studentischem Wissen mit Anspruch auf Vergleichbarkeit und Generalisierbarkeit geht. Ein Test wird als standardisiert bezeichnet, wenn Konstruktion, Administration und Auswertung nach uniformen Prozeduren vollzogen werden, sodass Testergebnisse über verschiedene Testeinsätze hinweg konsistent interpretiert werden können (Black & Duhon, 2003, S. 92). Ein in den Vereinigten Staaten von Amerika häufig eingesetzter Test ist der Major Field Test in Business (MFAT-B) (Educational Testing Service, 2011) (vgl. Abschnitt 3.4.1.).

Im Zusammenhang mit dem MFAT-B wurden im englischsprachigen Raum mehrere richtungsweisende Studien zur Struktur und zu den Determinanten betriebswirtschaftlichen Wissens durchgeführt, die im Folgenden vorgestellt

werden. Ein tabellarischer Überblick über die Befunde befindet sich in Tabelle 6.

In einer Studie von Mirchandani et al. (2001) erwies sich das Abschneiden im SAT für 114 Studierende der Betriebswirtschaft der Rowan Universität als deutlich vorhersagestärkster Faktor in Bezug auf die Testleistung. Von den Noten im Studium erwiesen sich die Noten der Module als vorhersagekräftig, die mit quantitativen Inhalten assoziiert sind (z. B. Mathematik, Accounting, Finance, Operations Management und Management Information Systems). Unter Kontrolle der Ergebnisse im SAT erwies sich der Einfluss der Modulnoten in quantitativen Fächern jedoch nur auf einem Level von $p < .10$ als signifikant. Mirchandani et al. (2001) berichteten über Geschlechtereffekte zu Ungunsten der Teilnehmerinnen, gingen aber nicht weiter auf die Ursachen dieses Unterschieds ein, sondern führten als Konsequenz, neben geschlechterübergreifenden Auswertungen, getrennte Berechnungen zur Vorhersage der Testleistung durch. Ein Vergleich der Vorhersagemuster zwischen Gesamtnotendurchschnitt und Testresultaten des MFAT-B interpretierten Mirchandani et al. (2001) als Indikator dafür, dass die Kontrolle der Lernzielerreichung sowohl anhand von universitätseigenen Leistungsprüfungen als auch durch standardisierte Tests stattfinden sollte. In einer aktuelleren Replikation der Studie zur Vorhersage der Testleistung im MFAT-B mit 169 Studierenden der Central Washington University fanden Bagamery, Lasik und Nixon (2005) ähnliche Vorhersagemuster. Die Teilnahme am SAT (die mit besseren Testwerten einhergeht), männliches Geschlecht und der bisherige Notendurchschnitt klärten 46 % der Varianz in den Ergebnissen des MFAT-B auf.

Black und Duhon (2003) sammelten in den Jahren zwischen 1996 und 1997 Daten von 456 Studierenden der Betriebswirtschaftslehre, die den MFAT-B als Teil ihrer Kursleistung bearbeiteten. Der Test erreichte eine Reliabilität von .89, was den Standards für gute wissenschaftliche Praxis im Bereich der Diagnostik entspricht (AERA[4], APA[5] & NCME[6], 2002). In Kombination mit den Daten des universitären Zulassungstests des American College Testing Programms (ACT) lagen den Autoren 297 komplette Datensätze vor. Ihr Modell zur Vorhersage der Testleistung klärte 58 % der Varianz auf und beinhaltete den Notendurchschnitt aller bis zum Testzeitpunkt absolvierten betriebswirtschaftlichen Module als stärksten Prädiktor. Darüber hinaus hatten die Gesamtpunktzahl im ACT, das Alter und männliches Geschlecht eine signifikant positive Vorhersagekraft bezüglich der Testleistung. Studierende mit dem Hauptfach Management schnitten im MFAT-B signifikant schlechter ab als Studierende mit anderen

4 American Educational Research Association
5 American Psychological Association
6 National Council for Measurement in Education

Hauptfächern. Zu den üblichen Variablen der Vorhersage fanden Cox, Charles, Chen und Totten (2011) entgegen ihrer Hypothese einen positiven Einfluss von Berufserfahrung (gemessen in Jahren) auf die Testleistung, aber nicht auf die Durchschnittsnote im Studium. Die Autoren vermuteten, dass der MFAT-B aufgrund seiner breiten fachlichen Ausrichtung mehr Zusammenhangswissen von den Studierenden fordert als die thematisch eng umrissenen Modulabschlussprüfungen. Sie mutmaßten, dass Studierende mit Berufserfahrung erstens reifer sind und zweitens einen besseren Überblick über verschiedene Themengebiete der Betriebswirtschaft haben und diese besser integrieren können.

Aktuelle Studien bestätigen im Wesentlichen die bekannten Vorhersagemuster von betriebswirtschaftlicher Testleistung durch Notendurchschnitt, Geschlecht und SAT (Mason, Coleman, Steagall, Gallo & Fabritius, 2011) und den Vorsprung von Studierenden mit quantitativ orientierter Schwerpunktsetzung im Studium (Settlage & Settlage, 2011). Mason et al. (2011) kontrastierten die Leistungen Studierender unterschiedlicher Schwerpunktfächer zu den Leistungen von Studierenden mit dem Schwerpunkt Management und fanden heraus, dass Studierende aller Schwerpunktfächer außer Marketing und Produktion und Logistik im MFAT-B signifikant besser abschnitten als Studierende mit dem Schwerpunkt Management.

Ein Überblick über ausgewählte Ergebnisse, die unter Einsatz des englischsprachigen MFAT-B erzielt wurden erfolgt in Tabelle 6.

Tab. 6: Überblick über Regressionsanalysen zur Identifikation relevanter Prädiktoren für die Leistung im MFAT-B aus ausgewählten Studien im englischen Sprachraum

Autoren Jahr	**Mirchandani et al. (2001)**	**Black und Duhon (2003)**	**Bagamery et al. (2005)**	**Cox et al. (2011)**	**Mason et al. (2011)**	**Settlage und Settlage (2011)**
N	114	297	169	60	1.411	229
Durchschnittsnote im Studium	.135 ($p < .10$) quantitative Fächer n. s. qualitative Fächer n. s Fächer in VWL	7.49***	10.270***alle Fächer 3.5*quantitative 4.54***Accounting 4.84**Management	14.15***	7.80**	13.002***Business n. s. alle Fächer
SAT/ACT Scores	.585**	1.51***	4.09* (nur Teilnahme)	–	0.04**	1.368***
männliches Geschlecht	–	3.79***	8.90***	–	4.90**	5.199***
Alter	–	0.71***	–	–	–	n. s.
Jahre Arbeitserfahrung	–	–	–	0.91*	–	–
Testvorbereitung	–	–	–	4.25*	–	–
Hauptfach	–	–3.57** Management	–	–	n. s. Marketing & Transport/Logistic 6.62***Accounting 5.27***Finance 3.80***Int. Business 7.66***Economics[1]	–5.274* Marketing –3.154*Business Administration[2]
Asiatische Herkunft	–	–	–	–	–2.13	–
korr. R^2	0.371	0.58	0.46	0.43	0.56	0.525

*$p < .05$; **$p < .01$; ***$p < .001$
n. s. steht für nicht signifikante Zusammenhänge, jeweils nicht erhobene Prädiktoren sind mit einem "-„gekennzeichnet
[1]Alle Hauptfächer verglichen mit Studierenden im Hauptfach Management
[2]Alle Hauptfächer verglichen mit Studierenden im Hauptfach Accounting

3.6 Zusammenfassung des Forschungsstands

Eine Ausgangsfrage zur Messung wirtschaftswissenschaftlichen Wissens ist, inwiefern volkswirtschaftliches und betriebswirtschaftliches Wissen getrennt modelliert werden können und welche Binnenstruktur betriebswirtschaftliches Wissen aufweist. Die Ergebnisse aus Studien, die den WBT und den BAKT einsetzten, wiesen darauf hin, dass volkswirtschaftliches und betriebswirtschaftliches Wissen zwar miteinander verbunden ist, jedoch am besten durch ein zweidimensionales Modell abgebildet wird. Eine anhand von MFAT-B Daten durchgeführte empirische Studie zur Binnenstruktur betriebswirtschaftlichen Wissens (Ling, 2012) untermauerte die Eindimensionalität des Konstrukts für betriebswirtschaftliche Inhalte des Bachelorstudiums. Studien, die auf Daten des BAKT beruhen, bestätigen die eindimensionale Struktur des betriebswirtschaftlichen Wissens ebenfalls weitgehend. Strukturelle Analysen von Bothe et al. (2005) wiesen darauf hin, dass Aufgaben mit Bezug zum Rechnungswesen möglicherweise eine eigene Dimension abbilden. Eine eindeutige empirische Untermauerung dieser Vermutung steht jedoch noch aus. Betriebswirtschaftliches Wissen wird bei der Erfassung von Lernergebnissen an Hochschulen entweder durch die Leistungen in Modulabschlussprüfungen oder durch standardisierte Tests erfasst; wobei in Deutschland mit dem BAKT (Bothe, 2003) zum jetzigen Zeitpunkt nur ein einziger standardisierter Test für betriebswirtschaftliches Wissen auf Hochschulniveau vorliegt. Im Gegensatz dazu liegen im englischsprachigen Bereich mehrere Studien vor, die sich mit der Vorhersage von Testleistungen in betriebswirtschaftlichen Leistungstests befassen und dazu den etablierten MFAT-B (ETS, 2012) nutzen. Die Testleistung stand in allen Studien in einem signifikanten Zusammenhang mit der allgemeinen akademischen Leistungsfähigkeit und ließ sich in der Regel sowohl durch Leistungsindikatoren vor dem Studium als auch durch Leistungsindikatoren während des Studiums vorhersagen. Dabei scheint eine quantitative Ausrichtung des Studiums der Wissensentwicklung im betriebswirtschaftlichen Bereich förderlich zu sein. Der positive Einfluss des Status als Muttersprachler wies darauf hin, dass betriebswirtschaftliches Wissen, neben quantitativen Fähigkeiten, sprachliche Fähigkeiten voraussetzt. Diese Annahme deckt sich mit dem kaufmännischen Domänenmodell von Winther (2010), das der betriebswirtschaftlichen Domäne rechnerische, sprachliche und technische Zugänge zuschreibt (vgl. Abschnitt 4.1). Der Zusammenhang zwischen Geschlecht und Testleistung ist eindeutig, seine Ursachen wurden bisher jedoch nicht hinreichend untersucht. In zahlreichen Studien wurde ein Effekt zu Ungunsten weiblicher Testteilnehmender festgestellt. Nur in wenigen Studien wurde dieser Zusammenhang nicht sichtbar (Bycio & Allen, 2007). Es liegen keine Studien vor, in denen weibliche Testteilnehmer den männlichen systematisch überlegen waren. Es lässt sich

demnach schlussfolgern, dass männliche Teilnehmer tendenziell über mehr testrelevantes betriebswirtschaftliches Wissen verfügen. Ob es sich dabei um eine Verzerrung handelt, die durch geschlechtsdiskriminierende Items entsteht (vgl. Ausführungen zu Differential Item Functioning in Abschnitt 7.2.4) oder ob dies einen tatsächlichen Wissensunterschied zwischen den Geschlechtern widerspiegelt, ist aus den bisherigen Forschungsarbeiten nicht abzuleiten. Möglicherweise unterliegt der Testbearbeitungsprozess bei weiblichen Teilnehmern anderen Prozessen als bei männlichen. So identifizierten Bielinska-Kwapisz und Brown (2013) kritisches Denken und motivationale Aspekte als Ursachen für geschlechtsspezifische Unterschiede. Denkbar sind auch leistungsrelevante Unterschiede in der Einschätzung der eigenen Leistungsfähigkeit in betriebswirtschaftlichen Fragestellungen. Eine theoretisch fundierte Begründung des Leistungsunterschieds zwischen den Geschlechtern in betriebswirtschaftlichen Testsituationen steht noch aus. Der Zusammenhang der Testleistung mit fachspezifischen und beruflichen Vorerfahrungen wurde in mehreren Studien bestätigt. Tendenziell weisen die empirischen Ergebnisse darauf hin, dass sich berufliche und fachliche Vorerfahrungen positiv auf die Testleistung in betriebswirtschaftlichen Tests auswirken. Diese Vermutung wird von den positiven Effekten einer Berufsausbildung auf die Ergebnisse im WBT (vgl. Abschnitt 3.1.1) untermauert.

Die nationalen und internationalen Ergebnisse zur Messung wirtschaftswissenschaftlichen Wissens geben zum einen erste Hinweise zur Struktur betriebswirtschaftlichen Wissens und zum anderen wichtige Hinweise zu dessen Zusammenhang mit relevanten Außenkriterien. Die empirischen Befunde der bisherigen Forschung dienen in der vorliegenden Arbeit der Kriteriums- und Konstruktvalidierung. Da theoretische Überlegungen und die empirische Messung von betriebswirtschaftlichem Wissen an deutschen Hochschulen noch nicht weit voran geschritten sind, wird in Kapitel 4 auf Konzepte der Kompetenzmessung in kaufmännischen Handlungsfeldern zurückgegriffen, um insbesondere die Testentwicklung der vorliegenden Arbeit wissenschaftlich zu fundieren.

4 Konzepte der Kompetenzmessung in kaufmännischen Handlungsfeldern

Aus der Problemstellung und dem Forschungsstand geht das Forschungsdesiderat hervor, einen Test mit betriebswirtschaftlichen Inhalten für den Einsatz an deutschen Hochschulen zu entwickeln, der über deklaratives Wissen hinaus anwendungsorientiertes Wissen erfasst. Dieser Test soll modernen psychometrischen Anforderungen genügen und eine Abbildung studentischen Wissens unterschiedlicher quantitativer und qualitativer Ausprägung ermöglichen. Der Test soll den in Abschnitt 3.1.2 vorgestellten deklarativen Test (den BAKT) um situative Items zur Abfrage weiterer Wissensformen ergänzen. Ziel ist dabei, eine anwendungsnahe Wissensdimension zu messen, die prozedurale Strukturen abbildet. Diese Anforderungen setzen ein wissenschaftlich fundiertes Vorgehen bei der Testentwicklung voraus. Aufgrund der inhaltlichen Nähe zu jüngeren Arbeiten aus dem Bereich der Kompetenzmessung in der kaufmännischen Bildung (z. B. Winther, 2010), orientiert sich die Testentwicklung neben den Standards der pädagogisch-psychologischen Diagnostik (AERA et al., 2008) an aktuellen Erträgen aus der berufs- und wirtschaftspädagogischen Forschung und Modellen der empirischen Bildungsforschung. Als Grundlage der in Kapitel 6 beschriebenen Testentwicklung werden in diesem Kapitel Konzepte der Kompetenzmodellierung und Messung in kaufmännischen Handlungsfeldern betrachtet.

Im ersten Abschnitt wird die Strukturierung der kaufmännischen Domäne in der wirtschaftspädagogischen Forschung vorgestellt. Das konzeptuelle Verständnis der Domäne ist ein grundlegendes Gestaltungskriterium für die Testentwicklung (Winther, 2010). Obwohl die betriebswirtschaftliche Domäne nicht mit der kaufmännischen Domäne gleichzusetzen ist, werden im folgenden Kapitel grundlegende Ansätze zur Strukturierung der kaufmännischen Domäne beschrieben, um diese im Rahmen der Testentwicklung auf den betriebswirtschaftlichen Kontext anzuwenden (vgl. Kapitel 6). Anschließend werden aktuelle Forschungsarbeiten zu Kompetenzmodellierung und -messung, unter besonderer Berücksichtigung situativer Aufgaben, vorgestellt. Das Kapitel schließt mit einer Zusammenfassung des Forschungsstands.

4.1 Zur Strukturierung der kaufmännischen Domäne

Während zur Struktur der betriebswirtschaftlichen Domäne kaum Forschungsarbeiten vorliegen, hat die berufs- und wirtschaftspädagogische Forschung in den letzten Jahren insbesondere Kompetenzen in unterschiedlichen kaufmän-

nischen Ausbildungsberufen untersucht. Zu diesen Forschungsarbeiten gehört auch eine differenzierte Betrachtung der kaufmännischen Domäne. Der Domänenbegriff umschreibt fachspezifische Leistungsbereiche, die sich über ausgewählte Anforderungssituationen charakterisieren lassen (Winther, 2010). Während im allgemeinbildenden Bereich in der Regel auf eine generelle übergeordnete Fachinhaltsstruktur abgestellt wird, öffnet die berufs- und wirtschaftspädagogische Forschung den Domänenbegriff und geht von einem übergeordneten, sinnstiftenden Handlungskontext aus (Winther & Achtenhagen, 2008). Damit beschreibt die Domäne das Handlungsfeld, in dem eine Person agiert und ihre Kompetenzen relevant sind. Wissen, in seinen unterschiedlichen Ausprägungsformen (vgl. Abschnitt 2.3), ist dabei ein wichtiger Teilaspekt der Bewältigung von Anforderungen innerhalb einer Domäne. Im Rahmen der Entwicklung von Leistungstests gibt die Domäne vor, welche Leistungsbereiche testrelevant sind und welche nicht. Dafür muss die Domäne im Vorfeld der Testentwicklung eingegrenzt und eine Struktur vorgegeben werden.

Prinzipiell bieten sich zwei Ansätze zur Eingrenzung und Strukturierung einer Domäne an: (1) über die Analyse der Arbeitsplatzanforderungen ausgewählter beruflicher Handlungsfelder und (2) über die Analyse vorhandener Curricula. Der curriculare Ansatz kann, zumindest im Rahmen der beruflichen Kompetenzmessung, in zwei weitere Zugänge differenziert werden. Zum einen kann eine curriculare Analyse anhand der Fachinhaltsstruktur vorgenommen werden. Zum anderen kann die Domäne anhand von Lernfeldern strukturiert werden. Lernfelder sind in den Kultusministerkonferenz-Rahmenlehrplänen beschriebene thematische Einheiten, die sich an konkreten beruflichen Aufgabenstellungen und Handlungsabläufen orientieren (Pätzold, 2003, S. 139). Sie zeichnen sich durch eine große Handlungsnähe aus und sind seit dem Jahr 1996 für alle neu geordneten Ausbildungsberufe vorgesehen (Schäfer, 1999).

Bei einer curricular orientierten Domänenstrukturierung stehen die Analysen des Lehrplans und der Lehrmaterialien im Mittelpunkt. Dabei sollte zwischen intendiertem Curriculum (was ist vom Lehrplan vorgesehen) und implementiertem Curriculum (was wird tatsächlich umgesetzt) unterschieden werden. Die Strukturierung des implementierten Curriculums ist weniger eindeutig, da als Informationsquelle häufig punktuelle Experteneinschätzungen herangezogen werden, die schwerer objektivierbar sind. Auch die Analyse von Lehrmitteln und Prüfungsmaterialien unterliegt der Restriktion, jeweils nur für die untersuchten Klassen und den Untersuchungszeitraum gültig zu sein. Eine lernfeldorientierte Strukturierung bietet den Vorteil, dass sie sich an tatsächlichen beruflichen Tätigkeiten orientiert. Eine Strukturierung anhand konkreter beruflicher Tätigkeiten ist jedoch nur dann sinnvoll, wenn die Domäne eng umrissen ist

(z. B. kaufmännische Kompetenzen in einer bestimmten Branche). Andernfalls besteht die Problematik, dass Domänenmodelle, die auf Grundlage von Arbeitsplatzanalysen innerhalb einer Branche entwickelt wurden, nicht auf andere Branchen übertragbar sein könnten.

Winther (2010) stellte für die kaufmännische Bildung ein Domänenmodell vor, das aus einer Mischform der beiden dargestellten Strukturierungsansätze beruhte. Für das Domänenmodell, das auf der Basis zweier kaufmännischer Ausbildungsgänge aufbaute, wurden sowohl Unternehmensprozesse als auch schulische Lern- und Arbeitsanforderungen analysiert und systematisiert (Winther, 2010). Zum einen fußte das Modell auf der vollzeitschulischen Ausbildung am Fachgymnasium Wirtschaft und zum anderen auf der dualen Ausbildung zum Industriekaufmann/zur Industriekauffrau (Winther, 2010). Bei der Erstellung des Domänenmodells ging die Autorin der Frage nach, was die Zugänge zu einem beruflichen Handlungsbereich sind und welche Denkfiguren und Begriffe zentral für das Verständnis des beruflichen Handlungsbereichs sind (Winther, 2010, S. 87). Die Identifikation von Lern- und Arbeitsanforderungen erfolgte bei Winther (2010) über die Dreiteilung der Unternehmensprozesse in Wertschöpfungsprozesse, Steuerungsprozesse und Unterstützungsprozesse. Vor dem Hintergrund dieser Prozesse ließen sich reale Arbeitsprozesse oder auch Lern- und Handlungsprozesse strukturieren (Winther, 2010). Eine solche Struktur ermöglichte die Zuordnung von Lern- und Arbeitsanforderungen zu den Unternehmensprozessen, die wiederum als Grundlage der Itementwicklung herangezogen wurden (Winther, 2010).

In der kaufmännischen Bildung liegen sowohl schulisch als auch betrieblich relativ homogene Anforderungen vor. Die universitäre, betriebswirtschaftliche Bildung stellt in Bezug auf die Strukturierung der Domäne, durch ihre heterogenen Anforderungen, eine besondere Herausforderung dar (Förster et al., 2012). Neben der von Winther (2010) gewählten Möglichkeit, die Domäne anhand einer Mischform zwischen curricularen Anforderungen und Arbeitsplatzanforderungen zu strukturieren, bietet sich für die Strukturierung der Domäne im Hochschulbereich ein curricular-orientierter Zugang an, wie er von Bothe (2003) bei der Entwicklung des BAKT verfolgt wurde (vgl. 3.1.2). Während eine Strukturierung der Domäne anhand von Geschäftsprozessen möglicherweise eine berufsbezogene, realitätsnahe Strukturierung der betriebswirtschaftlichen Domäne erlaubt, gewährleistet eine curriculare Strukturierung die Abbildung betriebswirtschaftlichen Wissens mit Bezug zu der in der universitären Bildung üblichen, fachspezifischen Didaktik. Bei diesem Zugang werden übergeordnete Fachinhaltsstrukturen innerhalb der Betriebswirtschaft als Ausgangsstruktur der Testentwicklung herangezogen. Eine direkte Übernahme der Struktur von Bothe (2003) (vgl. Abbildung 4) ist aus zwei Gründen nicht mög-

lich: (1) beruhen die curricularen Analysen des BAKT auf dem alten Studienmodell der Diplom-Abschlüsse und (2) soll der zu entwickelnde Test für Studierende der Wirtschaftspädagogik valide sein, was zu Einschränkungen in der Domänenstrukturierung führt.

Ein mit Bothe (2003) vergleichbares Vorgehen wählten Schumann et al. (2010) im Projekt „Ökonomische Kompetenzen von Maturandinnen und Maturanden" (OEKOMA). Sie nahmen unter der Bedingung heterogener Bildungsanforderungen eine umfassende Kategorisierung fachlicher Inhaltsstrukturen im Studium der Wirtschaftswissenschaften in der Schweiz vor. Die Analyse bezog sich auf Studienunterlagen (Vorlesungsskripte, Lehrmittel, Übungsmaterialien und Prüfungen) des ersten Studienjahres der Wirtschaftswissenschaften der Universität Zürich und der Universität St. Gallen. Das Ergebnis ist jedoch nur bedingt auf das deutsche Studiensystem übertragbar, da es nur das erste Studienjahr umfasst und stark durch schweizspezifische Hochschulkonzepte geprägt ist. Das Projekt OEKOMA wird in Abschnitt 4.2.1 weiterführend beschrieben.

Für die Strukturierung der Domäne im Bereich der hochschulischen Betriebswirtschaftslehre wird aus den oben angeführten Überlegungen folgendes Fazit abgeleitet: Entsprechend der an Fächern und nicht an Handlungsfeldern orientierten Struktur der universitären Bildung wird in der vorliegenden Arbeit die betriebswirtschaftliche Domäne vorwiegend aus curricularer Perspektive strukturiert. Um den Bezug zu möglichen beruflichen Tätigkeiten herzustellen, werden die den Curricula entnommenen Fachinhalte mit Anforderungen aus Stellenanzeigen abgeglichen und bei der Entwicklung der Aufgaben berücksichtigt. Weiterführende Erläuterungen zur Strukturierung der Domäne und zur Testentwicklung folgen in Kapitel 6.

Eng mit dem Domänenverständnis verknüpft sind Kompetenzmodelle. Kompetenzmodelle zeichnen sich einerseits durch Aussagen über die Art und Anzahl der zu unterscheidenden Kompetenzdimensionen innerhalb einer bestimmten Domäne (Strukturaspekt) und andererseits über die Bestimmung von Anforderungsdimensionen (Graduierungsaspekt) aus (Seeber, 2008).

Ein solches Kompetenzmodell liegt zum jetzigen Zeitpunkt für die Anforderungen an Bachelor-Absolventen betriebswirtschaftlicher Studiengänge nicht vor. Ebenso mangelt es an systematischen Beschreibungen der Domäne. Jedoch liegen für den kaufmännischen Bereich grundlegende Arbeiten vor, in denen empirisch fundierte Konzepte der Kompetenzmessung vorgestellt werden (z. B. Seeber, 2008; Winther & Achtenhagen, 2008; Winther, 2010). Die Erträge der berufs- und wirtschaftspädagogischen Forschung im Bereich der Kompetenzmodellierung werden im Folgenden vorgestellt.

4.2 Kompetenzmodellierung in der kaufmännischen Bildung

Neben der Systematisierung der Domäne sind Überlegungen zu Kompetenz-, insbesondere Wissensstrukturen, die für die Bewältigung von Anforderungen innerhalb der abgegrenzten Domäne benötigt werden, ein grundlegendes Element der Testentwicklung. Da für den Bereich des betriebswirtschaftlichen Hochschulwissens auf Bachelorniveau zum jetzigen Zeitpunkt keine veröffentlichten Kompetenzmodelle vorliegen, werden im Folgenden zentrale Befunde der berufs- und wirtschaftspädagogischen Forschung herangezogen, um das Verständnis kaufmännischer Kompetenzen zu schärfen. Im ersten Schritt werden Kompetenzstrukturmodelle vorgestellt, bei denen die Frage nach der Dimensionalität der Kompetenz im Mittelpunkt steht (Abschnitt 4.2.1). Im zweiten Schritt werden Konzepte der Kompetenzniveaumodellierung mit schwierigkeitsbestimmenden Aufgabenmerkmalen dargestellt.

4.2.1 Kompetenzstrukturmodelle

Eine erste umfassende Studie im beruflichen Bereich zur Modellierung beruflicher Fachkompetenzen für Ausbildungsberufe aus dem gewerblich-technischen und kaufmännischen Bereich erfolgte durch Lehmann und Seeber (2007) in den Untersuchungen von Leistungen, Motivation und Einstellungen in Abschlussklassen beruflicher Schulen (ULME). Das Testinstrument für den Bereich Wirtschaft und Verwaltung wurde so konstruiert, dass post-hoc Analysen zu Kompetenzniveaus und Kompetenzstruktur ermöglicht wurden. Testaufgaben zur Erfassung der kaufmännischen Fachkompetenzen wurden nach Wissens- und Verhaltensdimensionen in Anlehnung an Anderson und Krathwohl (2001) klassifiziert. In der leicht modifizierten Matrix zu kognitiven Strukturen wurde zwischen drei Wissenskategorien (deklaratives, konzeptuelles und prozedurales Wissen) und drei Verhaltensdimensionen (Reproduzieren, Verstehen/Anwenden und Kritisieren/Reflektieren) unterschieden (Hofmeister, 2005). Ziel war es, neben deklarativen Wissensbeständen auch konzeptuelle und prozedurale Wissensleistungen zu erfassen (Lehmann & Seeber, 2007). Im Rahmen der ULME-Studien (Lehmann & Seeber, 2007) überprüfte Seeber (2008) die Anforderungsstruktur eines berufsbezogenen Fachleistungstests für Bürokaufleute. Der Test war einer von 17 Fachleistungstests, die in ULME III eingesetzt wurden und bezieht sich auf die Erfassungen kontextspezifischer kognitiver Leistungsdispositionen (Seeber, 2008). Die Testaufgaben sollten möglichst die Breite berufsfachlicher Anforderungen abbilden und in unterschiedlichen Kontexten situativ eingebunden werden (Seeber, 2008). Die Auswahl der Aufgabeninhalte erfolgte durch Sichtung der vorliegenden Rahmenlehrpläne und

der Hamburger Bildungspläne für den jeweiligen Ausbildungsberuf, die der Lernfeld- und Kompetenzorientierung folgten (Seeber, 2008). Um neben dem intendierten Curriculum auch das implementierte Curriculum zu berücksichtigen, wurden Lehrbücher und Aufgaben der Kammerprüfungen in die Analyse einbezogen; zudem begleiteten Lehrkräfte den Prozess der Itementwicklung. Der Test enthielt (1) Aufgaben mit Bezug zu gesamtwirtschaftlichen Zusammenhängen (volkswirtschaftliche Dimension), (2) Aufgaben, die sich auf betriebswirtschaftliche Organisations- und Leistungsprozesse bezogen, (3) Aufgaben, die sich auf Rechtsnormen wirtschaftlichen Handelns bezogen und (4) Aufgaben, die sich stark auf Prozeduren aus dem Bereich Rechnungswesen bezogen und nur schwach mit betriebswirtschaftlichem Inhaltswissen verknüpft waren (Wertschöpfungsdimension) (Seeber, 2008, S. 79 f.). Für das Verständnis der kaufmännischen Domäne war speziell die Betrachtung der Dimensionalität des Tests von Bedeutung. Es galt zu prüfen, ob die Anforderungsstrukturen des Tests besser durch eine komplexe, aber homogene Fähigkeitsdimension abgebildet werden oder ein mehrdimensionaler Bezugsrahmen die Daten besser beschreibt (Seeber, 2008). Ein Vier-Faktoren-Modell, das die oben genannten Inhaltsbereiche als einzelne Faktoren modellierte, erwies sich im Vergleich zu einer einfaktoriellen Lösung als empirisch nicht haltbar (Seeber, 2008). Jedoch konnten zwei Inhaltsbereiche herauskristallisiert werden, die jeweils durch spezifische Verständnisleistung beeinflusst wurden. Zu diesen Inhaltsbereichen zählten betriebs- und volkswirtschaftliche sowie rechtliche Aspekte einerseits und Aspekte des Rechnungswesens andererseits (Seeber, 2008). Für die Verhaltensdimensionen wurde in der Testkonstruktion von einer ansteigenden Komplexität ausgegangen, die sich allerdings empirisch nicht eindeutig bestätigen ließ (Seeber, 2008; Lehmann & Seeber, 2007). Ähnlich unklare Ergebnisse dokumentierte Witt (2006) zuvor im Rahmen von Untersuchungen mit dem WBT, was den Nutzen der Verhaltensdimension nach Anderson und Krathwohl (2001) für die Modellierung von Kompetenzstufen in Frage stellt. Die Darstellung und kritische Betrachtung der kognitiven Anforderungen, die mit der Aufgabenschwierigkeit assoziiert sind, wird in Abschnitt 4.2.2 vorgenommen.

Winther und Achtenhagen (2008) stellen ein Kompetenzmodell vor, das zwischen domänenverbundenen und domänenspezifischen Kompetenzen differenziert. Zu den domänenverbunden Kompetenzen gehören die Bereiche „economic literacy" im Sinne einer kulturellen Teilhabe in wirtschaftsbezogenen Kontexten und „economic numeracy" im Sinne grundlegender mathematischer Kenntnisse im Rahmen von Unternehmensprozessen. Der domänenspezifische Kompetenzbereich wird „Geschäftsvorfall" genannt und bezieht sich auf die Bearbeitung komplexer ökonomischer Zusammenhänge auf Basis festgelegter Arbeits- und Geschäftsprozesse. Das Kompetenzstrukturmodell erhebt

den Anspruch, kaufmännisch-berufliche Kompetenzen über begründete Teilkompetenzen definieren und abbilden zu können (Winther & Achtenhagen, 2008). Die Autoren sehen eine zwischen domänenverbundenen und domänenspezifischen Kompetenzen differenzierende Konzeptualisierung empirisch bestätigt. In einer weiterführenden Arbeit (Winther & Achtenhagen, 2009) wurde auf der Basis dieses Kompetenzstrukturmodells für die kaufmännische Bildung die web-basierte Computersimulation ALUSIM entwickelt und mit 264 Auszubildenden im Ausbildungsberuf Industriekaufmann/Industriekauffrau getestet. Zentraler Bestandteil der Computersimulation waren ausgewählte authentische berufliche Anforderungssituationen, die innerhalb von Arbeits- und Geschäftsprozessen von Industrieunternehmen eingebettet wurden. Die Autoren legen eine Beschreibung beruflicher Handlungskompetenz über zwei Skalen nahe: die Skala der handlungsbasierten und die Skala der verstehensbasierten Kompetenz. Erstere zielt vorwiegend auf betriebliche Strukturen und Entscheidungsprozesse innerhalb authentischer betrieblicher Anforderungssituationen ab, während letztere die Bewältigung verstehensbasierter Anwendungsaufgaben vor dem Hintergrund betrieblicher Situationen widerspiegelt. Darüber hinaus zeigte sich, wie im Modell angenommen, auf inhaltlicher Ebene eine Differenzierbarkeit zwischen Wertschöpfungsprozessen und betrieblichen Steuerungsprozessen (Winther & Achtenhagen, 2009). Die Unterscheidung zwischen „handlungsnahen Kompetenzen" und „verstehensbasierten Kompetenzen" ist sowohl methodisch als auch theoretisch problematisch. Es ist bekannt, dass sogenannte „Methodeneffekte" die Messung eines latenten Konstrukts maßgeblich beeinflussen können (Schermelleh & Schweizer, 2012). Methodeneffekte beschreiben Varianzquellen, die sich über die gemessene Fähigkeit hinaus auf die Validität der Messung auswirken können. Mögliche Ursachen dafür sind z. B. Charakteristika des Messinstruments oder systematisch variierende Bedingungen der Messung. Die in einer Untersuchung gefundene systematische Varianz kann somit nicht nur von dem gemessenen Konstrukt resultieren, sondern darüber hinaus auch von der verwendeten Methode. Unter diesem Gesichtspunkt leistet die Umsetzung der zwei Kompetenzskalen bei Winther und Achtenhagen (2009) in sehr unterschiedlichen Itemformaten, nämlich einer computerbasierten Simulation für „verhaltensbasierte Kompetenzen" und einem papierbasierten Fragebogen für „verstehensbasierte Kompetenzen", der Vermutung Vorschub, dass die Unterscheidung der beiden Dimensionen durch ein methodisches Artefakt zustande gekommen sein könnte. Eine Möglichkeit, diese Methodeneffekte aus der Messung eines Konstrukts herauszurechnen, ist ein Multitrait-Multimethod-Design (siehe Campbell & Fiske, 1959).

Auf theoretischer Basis ist die Benennung der zwei Kompetenzdimensionen in „verstehensbasiert" und „handlungsbasiert" problematisch, da diese nicht im

Rahmen von bestehenden kognitionspsychologischen oder didaktischen Modellen des Wissens- oder Kompetenzerwerbs eingebunden sind. Die Bezeichnungen sind insofern missverständlich als dass „Verstehen" in den gängigen Wissenstaxonomien (z. B. Anderson & Krathwohl, 2001) eine Bezeichnung für eine kognitive Prozessdimension ist und keine eigene Wissensform.

Rosendahl und Straka (2011) führten Analysen zur wirtschaftlichen Fachkompetenz angehender Bankkaufleute durch. Sie bauten das Testmaterial zur Testung bankwirtschaftlicher Kompetenzen anhand von fünf bankspezifischen Situationen auf, z. B. Kontoeröffnung und Kontoführung oder Kontoführung im Todesfall. Zu den Situationen wurden 22 Teilaufgaben mit unterschiedlichen Antwortformaten gestellt, die mit dem Curriculum der Zielgruppe abgestimmt waren. Wirtschaftliche Fachkompetenzen wurden in eine allgemeinwirtschaftliche und eine bankwirtschaftliche Dimension unterteilt. Allgemeinwirtschaftliche Kompetenz wurde im Sinne einer wirtschaftlichen Grundbildung definiert, bankwirtschaftliche Kompetenz umfasst kognitive Leistungsdispositionen, die für die Bewältigung domänenspezifischer beruflicher Anforderungen (z. B. Führen von Konten, Anlageberatung) benötigt werden (Rosendahl & Straka, 2011). Diese im Vorfeld getroffene Unterscheidung, anstelle der Annahme einer übergreifenden wirtschaftlichen Leistungsdisposition, wurde in der Studie durch Modellvergleiche empirisch bestätigt. Darüber hinaus wurden Modelle geprüft, die Fachkompetenz weiter ausdifferenzierten. Diese Analysen bestätigen, dass Teilkompetenzen nach fachwissenschaftlichen Kategorien unterschieden werden können, wie bereits von (Seeber, 2008) gezeigt. Der post-hoc Charakter der Analysen und die anhand von Strukturgleichungsmodellen unbefriedigenden ermittelten Modellanpassungen schränken jedoch die Interpretierbarkeit dieser Ergebnisse ein (Rosendahl & Straka, 2011). Interessanterweise führte eine Strukturierung bankwirtschaftlicher Kompetenz nach Fallsituationen, die Anforderungen nach typischen Situationen des Bankalltages widerspiegeln, zu verbesserter Modellpassung. Dieses Ergebnis steht im Einklang mit den Befunden von Winther und Achtenhagen (2008). Die Autoren schlagen vor, zukünftig zu prüfen, ob Modelle, die eine kombinierte Dimensionierung nach fachwissenschaftlichen Aspekten und Arbeits-/Geschäftsprozessen abbilden (im Sinne einer „within-item-multidimensionlity"), die Zusammenhänge zwischen Aufgabenantworten besser erklären als die bisherigen separat durchgeführten Analysen.

Das Schweizer Projekt „Ökonomische Kompetenzen von Maturandinnen und Maturanden" (OEKOMA) hatte zum Ziel, ökonomische Kompetenzen als Gesamtheit des Wissens, der Fertigkeiten und Bereitschaften eines Individuums, wirtschaftliche Problemstellungen zu lösen, zu erfassen (Schumann et al., 2010). Als Itemformat wurden im Sinne einer „economic literacy" Itemstämme

in Form von Zeitungsartikeln entwickelt, denen vier bis fünf Einzelitems untergeordnet wurden (Schumann et al., 2010). Dieser Testkonstruktion lag die Annahme zu Grunde, dass mündige Wirtschafts- und Gesellschaftsbürger in der Lage sein sollten, authentische und alltagsbezogene Darstellungen zu wirtschaftsbezogenen Problemstellungen zu verstehen, zu analysieren und begründete Lösungen zu generieren (Schumann et al., 2010). Die Authentizität der Zeitungsartikel wurde durch eine aufwändige Medienanalyse abgesichert (Schumann et al., 2010). Mit 1500 bis 2000 Zeichen waren die modifizierten Zeitungsartikel jedoch verhältnismäßig lang und limitierten unter beschränkter Testzeit die Anzahl der einsetzbaren Aufgaben pro Testteilnehmer. Im Inhaltsbereich BWL wurden die drei Unterkategorien „Strategische Unternehmensführung", „Bereiche der Unternehmung" und „Corporate Finance" gebildet. Zum „Bereich Strategische Unternehmensführung" wurden zwei Itemstämme gebildet, zu den weiteren Unterkategorien jeweils 4. Auf die Unterkategorien verteilten sich 12–14 Items (Schumann et al., 2010). Eine empirische Abgrenzung der einzelnen Subkategorien ließ sich jedoch nicht eindeutig bestätigen (Schumann, 2013). Vor- und Nachteile einer szenariobasierten Umsetzung von Testaufgaben werden in Kapitel 6 zur Testentwicklung aufgegriffen und diskutiert.

In einer Studie im gewerblich-technischen Bereich wurden Simulationsaufgaben und situative paper-pencil Aufgaben am Ende des ersten Ausbildungsjahres eingesetzt, um fachspezifische Problemlösefähigkeiten zu messen (Gschwendtner, Geißel & Nickolaus, 2010). Die fachbezogene Problemlösefähigkeit bildete neben dem Fachwissen eine eigene Dimension (Gschwendtner et al., 2010). Prozedurales und deklaratives Wissen ließen sich in der Studie hingegen ebenso wie unterschiedliche fachinhaltliche Dimensionen nicht empirisch voneinander trennen (Gschwendtner et al., 2010). Ein interessanter Befund war, dass das Wissen von Kraftfahrzeugmechatroniker/innen zu Beginn des ersten Ausbildungsjahres noch in zwei fachspezifische Bereiche unterteilt werden konnte (fahrzeugmechatronisches und elektronisches Fachwissen). Diese Befundlage interpretierten Gschwendtner et al. (2010, S. 266) als eine „kognitive Integrationsleistung", die im ersten Ausbildungsjahr vorgenommen wird. Die Ergebnisse wiesen darauf hin, dass sich die kognitiven Wissensrepräsentationen über den Zeitverlauf ändern und mit voranschreitendem Lernprozess weiter vernetzen. Gleichzeitig machen die Ergebnisse deutlich, dass bisherige strukturelle Modelle den Aspekt der Kompetenzentwicklung über die Zeit noch weitgehend unberücksichtigt gelassen haben und diesbezüglich ein Forschungsbedarf besteht.

Eine Herausforderung der kaufmännischen Kompetenzmessung ist die Identifikation der für kompetentes Handeln relevanten Wissensbasis und der damit

verbundenen kognitiven Leistungen (Tramm & Seeber, 2006). Die Bemühungen, die kaufmännische Domäne so zu strukturieren, dass eine theoretisch fundierte und empirisch bestätigte Testentwicklung daraus abgeleitet werden kann, sind vor dem Hintergrund der bisherigen Befunde als noch nicht abgeschlossen zu bewerten. Die vorgestellten Kompetenzmodelle geben unterschiedliche Hinweise auf die Dimensionalität kaufmännischer Kompetenzen. Insbesondere variieren die Ergebnisse berufsfeldspezifisch. Diese Ergebnislage weist auf einen erheblichen Entwicklungsbedarf von diagnostischen Instrumenten hin, die für einen breiten Einsatz in verschiedenen Berufsfeldern geeignet sind (Seeber et al., 2010). Die zu leistenden Entwicklungsarbeiten schließen die Herausarbeitung konstitutiver Merkmale der kaufmännischen Domäne sowie deren Abgrenzung ein (Seeber et al., 2010). Möglicherweise muss dabei auch berücksichtigt werden, auf welcher Wissensintegrationsstufe sich der Lernende befindet. Darüber hinaus steht eine kriteriale Absicherung zwischen den bestimmten Kompetenzen und späterem Berufserfolg noch aus (Seeber et al., 2010).

Neben den strukturellen Überlegungen zu kaufmännischen Kompetenzen sollten auch schwierigkeitsbestimmende Aufgabenmerkmale bei der Testentwicklung in Betracht gezogen werden. Damit gehen Modelle über das Kompetenzniveau einher, welches es zu erfassen gilt. Bei der Kompetenzniveaumodellierung gilt es theoretisch-didaktisch zu begründen, welche Aufgaben innerhalb einer Domäne leicht und welche Aufgaben schwer sind. Aufgabenschwierigkeiten können zwar post-hoc durch empirische Berechnungen festgestellt werden. Eine a priori Festlegung der Aufgabenschwierigkeiten ist vor allem deshalb sinnvoll, weil sie die theoretische Basis der Testentwicklung stärkt und eine empirische Bestätigung der theoretisch angenommenen Schwierigkeitsstufen im Sinne der Konstruktvalidität des Tests interpretiert werden kann.

4.2.2 Kompetenzniveaumodelle

In der berufs- und wirtschaftspädagogischen Forschung hat sich die Vorabfestlegung von schwierigkeitsbestimmenden Aufgabenmerkmalen als ein fester Bestandteil der Testentwicklung etabliert (Schumann & Eberle, 2011). Dabei soll die Frage beantwortet werden, welche kognitiven Prozesse und Ressourcen zur Bewältigung eines spezifischen Items notwendig sind (Winther, 2010). Schumann und Eberle (2011) nennen zwei zentrale Vorteile, die mit einer solchen Vorabfestlegung einhergehen. Zum einen ist die vorherige Bestimmung für die ex-post Identifikation von Kompetenzniveaus absolut notwendig. Kompetenzniveaus ordnen die numerischen Werte auf einer Kompetenzskala konkreten, fachbezogenen Kompetenzen zu und bilden Skalenabschnitte, die dann qualitativ, kriteriumsorientiert beschrieben werden (Hartig, 2007). Zum anderen wird eine theorie- und regelgeleitete Itementwicklung erleichtert und

somit die Grundlage für eine sinnvolle Interpretation empirischer Itemschwierigkeitsparameter gelegt. Der zweite Grund ist für die vorliegende Arbeit von besonderer Bedeutung. Nachdem aktuelle Forschungsbefunde zu den Strukturen kaufmännischer Kompetenzen vorgestellt wurden, wird in diesem Kapitel der theoretische Rahmen für die Einbeziehung der Aufgabenschwierigkeit in die Testentwicklung dargestellt. Ziel der Einbeziehung schwierigkeitsbestimmender Merkmale ist es, vor dem Hintergrund eines theoretisch begründeten Rahmenmodells Items zu entwickeln, die unterschiedliche Schwierigkeiten abbilden. Schwierigkeitsbestimmende Aufgabenmerkmale können sowohl in Form von formalen Aufgabenmerkmalen beschrieben werden als auch in Form von inhaltsbezogenen Aufgabenmerkmalen (Schumann & Eberle, 2011). Da die formalen Aufgabenmerkmale des zu entwickelnden Tests aufgrund des festgelegten Itemformates nur geringfügig variieren, werden im Folgenden nur die inhaltsbezogenen Aufgabenmerkmale vorgestellt und deren Bedeutung für die Testentwicklung erläutert. Eine ausführliche Auflistung inhaltsbezogener Aufgabenmerkmale kann bei Schumann und Eberle (2011, S. 79 f.) nachgelesen werden. Aus der Vielzahl von inhaltsbezogenen Aufgabenmerkmalen werden im Folgenden diejenigen vorgestellt, die sich in aktuellen Studien der Berufs- und Wirtschaftspädagogik als theoretisch und empirisch relevant etabliert haben. Dazu gehören das kognitive Anforderungsniveau, die funktionale Modellierung und die inhaltliche Komplexität (Schumann & Eberle, 2011). Diese Merkmale sind weitgehend unabhängig voneinander und sollten vom Testentwickler systematisch variiert werden (Schumann & Eberle, 2011; Winther, 2010). Ziel der Einbeziehung schwierigkeitsbestimmender Aufgabenmerkmale ist in der vorliegenden Arbeit, jedoch nicht die Bildung von Kompetenzstufen (Beaton & Allen, 1992; Hartig, 2007). Vielmehr sollen diejenigen Merkmale bei der Testkonstruktion herangezogen werden, die Itemschwierigkeiten systematisch variieren und somit eine möglichst genaue Messung der Personenfähigkeit mit unterschiedlichen Fähigkeitsniveaus ermöglichen.

Kognitive Prozesse

Die Kategorisierung der Testitems entlang der kognitiven Prozessdimension (vgl. Tabelle 1) wird häufig als Rahmen für die Itementwicklung herangezogen (Hofmeister, 2005; Schumann & Eberle, 2011; Winther, 2010). Dies geschieht zumeist unter Rückgriff auf erweiterte Taxonomien der Wissensarten nach Bloom von Anderson und Krathwohl (2001) (z. B. Marzano & Kendall, 2008). Die Taxomonie beruht auf der Überlegung, dass verschiedene Anforderungssituationen Bearbeitungsprozesse mit unterschiedlichem Beanspruchungsniveau hervorrufen können (Winther, 2010). Wie in Abschnitt 2.3.2 beschrieben, unterscheiden Anderson und Krathwohl (2001) in Anlehnung an Bloom (1972)

sechs verschiedene Anforderungsstufen. Diese sogenannten kognitiven Anforderungsdimensionen lauten: Erinnern, Verstehen, Anwenden, Analysieren, Evaluieren und Kreieren (frei übersetzt nach Anderson & Krathwohl, 2001, S. 28). Die detaillierten Beschreibungen von Anderson & Krathwohl (2001) zu den Anforderungsstufen können von Testentwicklern genutzt werden, um Aufgaben unterschiedlicher Schwierigkeit zu generieren. Leichte Aufgaben erfordern lediglich den Abruf einer Information, wie bei der Abfrage einer zuvor gelernten Definition. Schwere Aufgaben erfordern es, auf der Grundlage bestehender Informationen neue Informationen zu generieren. Mit diesen Dimensionen geht der Anspruch einher eine Schwierigkeitshierarchie von Erinnern bis Kreieren abbilden zu können. Im Rahmen wirtschaftswissenschaftlicher Wissenstests spiegeln sich diese hierarchischen Taxonomiestufen häufig nicht in empirisch nachweisbaren unterschiedlichen Lösungshäufigkeiten wieder (Lehmann & Seeber, 2007; Seeber, 2008; Witt, 2006). Aus diesem Grund greifen viele Testentwickler auf eine grobstufigere Taxonomie zurück. Schumann und Eberle (2011) operationalisierten kognitive Verarbeitungsprozesse als dreistufiges Merkmal. Auf Stufe eins werden Informationen wiedergegeben/erinnert, auf Stufe zwei werden Informationen verstanden/verarbeitet und auf Stufe drei werden Informationen genutzt. Das Kompetenzstufenmodell der ULME Studien (Lehmann & Seeber, 2007), das bei Hofmeister (2005) erklärt wird, enthält ebenfalls drei kognitive Leistungsstufen. Diese wurden nah an den Begrifflichkeiten von Anderson und Krathwohl (2001) mit als Reproduzieren, Anwenden/Verstehen und Kritisieren/Reflektieren benannt. Jedoch erwies sich diese Kategorisierung in Bezug auf die empirische Aufgabenschwierigkeit als erklärungsschwach (Lehmann & Seeber, 2007; Seeber, 2008). Neben der Abstufung des kognitiven Anspruchsniveaus wird in der Regel eine zweite Dimension in die Taxonomie eingeführt, die die Wissensart beschreibt. Winther (2010) weist in diesem Zusammenhang darauf hin, dass über die Ebene der Wissensarten eher die Dimensionen der kaufmännischen Bildung und nicht die Itemschwierigkeiten abgebildet werden. Seeber (2008) identifizierte jedoch in einem Fachleistungstest für den Ausbildungsberuf „Bürokaufmann/Bürokauffrau“ die Unterscheidung zwischen Konzeptwissen und prozeduralem Wissen als schwierigkeitsbestimmend. Darüber hinaus verdeutlichte sie, dass eine Verknüpfung verschiedener Wissensarten, in diesem Fall von Konzept- und prozeduralem Wissen, innerhalb einer Aufgabe zu erhöhten Itemschwierigkeiten führen kann. Für den Ausbildungsberuf „Kaufmann/Kauffrau im Einzelhandel“ erwies sich die Unterscheidung zwischen deklarativem und prozeduralem Wissen als schwierigkeitsbestimmend (Lehmann & Seeber, 2007). Zusätzlich wiesen sich Aufgaben, die Verstehensleistungen sowie die Erstellung eines mehrdimensionalen Handlungsplans erforderten, durch hohe Itemschwierigkeiten und geringe Streuungen aus. Diese Befunde sprechen dafür, dass Schwierig-

keiten auch durch Interkationen der Wissens- und Prozessdimension determiniert werden können (Lehmann & Seeber, 2007). Da die empirische Befundlage zur Rolle der kognitiven Prozessdimension für die Schwierigkeit von Aufgaben nicht eindeutig ist, sollten neben der kognitiven Prozessdimension noch weitere Aufgabenmerkmale bei der Testentwicklung systematisch variiert werden, um eine begründbare Verteilung der Itemschwierigkeiten zu realisieren.

Funktionale Modellierung

Das Prinzip der funktionalen Modellierung beschäftigt sich mit den notwendigen Arbeitsschritten und Modellierungsleistungen vor dem Hintergrund des Abstraktionsgrades einer Anforderungssituation (Winther, 2010). Die funktionale Modellierungsleistung in einer Anforderungssituation bezieht sich auf die Komplexität des Entscheidungsmodells, das zur Lösung der Aufgabe benötigt wird (Schumann & Eberle, 2011). Der Testentwickler kann das Ausmaß der benötigten Modellierungsleistung variieren, indem Aufgaben mit unterschiedlicher Vorstrukturierung generiert werden (Winther, 2010). Besonders gut nachvollziehbar sind diese Abstufungen bei Aufgaben, die mit mathematischen Operationen gelöst werden. Geringe Modellierungsleistungen werden benötigt, wenn alle wichtigen Kennwerte in der Aufgabe vorgegeben werden. Fortgeschrittene Modellierungsleistung ist gefordert, wenn zur Lösung einer Aufgabe mehrere Berechnungsschritte notwendig sind. Eine hilfreiche Operationalisierung dieser Anforderungsdimension ist das Auszählen der vermuteten Lösungsschritte (Schumann & Eberle, 2011).

Inhaltliche Komplexität

Die inhaltliche Komplexität einer Aufgabe ist ein weiteres schwierigkeitsbestimmendes Merkmal. Während die kognitive Taxonomierung nach angeregten kognitiven Prozessen unterscheidet und das Prinzip der funktionalen Modellierung die benötigten Lösungsschritte einbezieht, wird die inhaltliche Komplexität über den Umfang der Zusammenhänge von fachlichen Inhaltsbereichen operationalisiert (Winther, 2010). Es soll der Grad der Vernetztheit von Wissen aus unterschiedlichen Inhaltsbereichen angegeben werden, der für die Lösung einer Aufgabe notwendig ist. Die inhaltliche Komplexität kann nur unter Kenntnis der jeweiligen fachwissenschaftlichen Struktur beurteilt werden. Somit geht einer Bewertung der inhaltlichen Komplexität die Analyse der Inhaltsbereiche der Domäne voraus. Für die Darstellung der inhaltlichen Komplexität einer Aufgabe eignen sich Advanced organizer oder Concept-Maps, die grafisch abbilden, wie viele Verknüpfungen von Inhaltsbereichen für die Lösung der Aufgabe benötigt werden (Bender, 2011; Winther, 2010).

4.3 Zusammenfassung zur kaufmännischen Kompetenzmessung

Die wirtschaftspädagogische Forschung ist in ihren Überlegungen zur Strukturierung der kaufmännischen Domäne und der Modellierung kaufmännischer Kompetenzen im Vergleich zur Hochschulforschung weit vorangeschritten. Trotzdem weisen die Befunde zur Strukturierung der Domäne und den daraus entstandenen Kompetenzmodellen sowohl theoretische als auch methodische Lücken auf (vgl. Absatz 4.2.1). Befunde zu schwierigkeitsbestimmenden Aufgabenmerkmalen in der Kompetenzmessung bieten zum Beispiel noch kein eindeutiges Bild darüber, welche Aufgabeneigenschaften die Schwierigkeit einer Aufgabe maßgeblich beeinflussen und welche als randständig zu betrachten sind. Die Arbeiten von Lehmann und Seeber (2007), Seeber (2008) und Witt (2006) wiesen darauf hin, dass die Modellierung der kognitiven Anforderungen im kaufmännischen Bereich nicht einzig über die häufig verwendeten Stufen der Taxonomie von Anderson und Krathwohl (2001) oder darauf basierenden reduzierten Taxonomien vorgenommen werden kann (Hofmeister, 2005). Darüber hinaus sollte die Forschung zu Kombinationseffekten verschiedener Anforderungen und zu Interaktionseffekten zwischen Wissens- und Prozessdimension weiter ausgebaut werden. Zusätzlich zeigen Befunde aus der Forschung im Bereich der gewerblich-technischen Ausbildungsberufe, dass die Vertrautheit mit den Aufgaben auch einen Beitrag zur Aufklärung der Schwierigkeit leistet (Gschwendtner, 2008). Damit wird deutlich, dass die Aufgabenschwierigkeit nicht nur in der Aufgabe selbst, sondern darüber hinaus auch in der Person des Testteilnehmers begründet ist. Persönliche Stärken und Schwächen oder stichprobenspezifische Sozialisation im Umgang mit bestimmten Aufgaben, interagieren in Bezug auf die Schwierigkeit möglicherweise mit den Eigenschaften einer Aufgabe.

Kritisch hinzuzufügen ist, dass die Klassifikation von Aufgaben trotz der Einbeziehung von Experten und detaillierten Kodierhandbüchern von subjektiven Einschätzungen beeinflusst ist. So beschrieb Seeber (2008) beispielsweise, dass die a priori Itemkategorisierung, die von Experten des Instituts für Berufs- und Wirtschaftspädagogik der Universität Hamburg durchgeführt wurden, aufgrund der empirischen Daten noch einmal überarbeitet werden musste.

Eine deduktive Testentwicklung stellt den Königsweg sowohl im Bereich der Leistungs- als auch im Bereich der Persönlichkeitsdiagnostik dar (Hartig & Jude, 2007). Eine vollständig modellgeleitete und empirisch bestätigte Testentwicklung wurde in der Forschung zu kaufmännischen Kompetenzen noch nicht erreicht. Die berufs- und wirtschaftspädagogische Forschung hat jedoch erste entsprechende Modelle aufgestellt, die als Arbeitsgrundlage genutzt werden

können und die für die Testentwicklung in dieser Arbeit rahmengebend sind. Zwar lassen sich die Modelle nicht ohne Weiteres auf die Erfassung von betriebswirtschaftlichem Wissen an Hochschulen übertragen, sie geben jedoch wichtige Orientierungspunkte in Hinblick auf die Testentwicklung der vorliegenden Arbeit.

In einer ersten Annäherung an die betriebswirtschaftliche Domäne wird für die vorliegende Studie ein curricularer Zugang gewählt, der auf der Analyse von Studienordnungen und Modulhandbüchern basiert. Damit soll das Wissen erfasst werden, das kompetentes Handeln in späteren Anforderungssituationen ermöglicht. Um möglichst reale Anforderungssituationen aus dem späteren beruflichen Tätigkeitsfeld der Studierenden abzubilden, werden zudem Stellenanzeigen in die Testentwicklung einbezogen.

Zusammenfassend kann gesagt werden, dass die situative Einbettung von Testaufgaben, speziell im Bereich der Kompetenzmessung, eine zentrale Rolle spielt. Rost (2008) argumentiert, dass die Kontextgebundenheit der Kompetenzen es sogar notwendig macht, dass Aufgaben in einen situativen Rahmen eingebettet werden. Insbesondere dann, wenn Wissen mit prozeduralem Charakter abgefragt werden soll, ist eine situative Einbettung der Aufgaben unerlässlich, da sonst kein Transfer stattfinden kann (Hofmeister, 2005). Situative Aufgaben eigenen sich demnach gut, um berufsnahes prozedurales Wissen zu erfassen.

Empirische Vergleiche der Itemschwierigkeiten prozeduraler und deklartiver Tests fielen bislang gemischt aus. Auf der Grundlage von Daten aus der ULME III Studie waren prozedurale Aufgaben mit erhöhter Schwierigkeit verbunden. Bei Winther (2010) weisen ebenfalls die prozeduralen Items höhere Schwierigkeitsindizes auf und die geschätzten Personenparameter fallen geringer aus. Bei Winther und Achtenhagen (2009) hingegen erweisen sich die Aufgaben, die prozedurales Wissen erfassen sollen, im Vergleich zu den Items des deklarativen Tests als leichter. Möglicherweise ergeben sich die Unterschiede aus verschiedenen Arten der Operationalisierung prozeduralen Wissens. Tendenziell erwiesen sich prozedurale Items, die simulationsbasiert umgesetzt wurden als leichter (Winther, 2010), während eine papierbasierte Testung zu erhöhten Schwierigkeiten führte (Lehmann & Seeber 2007; Seeber, 2008).

Aus theoretischer Sicht kann argumentiert werden, dass situative Aufgaben eine höhere Schwierigkeit aufweisen sollten, weil im Vergleich zu nicht-situativen Aufgaben mehr Schritte zur Lösung der Aufgabe notwendig sind. Zur Lösung einer situativen Aufgabe muss die Situation analysiert werden und die relevanten Informationen aus der Situationsbeschreibung müssen erkannt und verstanden werden. Die Bearbeitung komplexer situativer Aufgaben erfordert

die Einbeziehung von Wissen über den Kontext, was eine erhöhte Vernetztheit des Wissens erfordert und mit erhöhter Aufgabenschwierigkeit assoziiert ist (Gschwendtner, 2008).

Die situative Einbettung von Aufgaben kann sowohl in papierbasierten Szenarien als auch durch computerbasierte Simulationen umgesetzt werden. Vor der Wahl einer Umsetzungsform sollten verschiedene Aspekte situativer Aufgaben kritisch betrachtet werden, denn mit dem Einsatz von situativen Aufgaben für die Erfassung von Wissen gehen sowohl Vorteile als auch Nachteile einher. Im Folgenden werden das situative Testformat und seine spezifischen Eigenschaften beschrieben. Aus der kritischen Betrachtung situativer Aufgaben werden Konsequenzen für die Testentwicklung abgeleitet.

5 Situative Messverfahren und -instrumente

Situative Testitems sind dadurch gekennzeichnet, dass die Aufgabenbeschreibung im Itemstamm in eine Situation eingebettet wird. In der Regel geht es darum, dass sich die Testperson in die beschriebene Situation hineinversetzt (Rost, 2004). Darauf aufbauend werden das Erleben und das Verhalten in dieser Situation erfragt. Als Antwortmöglichkeiten können offene Antworten gefordert oder Antwortalternativen in einem geschlossenen Format vorgegeben werden. Situative Aufgaben werden sowohl in Leistungs- als auch in Persönlichkeitstests eingesetzt. Ein Großteil der Grundlagenforschung über situative Aufgaben kommt aus dem angelsächsischen Sprachraum. Eine frühe Form der situativen Aufgaben mit offenen Antwortoptionen ist das situative Interview (Latham, Saari, Pursell & Campion, 1980). In einem situativen Interview werden die Probanden gefragt, wie sie sich in einer vom Interviewer geschilderten Situation verhalten würden. Die frei formulierten Antworten werden anhand eines Bewertungsschlüssels mit dem Fokus auf das interessierende Konstrukt ausgewertet. Die Auswertung freier Antworten ist sehr zeitintensiv und im Vergleich zu geschlossenen Antwortformaten ist die Auswertungsobjektivität meist geringer (Rost, 2004). Aus diesen Gründen werden Aufgaben für standardisierte Tests häufig mit geschlossenem Antwortformat entwickelt. Auch wenn mit offenen Aufgabenformaten spontane Reaktionen und kreative Leistung besser erfasst werden können (Rost, 2004), werden im Hinblick auf Auswertungsökonomie und -objektivität für den vorliegenden Test ausschließlich Aufgaben mit geschlossenem Antwortformat entwickelt. Neben der Wahl des Grades der Vorstrukturierung der Antworten gibt es zwei grundlegend unterschiedliche Ansätze der Umsetzung situativer Aufgaben. Die traditionelle Form der Umsetzung erfolgt papierbasiert und wird in Abschnitt 5.1 beschrieben. Moderne Ansätze nutzen Computertechnologie und betten situative Testinhalte in simulierte Unternehmensumgebungen oder Videosequenzen ein, wie in Abschnitt 5.2 beschrieben wird. Das Kapitel schließt mit einer Herausarbeitung der besonderen Eigenschaften situativer Testaufgaben (Abschnitt 5.3.2).

5.1 Papierbasierte situative Tests

Eine vornehmlich in den Vereinigten Staaten von Amerika beforschte Gruppe von papierbasierten situativen Tests mit geschlossenen Antwortformaten sind die sogenannten Situational Judgement Tests (SJTs) (Christian et al., 2010). Die Testaufgaben eines typischen SJTs bestehen aus einem situationsorientierten Aufgabenstamm und einer Vorauswahl an möglichen Antwortalternativen, die als Handlungsoptionen formuliert sind. Die Anzahl der vorgegebenen Antwor-

talternativen ist nicht festgelegt, sollte aber mindestens zwei Optionen umfassen und beläuft sich zumeist auf vier bis sechs Antwortoptionen pro Aufgabe (Weekley & Ployhart, 2006). Im Unterschied zu klassischen Multiple-Choice-Aufgaben zeichnen sich SJTs durch ihre Situierung und durch unterschiedliche Antwortszenarien aus, die eine angemessenere oder weniger angemessene bzw. nicht akzeptable Strategie des Umgangs mit der zu lösenden Aufgabe darstellen. In vielen SJTs gibt es nicht nur eine eindeutig richtige Lösung, sondern mehrere. Die Entwicklung der Items und die Punktevergabe auf die unterschiedlichen Antwortoptionen können nach verschiedenen Regeln erfolgen und werden im Folgenden beschrieben.

SJTs werden oft im Kontext der Personalauswahl oder Personalbeurteilung eingesetzt (Möller, 2010). Dementsprechend beginnt die Entwicklung eines SJTs in der Regel mit der Sammlung erfolgskritischer Situationen am Arbeitsplatz. Aus diesen Situationen werden anschließend Aufgabenstämme und Antwortoptionen abgeleitet (Möller, 2010). Die Bewertung der Effektivität und Richtigkeit der Antwortoptionen erfolgt vorzugsweise über ein Expertenrating (Pöttker, 2009). So können berufliche Handlungssituationen vorgegeben werden, die Experten (Wissenschaftler, Praktiker) hinsichtlich verschiedener Handlungsoptionen nach dem Grad der Angemessenheit oder Effizienz beurteilen. Prinzipiell gibt es drei verschiedene Methoden der Erstellung von Antwortschlüsseln für SJTs: Theoriegeleitet, empirisch und durch Berufsexperten (Pöttker, 2009). Die Instruktionen für die Beantwortung der Items durch die Testteilnehmer und die Bewertung der Ergebnisse kann ebenfalls nach unterschiedlichen Prinzipien angelegt werden. Pöttker (2009) untersuchte die Auswirkung unterschiedlicher geschlossener Antwortformate und deren Bepunktung (Scoring) auf das Antwortverhalten (Pöttker, 2009). Ein einfaches Vorgehen ist das Ankreuzen einer Antwort durch den Probanden, wobei die gewählte Antwortoption anschließend anhand eines Lösungsschlüssels als richtig oder falsch bewertet wird (Weekley, Ployhart & Holtz, 2006). Möglich ist auch die Vergabe von Punkten für teilrichtige Antworten (Masters, 1982), wenn der Proband eine Antwort wählt, die zwar nicht vollständig richtig, aber auch nicht vollständig falsch ist (Weekley et al., 2006). Weekley et al. (2006) weisen jedoch darauf hin, dass dieser recht neue Ansatz noch weiterer Forschung bedarf.

Je nachdem, ob ein Test mehr Wissensanteile oder eher Persönlichkeitseigenschaften erfassen soll, können die Instruktionen unterschiedlich formuliert werden (Pöttker, 2009). Es wird zwischen zwei Instruktionstypen unterschieden: Die erste Möglichkeit besteht darin, die Probanden zu fragen was sie in der entsprechenden Situation tun *würden*. Die zweite Möglichkeit, ist zu fragen, was die Testteilnehmer denken, was sie tun *sollten*. Empirische Untersuchun-

gen wiesen darauf hin, dass unter der ersten Instruktion gegebene Antworten eher von Aspekten der Persönlichkeit beeinflusst werden. Instruktionen, die danach fragen was getan werden *sollte*, führen hingegen zu Antworten, die stärker mit kognitiven Dispositionen wie Intelligenz korreliert waren (McDaniel, Hartman, Whetzel & Grubb, 2007). Eine Erklärung für den Befund ist, dass die Antworten auf die erste Instruktion größere Nähe zum eigenen Verhalten aufweisen, während die zweite Instruktion eher Antworten generiert, die durch die wahrgenommene Norm der Testteilnehmer beeinflusst wird.

Papierbasierte Tests weisen den großen Vorteil auf, dass sie kostengünstig in der Entwicklung und Implementierung sind und ohne technische Voraussetzungen an unterschiedlichen Standorten eingesetzt werden können. Ein Nachteil ist jedoch, dass Situationsbeschreibungen auf Papier vieler Worte bedürfen und möglicherweise die Realität nicht hinreichend abgebildet wird. Zudem setzt die ausführliche Beschreibung einer Situation ein hohes Maß an Lese- und Konzentrationsfähigkeit voraus. Für bestimmte Zielgruppen, z. B. Personen mit Leseschwäche oder Migrationshintergrund, sind papierbasierte Aufgaben entsprechend nur bedingt geeignet. Da die in dieser Arbeit anvisierte Stichprobe durch den Erwerb der Hochschulzugangsberechtigung vorselektiert ist, kann davon ausgegangen werden, dass solche Beeinträchtigungen eher eine geringfügige Rolle spielen. Trotzdem weisen computergestützte Tests dahingehend Vorteile auf, dass Situationen durch das Einbinden von Videos realistisch dargestellt werden können. Zudem ermöglichen sie interaktive Elemente, wie zum Beispiel die adaptive Anpassung der Fragen an das Leistungsniveau der teilnehmenden Person. Weitere Vorteile computerbasierten Testens werden im Folgenden beschrieben.

5.2 Videobasierte Tests und Computersimulationen

Neben dem klassischen Papierformat werden vermehrt videobasierte SJTs eingesetzt und validiert (z. B. Kanning, 2008; Möller, 2010). Studien weisen darauf hin, dass eine computergestützte Umsetzung verschiedene Vorteile hat. Dazu zählen im Speziellen Aspekte der administrativen und psychometrischen Effizienz, eine Erweiterung der inhaltlichen Spannbreite von Tests, die Erfassung neuartiger abhängiger Variablen (z. B. exakte Messung der Bearbeitungszeit) und eine erhöhte Augenscheinvalidität (Preuß & Wehrmaker, 2008). Die Augenscheinvalidität beschreibt, inwiefern die Teilnehmer glauben, dass der Test geeignet ist, das von ihnen vermutete Zielkonstrukt zu erfassen. Cattell (1974, S. 108) spricht in diesem Zusammenhang von „faith validity", weil es unabhängig von der tatsächlichen Validität darum geht, zu erfassen, was die Teilnehmenden über den Test und dessen Validität denken. Computergestütz-

ten Assessments gelingt es in der Regel besser reale Anforderungen abzubilden (z. B. durch das Einbinden von Bildern und Videos oder durch simulierte Arbeitsplätze), sodass die Augenscheinvalidität in Bezug auf spätere berufliche Tätigkeiten tendenziell als hoch eingeschätzt wird. Sofern die Testteilnehmer das Zielkonstrukt eines Tests richtig erkannt haben, kann die Augenscheinvalidität ein Hinweis darauf sein, ob der Test prognostisch valide ist.

Lievens und Sackett (2006) untersuchten, inwiefern sich videobasierte und textbasierte SJTs bezüglich ihrer prognostischen Validität voneinander unterscheiden. Über 1000 Studierende bearbeiteten jeweils eines der beiden Testformate. Die beiden Tests wurden zwischen den Jahren 1997 und 2002 im Rahmen der Aufnahmeprüfung zum Medizinstudium in Belgien administriert. Ziel des SJTs war es, interpersonelle und kommunikative Kompetenzen zu messen. Als Validierungskriterien wurden der Notendurchschnitt nach einem Jahr Studium und die Note in einem interpersonellen Training erhoben. Zudem wurden die kognitive Fähigkeit der Teilnehmer und die Augenscheinvalidität des Tests über Fragebögen erfasst. Die Korrelationen mit dem kognitiven Fähigkeitstest fielen für den videobasierten Test geringer aus als für den textbasierten Test. Dieser Befund ist möglicherweise durch Effekte der Lesekompetenz zu erklären. Das interpersonelle Kriterium konnte besser durch den videobasierten Test vorhergesagt werden. Bezüglich der Augenscheinvalidität unterschieden sich die beiden Umsetzungsformen jedoch nicht.

In einer Studie von Kanning (2008) wurden elf textbasierte SJT Items mit elf inhaltlich identischen videobasierten SJT Items hinsichtlich ihrer Wahrnehmung durch die Testteilnehmer verglichen. Es zeigt sich, dass Items, die einen videobasierten Aufgabenstamm hatten, von den Probanden tendenziell als nützlicher, gerechter (im Sinne der Testfairness), realitätsnäher und mit größerer Akzeptanz bewertet wurden. Wurden die Antwortoptionen ebenfalls als Video präsentiert, verbesserte das die subjektive Wahrnehmung zusätzlich. Das Hinzufügen einer interaktiven Komponente sollte laut Kanning (2008) unter Berücksichtigung ökonomischer Abwägungen weiter untersucht werden. Die ökonomische Abwägung spielt deshalb eine große Rolle, weil das Herstellen und Implementieren von Videosequenzen sehr zeit- und kostenintensiv ist. Zudem erfordert diese Technik, dass bei der Erhebung jeder Testteilnehmer einen Computerarbeitsplatz mit Audiozugang zur Verfügung hat, was möglicherweise den Stichprobenzugang verzerrt.

Die Problematik, situative Testsituationen zu generieren, die nicht langer textbasierter Beschreibung bedürfen und trotzdem realitätsnahen Charakter aufweisen, umgingen Winther und Achtenhagen (2009) durch die Entwicklung einer einfachen Unternehmenssimulation (vgl. Absatz 4.2.1). Die Simulation ALUSIM wurde für den Ausbildungsberuf Industriekaufmann/Industriekauffrau

entwickelt und soll ausgewählte Arbeits- und Geschäftsprozesse von Industrieunternehmen visualisieren (Winther & Achtenhagen, 2009). Die Aufgaben sollen für die Auszubildenden authentisch sein und betriebliche Kommunikations- und Handlungsabläufe widerspiegeln (Winther & Achtenhagen, 2009). Die Simulationsaufgaben, die eine prozedurale Struktur aufwiesen, wurden durch verstehensbasierte Aufgaben ergänzt. Diese Unterteilung wurde in Abschnitt 4.2.1 bereits kritisch beleuchtet. Die zentralen Überlegungen der Simulationsentwicklung werden hier jedoch dargestellt, um das Potenzial simulationsbasierter Ansätze zur Messung unterschiedlicher Wissensfacetten zu verdeutlichen. Verstehensbasierte Aufgaben erfordern die Lösung eines definierten beruflichen Problems durch Aktivierung deklarativer Wissensbestände und wurden in Papierform dargeboten (Winther, 2010). Handlungsbasierte Aufgaben und verstehensbasierte Aufgaben ließen sich in der Studie von Winther (2009) als weitgehend getrennte Dimensionen abbilden. Aus den Ergebnissen einer Studie mit 264 Auszubildenden im 3. Ausbildungsjahr zeigte sich deutlich, dass im handlungsbasierten Testteil bessere Leistungen erzielt wurden als im verstehensbasierten Teil (Winther & Achtenhagen, 2009). Winther und Achtenhagen (2009) schlussfolgerten aus ihren Ergebnissen, dass Simulationen sehr gut geeignet sind, um handlungsbasierte Kompetenzen im beruflichen Bereich abzubilden. Es muss jedoch einschränkend vermerkt werden, dass die verwendeten Simulationsaufgaben nur einen sehr engen Teilbereich kaufmännischer Tätigkeiten erfasst haben. Zudem bleibt zu klären, inwiefern die Simulation Problemlöseleistungen anstelle von fachspezifischer Kompetenz erfasst. Weitere kritische Aspekte der Studie wurden bereits in Kapitel 4 abgehandelt. Trotz der Kritik an ALUSIM ist der Vorteil von Unternehmenssimulationen, dass geprüft werden kann, inwiefern Testteilnehmer zum Beispiel mit anforderungstypischer Software umgehen können (z. B. Microsoft Office oder SAP). Jedoch müssen Unternehmenssimulationen einen hohen Detaillierungsgrad aufweisen, damit sie realitätsnahe Anforderungen abbilden können. Mit zunehmender Detaillierung geht jedoch der Nachteil einer sinkenden Generalisierbarkeit einher.

Bei einer professionellen Umsetzung sind Computersimulationen geeignet, um prozedurales Wissen in einem relativ eng umrissenen Handlungsfeld zu messen (Winther, 2010). Die hohe Augenscheinvalidität dieser Verfahren trifft in erster Linie bei technikversierten Testteilnehmern auf große Akzeptanz. Benachteiligt werden im Gegenzug diejenigen Testteilnehmer, die keine oder geringe Computerkenntnisse haben oder sich im Umgang mit Computern unsicher fühlen (Parshall, Spray, Kalohn & Davey, 2002). Die Entwicklungskosten einer komplexen Unternehmenssimulation sind als sehr hoch einzustufen. Der Aufwand der Erhebung ist ungleich höher als bei klassischen papierbasierten Verfahren. Wichtige Funktionen für die computerbasierte Testung, wie zum Beispiel die

Möglichkeit, den Internetzugriff für den Zeitraum der Testung zu unterbinden (um Internetrecherchen zur Lösung der Aufgaben zu unterbinden), ohne dabei mögliche serverbasierte Funktionen der Simulation einzuschränken, können nicht vorausgesetzt werden. Auf der anderen Seite können durch automatisierte Auswertung Ressourcen bei der Dateneingabe und Datenauswertung gespart werden. Weitgehend unabhängig von der technischen Umsetzung, weisen situative Testaufgaben besondere Eigenschaften auf, die im folgenden Abschnitt beschrieben werden.

5.3 Besondere Eigenschaften situativer Items

Im Rahmen der Forschung zu situativen Aufgaben wurde vorwiegend im englischsprachigen Forschungsraum der Frage nachgegangen, welches Konstrukt mit situativen Aufgaben gemessen wird oder gemessen werden kann. Während die berufs- und wirtschaftspädagogische Forschung die Stärke von situativen Aufgaben in der Erfassung von Kompetenzen sieht (vgl. Kapitel 4), ist die Konstruktdiskussion um situative Aufgaben im englischsprachigen Raum noch nicht abgeschlossen und wird in Abschnitt 5.3.1 nachgezeichnet. In Abschnitt 5.3.2 erfolgt eine zusammenfassende kritische Betrachtung der Vor- und Nachteile unterschiedlicher Umsetzungsformen situativer Testitems, auf deren Grundlage die Rahmenbedingungen der Testentwicklung des vorliegenden Forschungsvorhabens festgelegt werden.

5.3.1 Die Konstruktdiskussion

Mit der Entwicklung und dem Einsatz von wissenschaftlich fundierten Tests geht die Frage einher, welches Konstrukt durch die Aufgaben erfasst werden soll. In der Regel sollte der Testentwicklung ein klar umschriebenes theoretisches Konstrukt zugrunde liegen, auf dessen theoretischen Implikationen Indikatoren für das zu messende Konstrukt entwickelt werden (Bühner, 2011; Rost, 2004). In der Vergangenheit wurde bei der Entwicklung von SJTs jedoch häufig auf die Konstruktdefinition und -abgrenzung verzichtet (Christian, Edwards & Bradeley, 2010). Stattdessen wurden Aufgaben generiert, die sich nah am tatsächlichen Tätigkeitsfeld eines Bewerbers oder Testteilnehmers befinden, ohne das zu erfassende Konstrukt oder die entsprechenden Konstrukte zu benennen. Diese Theoriearmut vieler SJTs wurde vielfach kritisiert und gilt als Schwachpunkt in der Forschung zu SJTs (Christian et al., 2010; Lievens, Peeters & Schollaert, 2008; Schmitt & Chan, 2006). Fundiertes Wissen über Konstrukte, die mit SJTs erfasst werden, ist als Legitimationsgrundlage für deren weitere Nutzung in Forschung und Praxis unerlässlich (Stemler & Sternberg, 2006). Darüber hinaus ist die Konstruktdefinition von besonderer Bedeutung,

um den Transfer von SJTs für ein Konstrukt von einem Anwendungsfeld auf das nächste zu ermöglichen. Schmitt und Chan (2006) warfen die Frage auf, ob es sich bei SJTs um eine Methode oder ein Konstrukt handelt. Damit ist gemeint, ob SJTs eine Methode darstellen, mit der unterschiedliche Konstrukte gemessen werden können oder, ob SJTs ein eigenes „situational judgement"-Konstrukt messen, das sich deutlich von etablierten Konstrukten, wie Persönlichkeit oder kognitiver Fähigkeit, absetzt. Sie kamen zu dem Schluss, dass beides möglich ist. SJTs können entwickelt werden, um unterschiedliche Konstrukte zu erfassen. Schmitt und Chan (2006) weisen darauf hin, dass die spezifischen Eigenschaften des Testformats die Auswahl an erfassbaren Konstrukten einschränkt. Darüber hinaus verdeutlichten sie, dass SJTs in der Regel nicht nur ein isoliertes Konstrukt erfassen, sondern zumeist davon auszugehen ist, dass unterschiedliche Fähigkeiten bei der Lösung eines situativen Items zum Tragen kommen. Christian et al. (2010) führten, angestoßen durch die Konstruktdiskussion um SJTs, eine Metaanalyse durch. Die Analyse von 161 veröffentlichten Studien aus den Jahren von 2005 bis 2008 zeigt auf, dass 33 % der analysierten Artikel zu SJTs keine Angaben zu dem gemessenen Konstrukt machten. Von den verbleibenden Studien gab ein Großteil an, Führungskompetenzen und soziale Kompetenzen zu messen. Studien, in denen explizit berufsbezogenes Wissen erfasst wurde, wurden in diesem Zeitraum hingegen nur selten durchgeführt. Schmitt und Chan (2006) nennen exemplarisch weitere Konstruktbeschreibungen aus SJT-Studien, wie lebenslanges Lernen, Multikulturalität, Karriereorientierung, Anpassungsfähigkeit, Durchhaltevermögen und Integrität. Ein international zunehmend ins Blickfeld geratener Einsatzort für SJTs ist der medizinische und zahnmedizinische Bereich (Ahmed, Rhydderch & Matthews, 2012; Patterson, Ashworth, Mehra & Falcon, 2012). Dabei werden SJTs vorrangig genutzt, um professionelles Verhalten im Umgang mit Patienten zeitökonomisch zu erfassen (Schubert et al., 2008).

Losgelöst von den oben beschriebenen Inhalten bestehender SJTs wird die Konstruktdebatte auch auf Ebene der erfassten Wissensarten geführt. Stemler und Sternberg (2006) gehen davon aus, dass SJTs ein Maß für praktische Intelligenz darstellen. Diese praktische Intelligenz wird oft als „tacit knowledge" (stilles Wissen) (Sternberg, Wagner, Williams & Horvath, 1995) beschrieben, weil diese schwer zu verbalisieren ist. Tacit knowledge ist aus Sicht von Sternberg et al. (1995) implizites Wissen, das in der Regel über situative Aufgaben erfasst wird. Es zeichnet sich durch drei Charakteristika aus: (1) Es wird durch wenig Unterstützung von außen gelernt, (2) es ist im alltagssprachlichen Sinne nützlich, (3) es ist prozedural (Leonard & Insch, 2010). Schmitt und Hunter (1993) lehnen diese Wissensbezeichnung strikt ab und argumentieren, dass SJTs „job knowledge" erfassen und es sich somit um ein Maß genereller kognitiver Leistungsfähigkeit handelt. Stemler und Sternberg (2006) greifen den

Begriff „tacit knowledge“ auf und ordnen ihn als Unterform des prozeduralen Wissens ein. Die Rolle des prozeduralen Wissens bei der Lösung situativer Aufgaben wurde auch in aktuellen Studien vermehrt aufgegriffen (Motowidlo, Hooper & Jackson, 2006a; Motowidlo & Beier, 2010; Motowidlo, Crook, Kell & Naemi, 2009). Motowidlo et al. (2006b) argumentieren, dass prozedurales Wissen das Wissen darüber ist, wie man sich in situationsbasierten Aufgabenstellungen verhält. Zudem gehen sie davon aus, dass zur Lösung einer situativen Aufgabe auch Wissen darüber benötigt wird, unter welchen Umständen es angemessen ist, bestimmte Persönlichkeitseigenschaften bei der Lösung einer Aufgabe durchblicken zu lassen (sogenannte implicit trait theories). Damit schreiben Sie situativen Aufgaben die Funktion einer „low-fidelity simulation“ zu (Motowidlo, Dunnette & Carter, 1990). Damit ist gemeint, dass situative Aufgaben wie kleine Unternehmenssimulationen funktionieren, dabei aber im Vergleich weniger realistisch sind. Stemler und Sternberg (2006) argumentieren, dass SJTs sowohl deklarative als auch prozedurale Wissensanteile erfassen. Motowidlo et al. (2006b) bleiben eine kognitionstheoretische Untermauerung ihrer Annahmen zu situativen Aufgaben und prozeduralem Wissen schuldig. Sternberg et al. (1995) greifen hingegen die in Kapitel 2 bereits eingeführte klassische Unterteilung des Philosophen Ryle (1949) „knowing how“ und „knowing that“ auf, um die Rolle des prozeduralen Wissens bei situativen Aufgaben zu verdeutlichen. „Knowing how“ hat einen engen Bezug zu direkten Handlungen und gilt deshalb als prozedural. Prozedurales Wissen wird durch deklaratives Wissen (knowing that) kontrastiert. Darüber hinaus geben Sternberg et al. (1995, S. 917) an, dass prozedurales Wissen in sogenannten Bedingung-Aktion-Paaren repräsentiert ist. Die einfache Form dieser Wissensrepräsentation lautet:

WENN <vorhergehende Bedingung> DANN <darauf folgende Aktion>.

Als Beispiel nennen Sternberg et al. (1995, S. 917) folgende Repräsentation:

WENN <die Ampel rot ist> DANN <stopp>.

Die Ausführungen von Sternberg et al. (1995) decken sich mit den Überlegungen der kognitiven ACT-Theorien von Anderson (1983), die von einem prozeduralen und einem deklarativen Wissenssystem ausgeht und eine Repräsentation prozeduralen Wissens in Form von Problemlöseproduktionen postuliert. In der Regel sind die Spezifikationen von Präkonditionen und Aktionen, die prozedurales Wissen ausmachen, jedoch deutlich komplexer als im oben genannten Beispiel aufgezeigt. Eine solche typische Problemlöseproduktion wurde in Abschnitt 2.3.1 vorgestellt und enthält ein Ziel, eine Überprüfung der Anwendbarkeit der Regel und eine Aktion (Anderson, 2001).

Zusammenfassend kann gesagt werden, dass situative Aufgaben für unterschiedliche Inhaltsbereiche konzipiert werden können und je nach Konstrukti-

onsprinzip sowohl deklarative als auch prozedurale Wissensaspekte erfasst werden können. Die Konstruktdebatte um SJTs hat bisher kein eindeutiges theoretisches Rahmenwerk für SJTs hervorgebracht. In der Regel werden innerhalb eines situativen Items nicht nur ein Konstrukt, sondern mehrere Konstrukte abgebildet. Zum Beispiel muss die ZFA im Beispielitem 79 aus den ULME-Studien in Abbildung 3 sowohl soziale als auch fachliche Aspekte für die Lösung der Aufgabe berücksichtigen. Diese Testeigenschaft erschwert die Konstruktdefinition, birgt aber auch konkrete Vorteile, denn die Komplexität und die Handlungsnähe situativer Items ermöglichen es z. B. berufliche Kompetenzen zu erfassen (Rost, 2008). Dementsprechend sind situative Aufgaben ein fester Bestandteil der berufs- und wirtschaftspädagogischen Forschung, in der situative Aufgaben genutzt werden, um Kompetenzen im beruflichen Bereich zu erfassen (vgl. Abschnitt 4.3).

Situative Aufgaben weisen großes Potenzial für die Messung von Kompetenzen (Rost, 2008) und prozeduralem Wissen auf (Hofmeister, 2005; Stemler & Sternberg, 2006). Neben diesen Potenzialen weisen sie jedoch auch Besonderheiten auf, die einer kritischen Betrachtung bedürfen.

5.3.2 Kritische Betrachtung situativer Items

Bei der papierbasierten Testung mit situativen Items sind unterschiedliche Aspekte zu beachten. Zum einen ist zu berücksichtigen, dass die schriftliche Beschreibung komplexer Szenarien eine hohe Lesekompetenz bei den Testteilnehmern erfordert. Der mögliche Einfluss von Lesekompetenz und Textverständnis auf die Testleistung sollte insbesondere berücksichtigt werden, wenn Nicht-Muttersprachler an der Testung teilnehmen. Lange textbasierte Aufgabenstämme führen dazu, dass weniger Items in einer zeitlich begrenzten Testung vorgegeben werden können, was sich zu Ungunsten der Reliabilität auswirken kann (Rost, 2004). Das Projekt „Ökonomische Kompetenzen von Maturandinnen und Maturanden“ (OEKOMA) (Schumann et al., 2010) aus der Schweiz umging dieses Problem, indem mehrere Sub-Fragen zu einem übergeordneten Szenario gestellt wurden. Dieses Vorgehen ist jedoch aus testtheoretischer Sicht nicht unproblematisch, da es möglicherweise die lokale stochastische Unabhängigkeit der Items gefährdet (Rost, 2004). Die lokale stochastische Unabhängigkeit ist eine Voraussetzung für die Auswertung von Tests mittels Item-Response-Theorie und besagt, dass die Lösungswahrscheinlichkeit eines Items keinen Einfluss auf die Lösungswahrscheinlichkeit eines anderen Testitems haben darf (Strobl, 2012). Wird eine Szenariotechnik gewählt, so muss dringend darauf geachtet werden, dass mit der Lösung eines Items nicht die Lösung darauffolgender Items erleichtert oder erschwert wird.

Ein weiterer Kritikpunkt an typischen geschlossenen situativen Aufgaben ist, dass sich geschlossene Antwortmöglichkeiten zwangsläufig auf bestimmte Handlungsalternativen beschränken und damit eine Reihe weiterer denkbarer Handlungsoptionen ausgeschlossen werden. Die gewählten Antwortoptionen bedürfen dementsprechend einer gut fundierten Begründung. Diese Begründung sollte zum einen auf theoretischer Basis und zum anderen durch Fach- und Testentwicklungsexperten vorgenommen werden. Im Anschluss an eine Pilotierung können Distraktoranalysen herangezogen werden, um die vorgegebenen Antwortalternativen zu optimieren.

Folgt man Rost (2008), so widersprechen situative Aufgaben für Kompetenzmessung in vielerlei Hinsicht der langen Zeit gültigen Maxime für die Konstruktion von Leistungsitems. Um eine angemessene Testreliabilität zu erzielen, muss ein Test möglichst viele homogene Items aufweisen, damit die Eindimensionalität des Tests nachgewiesen werden kann (Rost, 2004). Wenn Kompetenzen jedoch in ihrer Komplexität und in authentischen Kontexten erfasst werden sollen, dann sind die Aufgaben nicht homogen, da unterschiedliche Fähigkeiten bei der Lösung einer Aufgabe zusammenspielen. Dieser Problematik ist geschuldet, dass viele situative Tests eine nach üblichen Maßstäben nur unbefriedigende Reliabilität aufweisen (z. B. Abele et al., 2012; Schmitt & Chan, 2006).

Abele und Nickolaus (2013) entwickelten im beruflichen Bereich neue Lösungsansätze zur reliablen und validen Erfassung von Kompetenzen mit komplexen und problemhaltigen berufsfachlichen Aufgaben. Um die Reliabilität zu steigern, ohne die Testzeit übermäßig zu verlängern, versuchten sie Teilleistungen aus komplexen Problemlöseprozessen zu berücksichtigen und kleinschrittig über mehrere Items abzufragen. Dafür wurden komplexe Aufgaben von Kraftfahrzeug-Mechatronikern in weniger komplexe Teilaufgaben untergliedert und auf ihre psychometrischen Eigenschaften geprüft. Wie intendiert, konnte eine deutliche Reliabilitätssteigerung im Vergleich zu den komplexen Aufgaben erreicht werden (Abele & Nickolaus, 2013). Es bleibt jedoch zu klären, inwiefern eine Zerlegung des Lösungsprozesses in Teilschritte die Qualität des Lösungsprozesses beeinflusst (Abele & Nickolaus, 2013) und inwiefern die lokale stochastische Unabhängigkeit (vgl. Abschnitt 7.1.2) der Teilitems gewährleistet bleibt. Das Potenzial einer Unterteilung komplexer Aufgaben in weniger komplexe Teilschritte liegt neben der Reliabilitätssteigerung in der Möglichkeit die Items unterschiedlichen Dimensionen zuzuordnen. Ein ursprünglich eindimensional modelliertes heterogenes Konstrukt, ließe sich so möglicherweise durch mehrere Subdimensionen präziser abbilden.

Im Zusammenhang mit heterogenen Konstrukten hat sich die Nutzung mehrdimensionaler Modelle bewährt (Hartig & Höhler, 2008). Dabei wird berücksich-

tigt, dass nicht alle Items nur von einem latenten Faktor beeinflusst werden, sondern dass es möglich ist, innerhalb eines Modells Parameter für mehrere Faktoren zu schätzen. Bei der Modellierung von mehrdimensionalen Modellen gibt es zwei Möglichkeiten: (1) die Faktoren werden so modelliert, dass sie jeweils nur auf einen der mindestens zwei Faktoren laden („between-item multidimensionality") und (2) die Faktoren werden so modelliert, dass ein Item auf mehrere Faktoren laden kann („within-item multidimensionality") (Hartig & Höhler, 2008). Eine Modellierung situativer Aufgaben über multidimensionale Modelle scheint in Anbetracht der oben beschriebenen Herausforderungen ein vielversprechender Ansatz. Jedoch setzt eine solche Modellierung voraus, dass begründete Annahmen darüber bestehen, welches Item von welchen latenten Dimensionen beeinflusst wird. Solche Modelle liegen jedoch zum jetzigen Zeitpunkt der Forschung zur Messung betriebswirtschaftlichen Wissens an Hochschulen noch nicht vor. Stattdessen wird der Versuch unternommen, die Items so zu gestalten, dass sie neben dem anvisierten betriebswirtschaftlichen Wissen auf Bachelorniveau möglichst keine weiteren Dimensionen (wie zum Beispiel bestimmte Persönlichkeitseigenschaften oder Sozialkompetenzen) erfassen. Durch dieses Vorgehen büßen die situativen Aufgaben an realitätsnaher Komplexität ein, versprechen aber psychometrische Eigenschaften, die mit dem Rasch-Modell kompatibel sind.

Die kritischen Überlegungen zur Entwicklung und Verwendung von situativen Items werden im folgenden Kapitel aufgegriffen und es werden daraus Implikationen für die vorliegende Arbeit abgeleitet. Anschließend werden die Anforderungen, die der Test abbilden soll, beschrieben und das Rahmenmodell der Testentwicklung daraus abgeleitet.

6 Entwicklung eines situativen betriebswirtschaftlichen Wissenstests

In den vorangegangenen Kapiteln wurden zentrale Befunde der Kompetenzmessung und Kompetenzmodellierung aus dem Bereich der berufs- und wirtschaftspädagogischen Forschung vorgestellt. Zudem wurden die Besonderheiten situativer Aufgaben im Rahmen diagnostischer Aufgabenstellungen herausgearbeitet. Im folgenden Abschnitt werden aus den bisherigen Ausführungen die Implikationen für die Entwicklung eines situativen betriebswirtschaftlichen Wissenstests abgeleitet. Das Kapitel beginnt in Abschnitt 6.1 mit der Beschreibung der methodischen Grundlagen der Testentwicklung. In Abschnitt 6.2 werden die Anforderungen des Lernens und Arbeitens an Studierende betriebswirtschaftlicher Studiengänge herausgearbeitet. Aus diesen Beschreibungen folgt das Rahmenmodell der Itementwicklung in Abschnitt 6.3, in dem die entwickelten Items und ihre spezifischen Charakteristika systematisch beschrieben werden. Das Kapitel schließt mit einer Zusammenfassung der Testentwicklung in Abschnitt 6.4.

6.1 Methodische Grundlagen der Testentwicklung

Die Entwicklung eines wissenschaftlichen Tests sollte nicht ohne die Berücksichtigung nachvollziehbarer Regeln vorgenommen werden. Modelle der Testentwicklung geben dabei eine Orientierung darüber, welche Gestaltungskriterien einer wissenschaftlichen Testentwicklung zugrunde liegen sollten. Es geht insbesondere darum, den Kausalitätsschluss zwischen dem beobachteten Verhalten (z. B. der Antwort auf eine Frage im Test) und der Bewertung und Interpretation dieses Verhaltens im Hinblick auf die zu messende Fähigkeit abzusichern. In der Literatur liegen verschiedene Modellüberlegungen zum Testentwicklungsprozess vor, die sich durch ihr jeweiliges Vokabular und leicht abweichende Schwerpunktsetzungen unterscheiden, jedoch im Kern die gleichen Abläufe bei der Testentwicklung implizieren. Drei dieser Modelle werden im Folgenden beschrieben.

6.1.1 Modelle der Testentwicklung

Pellegrino, Chudowsky und Glaser (2001, S. 44) stellten ihre Grundüberlegungen einer modellgeleiteten Testentwicklung in Form eines Dreiecks als „Assessment Triangle" dar. Die Eckpunkte des Dreiecks werden mit den Begriffen „Kognition", „Observation" und „Interpretation" benannt. Unter dem Begriff Kognition ist das Modell zusammengefasst, das der Testentwickler vom Lernen

und Leisten im Rahmen des Assessments hat. Dabei geht es um Unterrichtsinhalte, fachdidaktische Inhalte und methodische Aspekte in Lehr-Lernprozessen (Winther, 2010, S. 60). Unter dem Eckpunkt der „Observation" wird beschrieben, wie die zuvor angesprochenen Kognitionen gemessen werden können. Der Testentwickler soll dabei Aufgaben unter Bezugnahme auf das zuvor angenommene kognitive Modell entwickeln (Winther, 2010). Die Interpretation beschreibt die Methode, die genutzt wird, um von den Beobachtungen im Test auf die Kognitionen des Testteilnehmers rückzuschließen.

Während das „Assessment Triangle" (Pellegrino et al., 2001) auf relativ abstrakter Ebene das Zusammenspiel von kognitivem Modell, Beobachtung und Interpretation im Rahmen von Assessments beschreibt, differenziert das Evidence-Centered Design (ECD) (Mislevy & Risconscente, 2005) einzelne Schritte der Testentwicklung weiter aus. Im Vergleich zum „Assessment Triangle" (Pellegrino et al., 2001) wird im Rahmen des ECD die Wichtigkeit der empirischen Testvalidierung herausgestellt. Die Autoren sprechen in diesem Zusammenhang von „Schichten" eines Assessments, die im Folgenden als „Schritte" interpretiert werden. Die Autoren stellen fünf Schritte der Testentwicklung heraus: (1) Modellieren der Domäne, (2) Analysieren der Domäne, (3) Entwickeln eines Rahmens der Assessmentkonstruktion, (4) Umsetzung des Assessments und (5) Einsatz des Assessments. Jeder dieser Schritte wird über seine Funktion, Schlüsselmerkmale und beispielhafte Repräsentationsformen beschrieben (Mislevy & Risconscente, 2005, S. 6). Insbesondere das „Conceptual Assessment Framework" als Bestandteil des dritten Entwicklungsschritts wird ausführlich beschrieben. An dieser Stelle der Testentwicklung findet die Operationalisierung des zuvor definierten Messgegenstands statt. Ähnlich wie bei dem „Assessment Triangle" basiert das „Conceptual Assessment Framework" auf einem „Student Model", das Angaben darüber macht, was bei einer Person mit welchen Ausprägungen gemessen werden soll. Zwischen diesem „Student Model" und dem „Task Model", das beschreibt, wie genau die Items der Messung aussehen sollen, liegt das „Evidence Model", in dem die Umstände der Messung festgelegt werden (z. B. darüber, wie Punkte im Test vergeben und verrechnet werden). Das sogenannte „Assembly Model" macht zudem Angaben darüber, wie viele Beobachtungen vom Testteilnehmer benötigt werden, um das Ziel des Assessments zu erfüllen.

Einer ähnlichen Systematik bei der Beschreibung des Testentwicklungsprozesses folgt Wilson (2005) mit seinem Konzept der „Four Building Blocks". Wilson (2005) beschreibt den Testentwicklungsprozess in vier Schritten. Zu Beginn jeder Testentwicklung steht die sogenannte „Construct Map". In dieser ersten Testentwicklungsphase wird das Konstrukt spezifiziert und die Metrik, auf der das Konstrukt gemessen werden soll, verankert. Im nächsten Schritt wird das

„Item Design" an die Vorgaben der „Construct Map" angepasst. Anschließend wird der „Outcome Space" definiert. Damit ist gemeint, dass Regeln der Bepunktung im Einklang mit der „Construct Map" und dem „Item Design" festgelegt und definiert werden. In Abhängigkeit des „Outcome Space" wird das „Measurement Model" gewählt. Zwischen dem Konstrukt und den Items wird ein Kausalzusammenhang impliziert. Nur wenn dieser Kausalzusammenhang besteht, können auf der Grundlage der Testergebnisse über das Messmodell durch einen Inferenzschluss Aussagen über das zuvor definierte Konstrukt getroffen werden.

Das theoretische Modell der „Four Building Blocks" lässt sich in einem Zyklus der Instrumentenentwicklung abbilden, der in Abbildung 6 dargestellt ist.

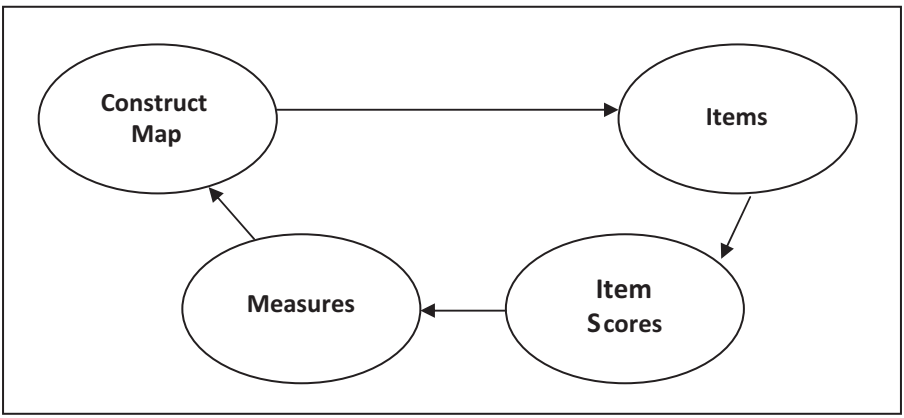


Abb. 6: Der Instrumentenentwicklungszyklus entlang der „Four Building Blocks"

(Quelle: Wilson, 2005, S. 19)

Zusammenfassend stellen alle Modelle die Wichtigkeit einer systematischen Testentwicklung heraus. Darüber hinaus wird deutlich, dass die Entwicklung von Aufgaben nicht ohne eine genaue Analyse und Definition des Zielkonstrukts vorgenommen werden kann. Erst wenn die Zielsetzung der Messung deutlich ist und der Gegenstand der Messung systematisch eingegrenzt und beschrieben wurde, können die Items auf das zuvor definierte Konstrukt abgestimmt werden. Ebenso muss die Bepunktung der Items im Einklang mit dem Messmodell gewählt werden, damit durch das Messmodell Rückschlüsse auf das anfangs definierte Konstrukt möglich werden. Die beschriebenen Testentwicklungsschritte decken sich mit den Angaben der Standards pädagogisch-psychologischer Diagnostik zum Testentwicklungsprozess (AERA, APA &

NCME, 2002). Dort wird die Testentwicklung in folgende vier Phasen unterteilt (frei übersetzt nach APA & NCME, 2002, S. 37):

(1) Darlegung des Ziels des Assessments und Darstellung der Domäne in Form von Inhalten und Fähigkeiten, die gemessen werden sollen.

(2) Entwicklung und Überprüfung von Testspezifikationen.

(3) Testentwicklung, Pilotierung, Bewertung und Selektion der Items und des Auswertungsprozesses.

(4) Zusammenstellung und Bewertung des Tests für die tatsächliche Nutzung.

Im Vergleich zu der globalen Beschreibung der Testentwicklung in den Standards pädagogisch-psychologischer Diagnostik werden in den Assessment-Modellen die einzelnen Entwicklungsschritte differenziert dargestellt und auf die besonderen Anforderungen von Assessments in Bildungskontexten bezogen. Die im Folgenden beschriebene Testentwicklung integriert die bildungswissenschaftlichen Modelle mit den Standards pädagogisch-psychologischer Diagnostik (vgl. Kapitel 7). Die Grundlage jeder Testentwicklung ist die Definition und Umschreibung des Zielkonstrukts. Diese wird im Folgenden vorgenommen.

6.1.2 Definition des Zielkonstrukts

Wissen wird als elementarer Bestandteil der beruflichen Handlungskompetenz in betriebswirtschaftlichen Berufen betrachtet (vgl. 2.2). Insbesondere werden das Wissen über Methoden und Prozeduren in betriebswirtschaftlichen Handlungsfeldern und das Wissen darüber, in welchen Situationen diese angewendet werden sollen, in Anlehnung an Anderson und Krathwohl (2001) als Zielkonstrukt anvisiert und dessen Einsatz in unterschiedlichen Situationen geprüft. Im Sinne von Wittmann, Süß und Oberauer (1996) handelt es sich dabei um Wissen über Prozeduren, die aus dem deklarativen Gedächtnissystem abgerufen werden. Das Konstrukt soll jedoch so handlungsnah wie möglich und durch die Aktivierung prozeduraler Wissensbestände abgebildet werden. Das Zielkonstrukt ist auf einem latenten Kontinuum verortet, das über eine geringe Ausprägung des Konstrukts bis zu einer hohen Ausprägung reicht. Thematisch beschränkt sich das zu testende Merkmal auf Inhalte des Bachelorstudiums. Für die Eingrenzung und Strukturierung der Domäne wird in Anlehnung an Schumann et al. (2010) und Bothe (2003) ein curricular geprägter Zugang gewählt. Die curricular strukturierten Anforderungen des Lernens an deutschen Universitäten werden im Abschnitt 6.2.1 beschrieben. Da der Test jedoch die Anwendung von Wissen auf beruflich relevante Situationen erfassen soll, werden neben den curricularen Anforderungen auch Anforderun-

gen zukünftiger Arbeitsplätze in die Testentwicklung miteinbezogen (eine genaue Beschreibung des Vorgehens erfolgt in Abschnitt 6.2.2). Zum Verhältnis von curricularen und beruflichen Anforderungen an Studierende ist aus wissenschaftlicher Sicht wenig bekannt. Eine Diskussion des möglichen Spannungsfeldes zwischen universitären und späteren beruflichen Anforderungen im Bereich der Betriebswirtschaftslehre erfolgt in Kapitel 10.

Die Konstruktspezifizierung erfolgt über die Eingrenzung und Systematisierung der Domäne im nachstehenden Abschnitt.

6.2 Anforderungen des Lernens und Arbeitens

Im Vergleich zur beruflichen Bildung ist der tertiäre Bildungssektor durch eine ausgeprägte Anforderungsheterogenität gekennzeichnet. Diese Heterogenität betrifft zum einen die Bedingungen des Lernens innerhalb der Universitäten und zum anderen die vielschichtigen Einmündungsmöglichkeiten von Hochschulabsolventen in den Arbeitsmarkt (Henning & Henning, 2009). Sie erschwert im Hinblick auf die Entwicklung von Tests im Hochschulbereich ein allgemeingültiges Domänenverständnis und somit auch die systematische Analyse domänenspezifischer Anforderungen des Lernens und Arbeitens. Nach welchen Kriterien Anforderungen klassifiziert und Aufgabenbündel zu Teilkompetenzen geschnürt werden, obliegt theoretisch begründeten Intentionen der Forschenden (Rosendahl & Straka, 2011), die im Folgenden dargelegt werden. Bevor im folgenden Abschnitt die universitären Anforderungen an Studierende betriebswirtschaftlicher Fächer verdeutlicht werden, soll kurz die Rolle wirtschaftspädagogischer Studiengänge in diesem Zusammenhang erläutert werden. Alle Angaben zum Studiengang beziehen sich auf Veröffentlichungen auf der Webseite der Universität Göttingen (2013). Das Studium der Wirtschaftspädagogik kombiniert wirtschaftswissenschaftliche und bildungspolitische Inhalte mit organisatorischen und didaktisch-methodischen Fragen der beruflichen Aus- und Weiterbildung. Dabei wird sowohl der schulische als auch der betriebliche und überbetriebliche Bereich bertachtet. Durch einen erfolgreichen Masterabschluss befähigt das Studium zum Eintritt in den Vorbereitungsdienst für das Lehramt an kaufmännischen berufsbildenden Schulen. Darüber hinaus gehört eine studienbezogene und arbeitsmarktbezogene Polyvalenz der Bachelor- und Masterabschlüsse an vielen Studienstandorten zu den Leitideen des Studiengangs (Zlatkin-Troitschanskaia & Breuer, 2010). Polyvalenz beschreibt den Umstand, dass Studierende der Wirtschaftspädagogik z. B. in Studiengänge des Fachs Wirtschaftswissenschaften wechseln können und auf dem Arbeitsmarkt dazu befähigt sein sollen, neben dem Lehramt an beruflichen Schulen in andere betriebswirtschaftliche Berufe einzumünden. Der An-

spruch auf Polyvalenz ist speziell für den Bachelorabschluss von Bedeutung, da dieser berufsqualifizierend sein soll (Dörfler, 2005), der Bachelorabschluss allein jedoch nicht ausreicht, um für den Vorbereitungsdienst an kaufmännischen beruflichen Schulen zugelassen zu werden. Entsprechend werden die Anforderungen an Studierende der Wirtschaftspädagogik und an Studierende der Betriebswirtschaftslehre im Folgenden unter der Perspektive betriebswirtschaftlicher Tätigkeiten außerhalb des Schuldienstes betrachtet.

6.2.1 Universitäre Anforderungen des Lernens

„Die Betriebswirtschaftslehre befasst sich mit der Analyse, Gestaltung, Führung eines Unternehmens sowie mit der wirtschaftlichen Entwicklung." (BLK[7] & BA[8], 2004, S. 246; zit. nach Ramm & Multrus, 2006, S. 17). Dabei werden betriebliche Vorgänge in einzelnen Unternehmen, aber auch zwischen verschiedenen Unternehmen betrachtet. Diese Fokussierung auf die Unternehmensebene ist das zentrale Unterscheidungsmerkmal zur Volkswirtschaftslehre (VWL), die vorwiegend wirtschaftliche Beziehungen auf staatlicher oder zwischenstaatlicher Ebene thematisiert (Ramm & Multrus, 2006).

Das Bachelorstudium der Betriebswirtschaftslehre in akkreditierten Studiengängen in Deutschland folgt standortübergreifend einer ähnlichen Systematik (Henning & Henning, 2009). Henning und Henning (2009, S. 80 f.) beschreiben folgende Studienstruktur: Zu den verpflichtenden fachwissenschaftlichen Modulen gehören Module zu Buchführung und Jahresabschluss inklusive Kostenrechnung, Informationsverarbeitung, Wirtschaftsrecht, Mikroökonomik und Makroökonomik. Grundlegende, verpflichtende BWL-Module sind in der Regel: Marketing, Produktion und Logistik, Finanzwirtschaft und Rechnungswesen. In den Wahl- und Spezialisierungsmodulen werden Themenbereiche wie Finanzmanagement, Investition und Unternehmensbewertung, Controlling, Kostenrechnungssysteme, Bilanzpolitik und -analyse, Grundlagen der Wirtschaftsinformatik, Grundlagen der Unternehmensbesteuerung, Grundlagen der Organisation und Grundlagen des internationalen Managements gelehrt. Zudem führen Henning und Henning (2009) Fremdsprachen, Schlüsselqualifikationen, verpflichtende Praktika und die abschließende Bachelorarbeit als Bestandteile des betriebswirtschaftlichen Bachelorstudiums auf.

Neben den inhaltlichen Anforderungen ist es bei der Testentwicklung wünschenswert, auch das kognitive Anforderungsniveau in Betracht zu ziehen, auf welchem die Inhalte des Studiums erlernt werden sollen. Die kurzen Beschreibungen der Modulhandbücher einzelner Universitäten (vgl. Anhang A) ermöglichen keine Ableitung darüber, welche kognitiven Anforderungsniveaus als

7 Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung
8 Bundesagentur für Arbeit

Lernziel vermittelt werden sollen. Einen empirischen Zugang zur Beantwortung der Frage, welche Wissensarten auf welchen Niveaustufen erlernt werden sollen, bietet die Analyse von Lehrmaterialen und Klausuren. Da jedoch Prüfer innerhalb eines Moduls häufig wechseln und somit auch die Klausuren und deren Anforderungsniveaus Fluktuationen unterliegen, ist eine solche Analyse mit methodischen Schwierigkeiten behaftet. Als Alternative zu Analysen der Lehrmaterialien werden am Endes dieses Abschnittes durch Experten generierte normative Lernziele und die damit verbundenen kognitiven Anforderungsniveaus dargestellt und für die Testentwicklung aufgegriffen (vgl. Tabelle 7).

Im Bereich der betriebswirtschaftlichen Studiengänge liegen keine verbindlichen Rahmencurricula vor, die die Studienstruktur vorgeben. Deshalb muss bei der Betrachtung der Domäne aus hochschulcurricularer Perspektive auf Studienordnungen und Modulbeschreibungen zurückgegriffen werden, die von den jeweiligen Universitäten selbst erstellt und zumeist online veröffentlicht werden (z. B. Universität Göttingen, 2010). In Anlehnung an das Domänenmodell kaufmännischer Bildung soll der Test situative Aufgaben enthalten, die sowohl Steuerungs- und Unterstützungsprozesse als auch Wertschöpfungsprozesse innerhalb einer Unternehmung abbilden (Winther, 2010). Darüber hinaus müssen die Aufgaben mit den deutschen Hochschulcurricula in den Studienfächern Wirtschaftspädagogik und Betriebswirtschaftslehre kompatibel sein. Das heißt, die innerhalb der situativen Aufgaben eingebundenen Fachinhalte und Anforderungen müssen sich mit dem Lehrangebot an deutschen Universitäten decken. Für die Eingrenzung dieser Inhaltsbereiche gilt es, diejenigen Module zu identifizierten, die studienstandortübergreifend verpflichtend sind. Da der Test sowohl für Studierende der Betriebswirtschaftslehre als auch für Studierende der Wirtschaftspädagogik curriculare Gültigkeit aufweisen soll, dürfen nur Inhalte aus Modulen in den Test aufgenommen werden, die in beiden Studiengängen verpflichtend sind.

Vor diesem Hintergrund wurde im Jahr 2010 am Lehrstuhl für Wirtschaftspädagogik und Personalentwicklung an der Georg-August-Universität eine Analyse der Studienordnungen aller Wirtschaftspädagogikstudiengänge in Deutschland vorgenommen und durch Analysen des Bachelorstudiengangs Betriebswirtschaftslehre ausgewählter Studienstandorte (Göttingen, Hamburg, Mannheim) ergänzt. Die Analysen bestätigten die von Henning und Henning (2009) dokumentierte Studienstruktur in den Grundzügen. Die im Bachelor Wirtschaftspädagogik und im Bachelor Betriebswirtschaftslehre zu absolvierenden Pflichtmodule belaufen sich auf:

- Statistik und Mathematik,
- Mikroökonomie,

- Makroökonomie,
- Einführung in die VWL,
- Recht,
- Finanzwirtschaft,
- Rechnungswesen,
- Marketing und Absatz,
- Produktion (mit Logistikanteilen),
- und Unternehmensführung.

Eine Abweichung zu den Angaben von Henning und Henning (2009) liegt bezüglich des Moduls Unternehmensführung, Organisation und Management vor, das an der Mehrzahl der Standorte zu den Pflichtmodulen gehört und nicht zu den Wahlpflichtmodulen. Der Themenbereich der Wirtschafsinformatik ist in den Bachelorstudiengängen der Betriebswirtschaftslehre weitgehend als Pflichtmodul etabliert, in den wirtschaftspädagogischen Studiengängen jedoch an einigen Standorten nicht und bleibt deshalb bei der Testentwicklung unberücksichtigt.

Mathematische, volkswirtschaftliche und rechtliche Grundlagen sollen explizit nicht Gegenstand des Tests sein. Auf Grundlage eigener Analysen der Modulhandbücher und Abgleiche mit bestehenden Arbeiten zu curricularen Analysen (Bothe, 2003; Schumann et al., 2010) wurden folgende Bereiche als großer inhaltlicher Überscheidungsbereich zwischen verschiedenen Standorten und den Bachelorstudiengängen Wirtschaftspädagogik und Betriebswirtschaftslehre identifiziert:

- Finanz- und Rechnungswesen,
- Marketing und Absatz,
- Produktion (mit Logistikanteilen),
- und Unternehmensführung.

Zu der Analyse der Bachelormodule ist anzumerken, dass (1) die Module mit ähnlichen Inhalten nicht immer den gleichen Namen tragen und (2) Module, die den gleichen Namen tragen, teilweise stark abweichende Beschreibungen der Modulinhalte aufweisen (vgl. Anhang A). Trotz dieser Schwierigkeiten lässt sich eine gemeinsame inhaltliche Schnittmenge identifizieren, und die Modulbeschreibungen der Kernfächer geben den inhaltlichen Rahmen der Testitems

vor. Ein Überblick über die Modulbeschreibungen von drei für die Testentwicklung ausgewählten Universitäten ist im Anhang A dokumentiert.

Neben zentralen Inhalten eines Bachelorstudiums sollten bei der Testentwicklung die gängigen Prüfungsformen an der Universität in Betracht gezogen werden (Winther, 2010). In diesem Rahmen sollte das kognitive Anforderungsniveau betrachtet werden, auf dem die Inhalte des Studiums abgefragt werden. Wie bereits erwähnt, ist dieses im Hochschulbereich nicht unproblematisch, denn Studienleistungen werden in Deutschland im Regelfall seitens der jeweiligen Hochschule zertifiziert (Teichler, 2007). In der Praxis ergibt sich daraus eine Vielfalt an Methoden der studentischen Leistungsüberprüfung. Die Dozierenden der einzelnen Fächer nutzen zumeist die große Entscheidungsfreiheit, wie sie die obligatorische Leistungsprüfung am Ende des Semesters vornehmen wollen. In der Regel handelt es sich dabei um Klausuren mit offenen oder geschlossenen Antwortformaten, 10 bis 20-seitige Hausarbeiten oder es werden mündliche Abschlussprüfungen durchgeführt. Hinsichtlich der Erfassung von handlungsnahen Lernzielen, wie sie in Tabelle 7 aufgelistet sind, bestehen jedoch noch große Defizite. Zwar werden in den Modulbeschreibungen der einzelnen Veranstaltungen in einigen Fällen Kompetenzen und handlungsnahe Lernziele schriftlich formuliert und veröffentlicht (Stratmann, Preussler & Kerres, 2009), konkrete Umsetzungen in den Lehrveranstaltungen finden jedoch häufig nicht statt. Stratmann et al. (2009) kritisierten, dass zumeist Erinnerungsleistungen von den Studierenden gefordert werden. Diese Art der Erfassung von Lernergebnissen reicht aber für die durch die Kompetenzorientierung geforderten Rückschlüsse auf Anwendung des Wissens nicht aus. Herkömmliche Prüfungsformen, wie zum Beispiel Klausuren und Aufsätze, stoßen für die Kompetenzerfassung laut Stratmann et al. (2009) an ihre Grenzen, weil diese Leistungen nur punktuell erfassen. Eine ausführliche Diskussion der Passung von bildungspolitischen und curricularen Zielvorgaben mit der Umsetzung von Lehre und Assessments an deutschen Universitäten erfolgt in Abschnitt 10.2.

Im Rahmen der Testentwicklung wurde geprüft, ob die Beschreibungen der Lernziele in Modulhandbüchern herangezogen werden können, um Anforderungsniveaus abzuleiten, auf denen die Inhalte gelernt werden sollen. Es hat sich jedoch herausgestellt, dass die Beschreibung der Lernziele nicht systematisch erfolgt und diese sowohl zwischen den Modulen als auch zwischen den Universitäten stark in ihrer Ausarbeitung variieren, sodass sie nicht auswertbar waren. Eine Analyse der veröffentlichten Lernziele und Prüfungsformen hat sich somit als wenig fruchtbar erwiesen. An dieser Stelle wäre zum Beispiel eine Delphi-Untersuchung (Linstone & Turoff, 1975) mit Fachexperten deutscher Universitäten eine mögliche Methode, einen Konsens über grundlegende Lernziele betriebswirtschaftlicher Studiengänge zu erzielen. Auf europäischer Ebe-

ne wurde ein solcher Ansatz bereits verfolgt, der im Folgenden vorgestellt wird.

In der AHELO-Studie der OECD wurde in einem Delphi-ähnlichen Verfahren ein solcher Konsens für den Bereich der volkswirtschaftlichen Hochschulbildung (OECD, 2011) erarbeitet. Das in Abschnitt 3.4.2 beschriebene AHELO-Projekt der OECD entwickelte unter Mitwirkung von Experten aus der universitären Forschung eine Auflistung an Lernzielen, die Studierende im Bereich der Wirtschaftswissenschaften während eines Bachelorstudiums erreichen sollten. Die Ausarbeitung der Lernziele bezieht sich auf den Bereich der Wirtschaftswissenschaften mit starkem Fokus auf volkswirtschaftliche Inhalte (economics). In Tabelle 7 werden die Lernziele der OECD frei übersetzt und angepasst an den betriebswirtschaftlichen Kontext dargestellt.

Tab. 7: Lernziele für Studierende der Wirtschaftswissenschaften (frei übersetzt und adaptiert nach OECD, 2011 S. 28–29)

Lernziel I: Studierende sollten in der Lage sein, Fachwissen und Fachverständnis zu demonstrieren.
Fachwissen und Fachverständnis können über folgende Anforderungen erfasst werden: • konsistenter und kohärenter Gebrauch betriebswirtschaftlicher Fachsprache • klare Definitionen von Standardbegriffen der Betriebswirtschaft • Erklärung grundlegender Konzepte der Betriebswirtschaft unter Berücksichtigung bekannter Kontroversen • Bekanntheit betriebswirtschaftlicher Prinzipien • Erklären, wie Marktteilnehmer zu Entscheidungen kommen und betriebswirtschaftliche Entscheidungsprobleme lösen können
Lernziel II: Studierende sollten Fachwissen und dessen Anwendung auf reale Problemstellungen demonstrieren.
Fachwissen und dessen Anwendung können über folgende Anforderungen erfasst werden: • effektive Anwendung von betriebswirtschaftlichem Schlussfolgern und Analysemethoden auf spezifische Anwendungsgebiete • Erkennen von Annahmen und ihren Implikationen für analytische Ergebnisse und wirtschaftliche Debatten • Nutzung betriebswirtschaftlichen Schlussfolgerns, um beratend tätig zu werden • die Konzepte aus Lernziel I sollen genutzt werden, um betriebswirtschaftliche Fragestellungen zu bewerten
Lernziel III: Studierende sollten demonstrieren, dass sie effektiv von relevanten Daten und quantitativen Methoden Gebrauch machen können.
Die Fähigkeit, effektiv von relevanten Daten und quantitativen Methoden Gebrauch machen können, sollte in folgenden Bereichen demonstriert werden: • Zugang zu relevanten Daten • Methoden der Datenanalyse

(Fortsetzung Tab. 7)

• Interpretation und Limitationen empirischer Daten
Lernziel IV: Studierende sollten demonstrieren, dass sie in der Lage sind, mit Fachleuten und Nicht-Fachleuten zu kommunizieren.
Die Fähigkeit, mit Spezialisten und Nicht-Spezialisten zu kommunizieren, sollte in folgenden Bereichen demonstriert werden: • Teilen von Informationen, Problemen und Lösungen • Gebrauch von Analysetools, wie Tabellen, Diagrammen, Graphen, Modelle usw., um mit den Zuhörern zu kommunizieren • Präsentation von quantitativen Informationen in nützlicher Form • Zusammenfassung von Daten, die in Rohdatenform zu komplex sind • Erklären von Ergebnissen für Spezialisten und Nicht-Spezialisten
Lernziel V: Studierende sollten demonstrieren, dass sie in der Lage sind, selbstregulativ zu lernen.
• Abstraktion, Analyse, Deduktion, Induktion, Quantifikation und Design, Framing

Trotz dieser abweichenden inhaltlichen Ausrichtung werden die von der OECD veröffentlichten Lernziele als richtungweisend angesehen, um Anforderungen an Studierende der Betriebswirtschaftslehre zu subsummieren.

Die Lernziele aus dem Bereich I werden weitgehend durch den bereits bestehenden Wissenstest (BAKT) erfasst (vgl. 3.1.2), der darauf abzielt, deklarative Wissensbestände zu erfassen und zum Beispiel Definitionen von betriebswirtschaftlichen Begriffen abfragt. Ziel des zu entwickelnden Tests soll es dementsprechend sein, Aufgaben zu generieren, die Fähigkeiten aus dem Bereich der Lernziele II und III erfassen. Die situativen Aufgaben erfordern die Fähigkeit, betriebswirtschaftliches Wissen in unterschiedlichen Anwendungsfeldern (Finanz- und Rechnungswesen, Unternehmensführung, Produktion und Marketing) einzusetzen. Zudem sollen die Fragen einen Bezug zu empirischer Datenbeschaffung, Auswertung und Interpretation aufweisen. Fähigkeiten aus dem Bereich des Lernziels IV (Kommunikation mit Fachleuten und Nicht-Fachleuten) werden nur indirekt in Aufgaben geprüft, in denen der Testteilnehmer gefragt wird, eine angemessene Präsentations- oder Kommunikationsform zu wählen. Die aktive Präsentation von Inhalten und Daten innerhalb des geschlossenen Antwortformates ist nicht möglich. Lernziel IV findet aufgrund der Anlage der Befragung als einmalige Erhebung keine Berücksichtigung. Auch wenn die von der OECD formulierten Lernziele (vgl. Tabelle 7) nicht systematisch entlang kognitiver Prozessdimensionen organisiert sind, so wird deutlich, dass von den Studierenden sowohl das Erinnern, Reproduzieren und Anwenden von Wissen als auch die Bewertung sowie kritische Reflexion von Inhalten erwartet werden. Für die Testentwicklung bedeutet das, dass die Aufgaben sich nicht nur

entlang der Wissensart sowie der inhaltlichen Anforderungsdimension orientieren sollten, sondern auch die kognitive Prozessdimension (vgl. Abschnitt 2.3.2) in einer zuvor festgelegten Variationsbreite systematisch widerspiegeln sollten. Dieses Desiderat in wird im Rahmenmodell der Testentwicklung der vorliegenden Arbeit berücksichtigt (vgl. Abschnitt 6.3).

Neben dem curricularen Zugang sollen bei der Entwicklung realistischer, situativer Items die Anforderungen des Arbeitsmarktes berücksichtigt werden. Ein Überblick über die Anforderungen, die an Bachelorabsolventen der Betriebswirtschaftslehre gestellt werden, erfolgt im folgenden Abschnitt.

6.2.2 Anforderungen des Lernens und Arbeitens in betriebswirtschaftlichen Handlungsfeldern

Die Anforderungen des Arbeitens an Hochschulabsolventen sind volatil und schwer zu quantifizieren. „Employability" und Schlüsselqualifikationen sind zwei zentrale Begriffe, die den Diskurs um die Anforderungen des Arbeitsmarktes an Hochschulabsolventen prägen (Schaeper & Wolter, 2008). Fachspezifische Kompetenzen geraten dabei zumeist aus dem Blickfeld. Neben Interviews mit potenziellen Arbeitgebern und Personalverantwortlichen bietet die Analyse von Stellenanzeigen die Möglichkeit abzuschätzen, welche betriebswirtschaftlichen Fachrichtungen und Fähigkeiten für die Bewältigung von betriebswirtschaftlichen Arbeitsplatzanforderungen relevant sind (Sailer, 2009).

Im Jahr 2004 analysierte Sailer (2009) 3787 Stellenanzeigen für Wirtschaftswissenschaftler im Hinblick auf den von Unternehmensseite geforderten Studienschwerpunkt. In seiner Analyse bestand die stärkste Nachfrage bei Wirtschaftsmathematikern und Wirtschaftsinformatikern, gefolgt von Studierenden mit Spezialisation im Marketing- bzw. Dienstleistungssektor und im Bereich Controlling. Relativ weniger häufig wurden die Schwerpunkte Materialwesen/ Logistik, Unternehmensführung/Strategie und Produktion nachgefragt (Rang 7 bin 9 nach Nennungshäufigkeit) (Sailer, 2009, S. 90).

Henning und Henning (2009) nannten mit Finanz- und Rechnungswesen, Vertrieb und Marketing ähnliche relevante Funktionen von Absolventen der Betriebswirtschaftslehre. Zudem führten sie auf Rang 4 Tätigkeiten im Bereich Management und Planung auf, die bei Sailer (2009) auf Platz 8 der Häufigkeitsliste rangierten (vgl. Tabelle 8). Die geforderten Schwerpunktsetzungen finden sich mit Ausnahme von Tätigkeiten im Personalbereich und allgemeinen verwaltenden Tätigkeiten weitgehend in den Pflichtmodulen der Bachelorstudiengänge in Betriebswirtschaftslehre und Wirtschaftspädagogik wieder (vgl. Abschnitt 6.2.1). Wirtschaftsmathematik (inkl. Operations Research) und Wirtschaftsinformatik sind an einigen Standorten nicht in den Studiengang der

Wirtschaftspädagogik integriert und werden daher trotz ihrer Wichtigkeit auf dem Arbeitsmarkt nicht in den Test aufgenommen.

Tab. 8: Die zehn häufigsten Funktionsbereiche von Absolventen der Betriebswirtschaftslehre (testrelevante Bereiche grau unterlegt)

Rang	Die 10 häufigsten Funktionsbereiche von Absolventen der BWL
1	Finanz- und Rechnungswesen
2	Vertrieb
3	Marketing/Public Relations/Werbung
4	Management Planung
5	Personalverwaltung
6	Softwareproduktion
7	Allgemeine Verwaltung/Koordination
8	Aus- und Weiterbildung
9	Materialwirtschaft/Logistik
10	Führungsebene/Unternehmensleitung

Aus den Analysen wird deutlich, dass insbesondere Absatz und Vertrieb sowie Finanz- und Rechnungswesen (Controlling) gefragte Aufgabenbereiche für Absolventen der Betriebswirtschaftslehre sind. Funktionen mit Bezug zu Materialwirtschaft und Produktion sind hingegen seltener und sollten entsprechend in dem Test auch durch weniger Aufgaben repräsentiert werden. Es ist jedoch zu beachten, dass Stellenanzeigenanalysen immer nur für den Zeitraum der Datenergebung gültig sind und somit zeitlicher Begrenzung und kurzfristigen Arbeitsmarkttrends unterworfen sind (Sailer, 2009).

Um die Anforderungen an Absolventen der Betriebswirtschaftslehre an aktuelleren Daten zu überprüfen, wurden im Rahmen einer Diplomarbeit an der Georg-August-Universität Göttingen im Jahr 2012 zwischen März und April 600 Stellenanzeigen von einer bekannten Onlinestellenbörse analysiert (Jahn, 2012). Es wurden nur Stellenanzeigen in Betracht gezogen, die ein betriebswirtschaftliches Studium und nicht mehr als fünf Jahre Berufserfahrung voraussetzten. Von der Analyse ausgeschlossen wurden Stellenanzeigen für Trainees und Praktikanten. Eine an vier Stichtagen durchgeführte Überprüfung der Anzahl der Stellenanzeigen pro Berufsfeld[9] ergab ein ähnliches Bild wie die Analysen von Henning und Henning (2009) sowie Sailer (2009) (vgl. Tabelle 9).

9 Berufsfelder werden von der Stellensuchmaschine vorgegeben und von Seiten der Stellenanbieter den Stellenanzeigen zugeordnet.

Tab. 9: Die zehn am häufigsten im Onlinestellenmarkt ausgeschriebenen Berufsfelder für BWL-Absolventen im April 2012 (Mehrfachnennungen möglich, testrelevante Bereiche grau hinterlegt)

Rang	Berufsfeld	Anzahl der Stellenanzeigen	Anteil %
1	Finanz- und Rechnungswesen inklusive Wirtschaftsprüfung	426	25.9
2	IT-Telekommunikation	416	25.3
3	Vertrieb, Handel & Einkauf	314	19.1
4	Marketing & Werbung	256	15.6
5	Unternehmensführung & Management	237	14.4
6	Ingenieurwesen & technische Berufe	190	11.5
7	Banken, Versicherungen & Finanzdienstleister	135	8.2
8	Personalwesen	108	6.6
9	Transport & Logistik	93	5.7
10	Kaufmännische Berufe & Assistenz	66	4.0
	Sonstige	67	4.1

Der von Arbeitgeberseite am häufigsten nachgefragte Einsatzbereich für Absolventen der Betriebswirtschaftslehre liegt im Bereich des Finanz- und Rechnungswesens, gefolgt von IT- und Telekommunikationsberufen, die in dem zu entwickelnden Test jedoch nicht berücksichtigt werden. Positionen im Vertrieb und Marketing wurden sowohl 2004 als auch 2012 häufig gesucht. Der Bedarf an Absolventen in Managementpositionen fällt hingegen etwas geringer aus. Alle Analysen bestätigen die relativ geringe Nachfrage nach BWL-Absolventen für die Bereiche Produktion, Materialwirtschaft und Logistik.

Aus den oben angeführten Anforderungen an Absolventen und der Auflistung häufiger Funktionen wird jedoch noch nicht deutlich, mit welchen fachlichen und situativen Anforderungen die Absolventen nach Beendigung ihres Studiums konfrontiert werden. Um einen Eindruck zu bekommen, welche konkreten Anforderungen Absolventen in bestimmten Tätigkeitsbereichen meistern sollen, analysierte Jahn (2012) die fachlichen Anforderungen, die innerhalb der fünf am häufigsten gefragten Berufsfelder an Absolventen und junge Berufstätige mit betriebswirtschaftlichem Studienabschluss (ab Bachelor) gerichtet waren. Die für ein Berufsfeld typischen Tätigkeiten wurden eklektisch in die situativen Aufgabenstellungen einbezogen. Eine Beschreibung dieses Vorgehens erfolgt in Abschnitt 6.3.

6.3 Rahmenmodell der Itementwicklung

Aus den vorangegangenen Kapiteln und der Betrachtung der Domäne in Abschnitt 6.2 wird das Rahmenmodell für die Itementwicklung abgeleitet. Im ersten Abschnitt werden die Implikationen der bisherigen Forschung für die Testentwicklung zusammengetragen. Anschließend werden in Anlehnung an Abschnitt 4.2.2 Anforderungen an die kognitiven Anforderungsniveaus der Testitems formuliert. Es folgt eine Beschreibung der Testentwicklungsschritte unter Einbezug der in Abschnitt 6.2 umrissenen Anforderungen des Lernens und Arbeitens an Studierende der Betriebswirtschaft. Das Kapitel schließt mit der tabellarischen Darstellung des entwickelten Testinstruments.

6.3.1 Implikationen der bisherigen Forschung für die Testentwicklung

Auf der Grundlage bisheriger Forschungsarbeiten wird der zu entwickelnde Test als papierbasierter Test mit situativen Aufgabenstämmen und vier möglichen Antwortalternativen umgesetzt. Die Entwicklung des Tests wird vor dem Hintergrund eines einfachen dichotomen Scoring-Modells vorgenommen. Zum einen sollen die Ergebnisse mit dem bereits bestehenden, dichotom gescorten BAKT (Bothe, 2003) in Beziehung gesetzt werden. Zum anderen sollen die Testergebnisse auf Grundlage der Item-Response-Theorie ausgewertet werden, die für dichotome Daten, speziell bei relativ kleinen Stichprobengrößen, die besten Auswertungsmethoden bietet (Bühner, 2011). Das einfache „pick best"-Antwortformat hat sich in Studien mit situativen Aufgaben zu unterschiedlichen Themen bewährt (z. B. Hunter, 2003; Stevens & Campion, 1994; Weekley et al., 2006). Der Überlegung, dass nicht alle falschen Antwortalternativen in gleichem Maße falsch sein müssen (Weekley et al., 2006), wird Rechnung getragen, indem bewusst mindestens eine Antwortalternative generiert wird, die teilrichtige Elemente enthält, jedoch sachlogisch begründbar der „richtigen" Antwort unterlegen ist. Die Darbietung der Items erfolgt papierbasiert. Die Umsetzung als Papierversion ist insofern sinnvoll, als dass der Test sich zunächst als Papierversion vor den Standards pädagogisch-psychologischer Diagnostik bewähren sollte (vgl. Kapitel 7), bevor Ressourcen in eine multimediale Umsetzung der Testitems investiert werden. Zudem sollte der Bezug zu bestehenden Arbeiten wie dem BAKT (Bothe, 2003) nicht durch abweichende Erhebungsmethoden erschwert werden. Variierte Erhebungsmethoden könnten einen Methodenbias (Schermelleh & Schweizer, 2012) begünstigen und zu Verzerrungen beim Vergleich zwischen den beiden Fragebögen führen. Studien belegen, dass die Art und Weise der Testadministration (computerbasiert vs. papierbasiert) einen substanziellen Einfluss auf die die Bearbeitungsstrategien und Testergebnisse haben kann (Pomplun, Frey & Becker, 2002). Zudem ist es wahrscheinlich, dass die Erhebungsmethode einen Einfluss auf motivationale

Prozesse während der Testbearbeitung hat (z. B. Kanning, 2008). Zuletzt erleichtert die Durchführung als Papierversion die Datenerhebung im Rahmen von universitären Lehrveranstaltungen, da keine Computerressourcen eingeplant werden müssen.

Auf der Grundlage der skizzierten Konstruktdiskussion (vgl. 5.3.1) wird der hier zu entwickelnde Test nach folgenden Überlegungen angelegt:

Im Sinne von Schmidt und Hunter (1993) soll der Test Wissen in der Domäne der Betriebswirtschaft auf dem Niveau des Bachelorabschlusses erfassen. In Anlehnung an Stemler und Sternberg (2006) sowie Hofmeister (2005) wird davon ausgegangen, dass situative Aufgaben sowohl deklarative als auch prozedurale Wissensbestände erfassen können. Situative Aufgaben werden als Maß kompetenznaher Konstrukte verstanden (Rost, 2008), das Handlungswissen auf deklarativer und prozeduraler Ebene erfassen kann.

Da mit dem BAKT (Bothe, 2003) bereits ein deklarativer Wissenstest zur Verfügung steht, sollen die Aufgaben so konstruiert werden, dass weitgehend Handlungswissen auf prozeduraler Ebene damit erfasst wird. Um dieses Ziel zu erreichen, wird die Systematik der Problemlöseproduktionen (Anderson, 2001; Sternberg et al. 1995) bei der Aufgabenkonstruktion berücksichtigt. Eine detaillierte Beschreibung dieses Vorgehens erfolgt in Abschnitt 6.3.

Da die Auswertung nach Item-Response-Theorie (vgl. Abschnitt 7.1.2) ein zentraler Bestandteil der vorliegenden Arbeit sein soll, wird von einer Szenariotechnik mit Sub-Fragestellungen abgesehen. Stattdessen werden die Situationen zugunsten einer ausreichenden Itemanzahl stark verkürzt dargestellt. Dieses Vorgehen hat den Nachteil, dass die Situationen weniger detailliert beschrieben werden und somit an Authentizität einbüßen. Solche abstrakten Situationsbeschreibungen bieten jedoch den Vorteil, dass die Relevanz der Aufgaben für breitere Tätigkeitsfelder erhalten bleibt. Für die Zielgruppe der Hochschulstudierenden ist ein solches Vorgehen angemessen, da das Bachelorstudium auf ein heterogenes berufliches Aufgabenfeld vorbereit.

Wie in der kritischen Betrachtung situativer Aufgaben am Ende von Kapitel 5 herausgearbeitet wurde, besteht ein sogenanntes Reliabilitäts-Validitäts-Dilemma (Rost, 2004). Das Dilemma beschreibt die Situation, dass mit steigender Heterogenität der Items die kriteriale Validität steigt, die Reliabilität aber sinkt. Mehrdimensionale Modelle (insbesondere „within-item"-Modelle) bieten hier die Möglichkeit das Problem zu begrenzen, erfordern aber eine theoretische und empirische Basis, die zum jetzigen Zeitpunkt der Testentwicklung noch nicht gegeben ist. Kritisch ist zu betrachten, dass die Situationsbeschreibungen Verkürzungen einer realen Situation darstellen und keine dynamischen Elemente enthalten. Dynamische Testdesigns sind wünschenswert, erfordern aber

ebenso wie eine multidimensionale Modellierung eine genaue Kenntnis der Itemeigenschaften, die zum jetzigen Zeitpunkt der Testentwicklung noch nicht vorliegen.

Als Konsequenz aus den obigen Überlegungen werden folgende Rahmenbedingungen für die Itementwicklung vorgegeben:

- Der Test wird papierbasiert umgesetzt.
- Pro Itemstamm werden vier Antworten vorgegeben.
- Situationen im Itemstamm werden so real wie möglich, aber verkürzt dargestellt.
- Die Iteminhalte orientieren sich in erster Linie an den Pflichtinhalten des universitären Bachelorcurriculums für Wirtschaftspädagogik und Betriebswirtschaftslehre.
- Die Iteminhalte orientieren sich in zweiter Linie an Anforderungen aus Stellenanzeigen für Absolventen der Betriebswirtschaftslehre.
- Die Items befinden sich auf unterschiedlichen kognitiven Anforderungsstufen.

Die Hintergründe der Bestimmung der Aufgabenanforderungen werden im folgenden Abschnitt näher beleuchtet.

6.3.2 Spezifikation der kognitiven Anforderungen

Das zu erfassende Merkmal wird, in Anlehnung an Anderson und Krathwohl (2001), als Wissen über Methoden und Prozeduren der Betriebswirtschaftslehre und das Wissen darüber, in welchen Situationen diese angewendet werden sollen, definiert. Die Items sollten so konstruiert werden, dass das kognitive Anforderungsniveau der Mehrzahl der Items auf dem Niveau des Anwendens und Verstehens nach Hofmeister (2005) in Anlehnung an Anderson und Krathwohl (2001) liegt. Beim Anwenden prüft der Testteilnehmer die vorliegenden Situationsbeschreibungen und sucht aus bereits bestehenden Handlungsschemata passende Reaktionen aus (Hofmeister, 2005). Die situative Einbettung ist für Aufgaben auf Anwendungsniveau erforderlich, da sonst keine Transferleistung abgefragt werden kann (Hofmeister, 2005). In den Aufgaben sollen also bekannte Methoden und Prozeduren in neuen Situationen angewandt werden. Wünschenswert für den Testaufbau sind einige Items, die vom Testteilnehmer kritisches und reflexives Denken erfordern. Dabei geht es darum, über bestehende Handlungsschemata hinaus einen Sachverhalt systematisch hinsichtlich relevanter Komponenten zu untersuchen (Hofmeister, 2005). Bei einem geschlossenen Antwortformat ist das Anregen kritischer und reflexiver

Denkprozesse durch die Vorgabe von Antwortoptionen erschwert. Nur Aufgaben, die eine Bewertung von Inhalten anhand von selbstgewählten Kriterien erfordern und dem Testteilnehmer eine eigene Einschätzung abverlangen, erfüllen die Anforderung der kritischen Reflexion (Hofmeister, 2005).

Neben dem kognitiven Anforderungsniveau sollte auch die inhaltliche Komplexität der Items systematisch variiert werden. Die inhaltliche Komplexität soll den Grad der Vernetztheit von Wissen aus unterschiedlichen Inhaltsbereichen widerspiegeln. In der vorliegenden Arbeit wird mit jeder in der Situationsbeschreibung dargestellten Bedingung (UND-Verknüpfung) von einer steigenden inhaltlichen Komplexität ausgegangen (vgl. Absatz 5.3.1). Zudem soll die Modellierungsleistung, die die Aufgaben erfordern, systematisch variiert werden. Dabei wird zwischen Aufgaben, deren Lösung eine mathematische Modellierungsleistung erfordert, und Aufgaben, die keine mathematische Modellierungsleistung erfordern, unterschieden.

Bezogen auf die in Tabelle 7 dargestellten Lernziele für Bachelorabsolventen sollen die Items der Überprüfung der Lernziele II (Anwendung von Fachwissen auf reale Probleme) und III (Gebrauch von quantitativen Methoden) dienen. Es wird davon ausgegangen, dass das Lernziel I (Fachwissen demonstrieren) hinreichend durch die Aufgaben des BAKT (Bothe, 2003) abgedeckt wird. Lernziel IV (Kommunikation mit Fach- und Nicht-Fachleuten) wird indirekt bei Aufgaben geprüft, die danach fragen, wie der Testteilnehmer bestimmte Sachverhalte präsentieren oder kommunizieren würde. Ein tatsächlicher Einsatz von Präsentationsmedien im gewählten Papier-Bleistift-Testformat wäre nicht umsetzbar gewesen. Das Lernziel V (selbstregulatives Lernen) wird nicht über die Testaufgaben erfasst.

6.3.3 Itementwicklung und resultierende Testcharakteristika

Aus den vorangegangenen theoretischen Überlegungen und Analysen der domänenspezifischen Anforderungen an Studierende wurden folgende Rahmenvorgaben für die Itementwicklung abgeleitet:

1. Die inhaltlichen Anforderungen der Items beziehen sich auf die Grundlagenmodule Unternehmensführung, Finanz- und Rechnungswesen, Produktion und Marketing.
2. Die inhaltlichen Anforderungen decken sich mit Anforderungen aus Stellenausschreibungen.
3. Der Aufgabenstamm ist so aufgebaut, dass er der Wenn-Komponente eines Bedingungs-Aktions-Paares entspricht. Der Aufgabenstamm enthält unterschiedlich viele Überprüfungen im Sinne einer oder mehrerer UND Bedingungen im Wenn-Teil der Produktion.

4. Für jede Situation werden 4 Antwortalternativen vorgegeben, die dem Dann-Teil der Produktion entsprechen. Es gibt nur eine vollkommen richtige Antwort und mindestens eine oder mehrere Antwortoptionen, die teilrichtige Elemente enthalten.
5. Die Aufgaben entsprechen der in Abschnitt 6.3.2 vorgegebenen Verteilung der kognitiven Anforderungen.

Die Umsetzung der definierten Entwicklungsschritte wird anhand eines Beispielitems aus dem Bereich des internen Rechnungswesens vorgestellt. Das in Abbildung 7 dargestellte Item bezieht sich auf den Modulinhalt „Entscheidungsprobleme und Entscheidungsrechnung" (vgl. Tabelle A1–1 im Anhang A). Eine typische Anforderung aus Stellenanzeigen in diesem Fachbereich lautet: „Entscheidungsvorlagen liefern" (Jahn, 2012, S. 53). Der Bezug zu arbeitsplatztypischen Anforderungen war im Itementwicklungsprozess der curricularen Verankerung nachgeordnet, was durch die gestrichelte Umrandung in Abbildung 7 gekennzeichnet ist. Die beiden Anforderungen wurden in Anlehnung an eine Fallstudie zu Scoring-Modellen (Horváth, Gleich & Voggenreiter, 2001, S. 103) in eine Situationsbeschreibung übersetzt. Um die Aktivierung einer Produktionsregel hervorzurufen, wird im Aufgabenstamm das Ziel der Operation festgelegt. Mit dieser Festlegung wird zudem vermieden, dass Testteilnehmer die Aufgaben vor dem Hintergrund unterschiedlicher Zielsetzungen beantworten und somit unsystematische Varianz generiert wird.

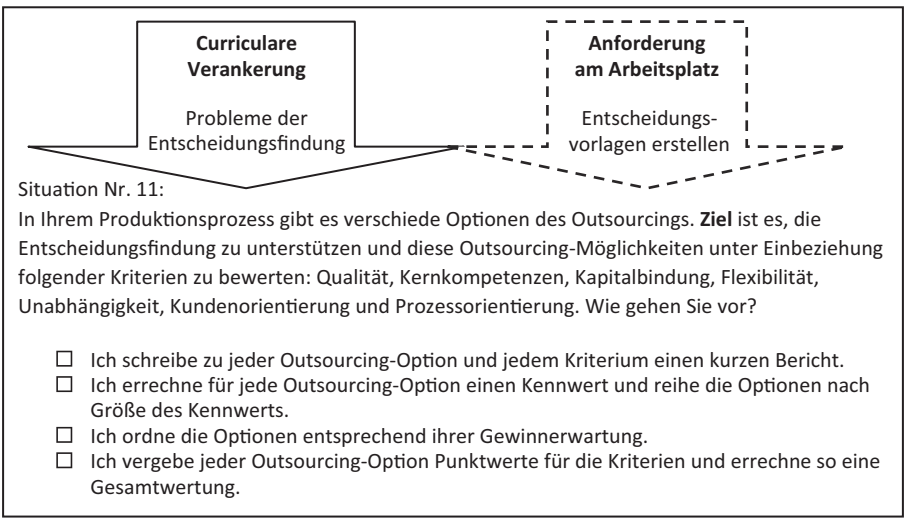


Abb. 7: Darstellung der Verankerung von Iteminhalten am Beispiel des Items S. 11 (siehe Anhang E für die Auflistung aller Items und Tabelle E-2 für die jeweiligen Verankerungen)

Das Item wird als Item mit geringer inhaltlicher Komplexität (I) eingestuft, da im Itemstamm keine zu prüfenden Bedingungen genannt werden. Dem Item würde eine höhere Komplexität zugesprochen werden, wenn im Itemstamm eine Gewichtung der Entscheidungskriterien erwähnt worden wäre. Obwohl Entscheidungsprobleme im Arbeitsleben nicht mit vordefinierten Antwortmöglichkeiten geliefert werden, wurde ein geschlossenes Antwortformat gewählt, um die höchstmögliche Auswertungsobjektivität zu gewährleisten (Bühner, 2011). Um die Mehrdeutigkeit von Situationen des wirklichen Lebens anzudeuten, enthält neben der richtigen Antwortmöglichkeit mindestens eine weitere Antwortmöglichkeit teilweise richtige Elemente, wird aber als falsch gewertet. Die so entwickelten 45 Items durchliefen einen mehrstufigen Beurteilungsprozess durch Studierende, Fachdidaktiker, Diagnostiker und Fachwissenschaftler der Universität Göttingen. Der Itementwicklungsprozess ist im Anhang B, Abbildung B-1 dargestellt.

Aus den oben beschriebenen theoretischen Überlegungen und dem skizzierten Entwicklungsprozess resultieren die in Tabelle 10 dargestellten Testcharakteristika.

Tab. 10: Tabellarische Darstellung der Testzusammensetzung und intendierte Itemeigenschaften
Kognitive Prozessdimension: 1 = Reproduzieren, 2 = Anwenden, 3 = Reflektieren
Komplexität bestimmt nach Anzahl der Wenn-Bedingungen im Aufgabenstamm

Modulbezug	Item Nr.	Inhaltliche Anforderung	Kognitive Prozess-dimension	Kom-plexität	Lernziele nach OECD	Mathema-tische Modellie-rung
Unterneh-mensführ-rung	1.1	Laterale Kommunikation umsetzten	1	1	II & IV	nein
	1.2	Konsequenzen von Dezentralisierungsprozessen kom-munizieren	3	1	II & IV	nein
	1.3	Planungstechniken in der strategischen Projektpla-nung anwenden	2	1	II	nein
	1.4	Hauptversammlung organisieren	1	1	II	nein
	1.5	BCG-Matrix in Produktpräsentation nutzen	3	2	II & IV	nein
	1.6	SWOT-Analyse durchführen	1	1	II	nein
	1.7	Leitbild verfassen	2	1	II	nein
	1.8	Geeignete Lohnart bestimmen	3	2	II	nein
	1.9	Preis bei Merger erhöhen	3	1	II	nein
	1.10	Outsourcing-Optionen beurteilen	2	1	II	nein
	1.11	Geeignete Leitungskonfiguration bestimmen	3	2	II	nein
Finanz- und Rechnungs-wesen	2.1	Deckungsbeitrag errechnen	2	2	III	ja
	2.2	Methode zur Darstellung von Ergebnissen unter ver-schiedenen Bedingungsfaktoren aufzeigen	3	1	III	nein
	2.3	Erstellen eines Berichtsformulars im Vertrieb	2	1	II & III	nein
	2.4	Soll-Ist-Vergleich Bearbeitungszeit in der Produktion	2	1	II& III	ja
	2.5	Periodenerfolgsrechnung im internationalen Unter-nehmen auswählen	3	2	II & III	nein

(Fortsetzung Tab. 10)

Modulbezug	Item Nr.	Inhaltliche Anforderung	Kognitive Prozess-dimension	Kom-plexität	Lernziele nach OECD	Mathema-tische Modellie-rung
	2.6	Outsourcing-Optionen anhand von Kriterien bewer-ten	2	1	II & III	nein
	2.7	Unternehmenskennzahlen bewerten	2	2	II	ja
	2.8	F & E-Quote bewerten	1	1	II	nein
	2.9	Variable Stückkosten berechnen	2	1	III	Ja
Produktion	3.1	ABC-Analyse durchführen	1	1	II	nein
	3.2	Kriterien für Wahl eines Produktionsverfahrens fin-den	1	2	II	nein
	3.3	Anpassungsstrategien finden	3	2	II	nein
	3.4	Gesamtkostenverlauf prognostizieren	3	2	II	nein
	3.5	Engpässe in Produktion beheben	2	1	II	nein
	3.6	Passende Methode zur Bestimmung von Produktions-mengen nennen	3	2	II & III	nein
	3.7	Collaborative Planning umsetzen	1	1	II	nein
	3.8	Arbeitsunfälle in der Produktion vermeiden	3	1	II	nein
	3.9	Auslieferungsformen anhand von Kriterien bewerten	2	2	II	nein
	3.10	Verbrauchmaterialien bestellen	1	1	II	nein
	3.11	Produktionssysteme modellhaft abbilden	3	2	II & III	nein
Marketing	4.1	Methode für Umfrage zur Akzeptanz von Onlineka-talogen finden	3	2	II	nein
	4.2	Kognitive Dissonanz beim Kunden reduzieren	2	1	II	nein
	4.3	Kriterien für Marktsegmentierung benennen	1	1	II	nein

(Fortsetzung Tab. 10)

Modulbezug	Item Nr.	Inhaltliche Anforderung	Kognitive Prozess-dimension	Kom-plexität	Lernziele nach OECD	Mathema-tische Modellie-rung
	4.4	Passenden Testmarkt wählen	3	2	II	nein
	4.5	Marketing-Slogan entwerfen	2	2	II	nein
	4.6	Produktpolitik umsetzen	2	1	II	nein
	4.7	Studiendesign für Analyse des Kaufverhaltens wählen	3	1	II & III	nein
	4.8	Family Branding umsetzen	1	1	II	nein
	4.9	Marktforschungsauftrag vergeben	3	1	II	nein
	4.10	Methode zur Bestimmung der Stärke des Einflusses von Werbebudget auf Absatz benennen	3	2	II &III	nein
	4.11	Erwartungswert errechnen	2	2	II & III	ja
	4.12	Preisbereitschaft bestimmen	3	2	II & III	nein
	4.13	Einzelpreis bestimmen	2	2	II & III	ja
	4.14	Bündelpreis bestimmen	2	2	II & III	ja

6.4 Zusammenfassung

In diesem Kapitel wurde die Entwicklung eines situativen betriebswirtschaftlichen Wissenstests bis zur Pilotierung beschrieben. Die Testentwicklung erfolgte modellbasiert. Dabei wurde sowohl auf eine systematische Auswahl der Testinhalte als auch auf eine systematische Verteilung des kognitiven Anforderungsniveaus über die Items geachtet. Zwar sind die genutzten Modelle des Lernens unvollständig und dementsprechend in ihrer Aussagekraft begrenzt, sie können jedoch einzelne Bereiche des Lösungsprozesses beschreiben und helfen, diese zu verstehen (Winther, 2010, S. 102). Die Auswahl der Inhaltsbereiche deckt ebenso nicht alle relevanten universitären Bildungsinhalte ab (es fehlen zum Beispiel Inhalte aus der Wirtschaftsinformatik). Die Bereiche Finanz- und Rechnungswesen, Produktion, Absatz und Marketing sowie Unternehmensführung können jedoch als Kernbereiche der betriebswirtschaftlichen Grundbildung von Studierenden der Wirtschaftspädagogik und Betriebswirtschaftslehre angesehen werden.

Nachdem sowohl die „Construct Map" als auch das „Item Design" vorab spezifiziert wurden (Wilson, 2005), müssen die Aufgaben an einer hinreichend großen Stichprobe pilotiert werden. Die Pilotierung dient in erster Linie der Itemselektion auf Basis empirischer Itemkennwerte. In zweiter Linie sollen erste Aussagen zu unterschiedlichen Aspekten der psychometrischen Güte des Tests getroffen werden. Bevor die Pilotierung durchgeführt wird, muss jedoch der „Outcome Space" des Tests bestimmt werden und ein passendes Messmodell spezifiziert werden (Wilson, 2005). Die Testoptimierung sowie die Wahl des Messmodells erfolgen anhand aktueller Standards der pädagogisch-psychologischen Diagnostik, die im folgenden Kapitel in ihren unterschiedlichen Facetten beschrieben werden.

7 Standards pädagogisch-psychologischer Diagnostik

Die wissenschaftliche Messung latenter Konstrukte wird von der Testentwicklung bis zur Testauswertung und Interpretation von Standards geleitet. Sogenannte psychometrische Modelle bieten den Rahmen für eine wissenschaftliche, modellgeleitete Entwicklung und Auswertung von Tests. Die beiden derzeit vorrangigen Testtheorien werden im folgenden Unterkapitel erläutert. Der Fokus der Beschreibung liegt dabei auf den modernen probabilistischen Testtheorien. Neben den probabilistischen Testentwicklungs- und Auswertungsmethoden spielen sogenannte klassische Gütekriterien eine tragende Rolle in der Beurteilung der diagnostischen Güte von Tests. Diese werden im zweiten Abschnitt des Kapitels vorgestellt.

7.1 Modelle zur Auswertung von Tests

Eine Testtheorie macht Aussagen zum Zusammenhang von Testverhalten und dem zu erfassenden latenten Merkmal (Rost, 2004) und darüber, wie sich die Testwerte aufgliedern (Bühner, 2011). Eine solche Theorie ist wichtig, weil bei der Auswertung eines Tests vom Antwortverhalten eines Probanden auf die Ausprägung des zu erfassenden persönlichen Merkmals geschlossen wird (Rost, 2004). Test- und messtheoretische Überlegungen sind ein Teil der bereits beschriebenen Modelle der Testentwicklung (vgl. Abschnitt 6.1.1). Die in diesem Abschnitt thematisierten Testtheorien nehmen jedoch spezifisch das Verhältnis von beobachtbaren Testwerten und dem zu messenden Konstrukt in den Blick. Der Gegenstandsbereich der pädagogisch-psychologischen Testtheorien ist in Abbildung 8 dargestellt.

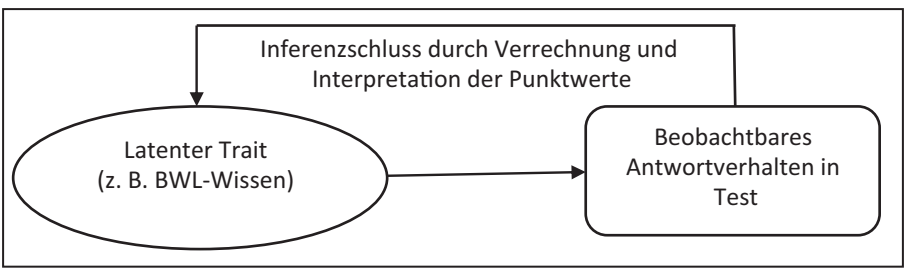


Abb. 8: Gegenstandsbereich pädagogisch-psychologischer Testtheorien

(eigene Darstellung in Anlehnung an Wilson, 2005, S. 13)

Die Entwicklung und Auswertung des vorliegenden Tests folgt in weiten Teilen den Überlegungen der probabilistischen Testtheorien (z. B. Rasch, 1960), ergänzt durch Standards der klassischen Testtheorie (KTT) (AERA et al., 2008).

Die klassische Testtheorie macht in erster Linie Annahmen darüber, wie sich der „wahre Testwert" einer Person zusammensetzt. Die Bestimmung des Messfehlers, der die wahren Werte überlagert, spielt in der KTT eine prominente Rolle. Im Rahmen der probabilistischen Testtheorien wird das Antwortverhalten für jedes einzelne Item unter Berücksichtigung der Personenfähigkeit und verschiedener Itemcharakteristika modelliert. Probabilistische Testtheorien werden deshalb oft als Item-Response-Theorien bezeichnet (IRT) und gelten in der empirischen Bildungsforschung als State of the Art (Rost, 2004). Beide Ansätze und ihre Bedeutung für die Testauswertung werden im Folgenden beschrieben, wobei die probabilistische Testtheorie im Vordergrund steht und zur klassischen Testtheorie (KTT) abgegrenzt werden soll.

7.1.1 Die klassische Testtheorie

Die klassische Testtheorie war viele Jahre lang die etablierte Theorie in der diagnostischen Praxis. Der klassischen Testtheorie liegt die Annahme zugrunde, dass Messwerte einzelner Personen über verschiedene Messungen hinweg variieren (Bühner, 2011). Daraus ergibt sich die Grundidee der klassischen Testtheorie, dass sich ein beobachtbarer Testwert aus einem wahren Wert und einen Messfehler zusammensetzt (Alagumalai & Curtis, 2005). Ziel der klassischen Testtheorie ist dementsprechend die Bestimmung des Messfehlers. Es wird davon ausgegangen, dass dieser Messfehler bei der Erfassung von latenten Merkmalen immer auftritt (Bühner, 2011). In der KTT werden nur unsystematische Messfehler in Betracht gezogen. Solche Messfehler können durch unsystematische innere und äußere Einflussfaktoren zustande kommen. Zu den inneren Messfehlerquellen gehören zum Beispiel Ermüdung oder schwankende Motivation bei der Testbearbeitung. Zu den äußeren Bedingungen, die Messfehler begünstigen, zählen zum Beispiel Lärm oder Wetterschwankungen während einer Testung.

Das zentrale Axiom der KTT lautet:

$$X_{vt} = \tau_v + \epsilon_{vt} \qquad \text{(Formel 1)}$$

X_{vt} stellt den beobachteten Wert einer Person dar, τ_v steht für den wahren Wert einer Person und ϵ_{vt} beschreibt den Fehlerterm (Bühner, 2011, S. 55).

Die KTT beruht auf einer Anzahl von logisch abgeleiteten Axiomen, die sich einer empirischen Prüfbarkeit entziehen. Zum einen hat der Erwartungswert der Fehler den Wert Null, der Messfehler ist mit dem wahren Wert unkorreliert und

der wahre Wert und Fehlerwert zweier verschiedener Tests sind unkorreliert. Zum anderen sind die Fehlerwerte von zwei verschiedenen Tests unkorreliert (Bühner, 2011).

Das heißt, die KTT macht keine Annahmen darüber, wie ein Item beantwortet wird, sondern zieht in Betracht, wie sich beobachtete Testwerte zusammensetzen. Aus diesem Grund wird die KTT auch oft als Messfehlertheorie bezeichnet (Bühner, 2011). Obwohl sich die KTT praktisch in zahlreichen Arbeiten bewährt hat, unterliegt sie vielfach kritischer Betrachtung.

Ein zentraler Kritikpunkt ist die Vernachlässigung der Verbindung zwischen dem zu messenden Merkmal und der Itembeantwortung (Bühner, 2011). Damit einher geht die Problematik der nicht erkannten Mehrdimensionalität. Denn erst wenn abgesichert ist, dass die Items eines Tests eindimensional sind, also nur ein latentes Merkmal erfassen, ist der wahre Wert der KTT sinnvoll interpretierbar. Eindimensionalität wird in der KTT jedoch nicht überprüft. Darüber hinaus sind die Axiome der KTT keiner empirischen Überprüfung zugänglich, sondern werden aus der Annahme der Zusammensetzung des Testwertes logisch geschlussfolgert. Einige der Annahmen müssen jedoch in Zweifel gezogen werden, wie zum Beispiel die Unabhängigkeit des wahren Wertes vom Messfehler oder die Annahme, dass Messfehler immer unabhängig voneinander sind (Bühner, 2011). Zudem wurden die Kennwerte der klassischen Testtheorie (z. B. Cronbachs Alpha) für intervallskalierte Daten entwickelt und können nicht ohne Weiteres auf die in Leistungstests typischen dichotomen Nominaldaten und Rangdaten übertragen werden (Bühner, 2011). Neben den oben aufgeführten Kritikpunkten fällt die Kritik der Stichprobenabhängigkeit der KTT-Kennwerte besonders ins Gewicht. So variieren in der KTT die Itemkennwerte mit der Stichprobe und die Ergebnisse unterschiedlicher Tests, die das gleiche Konstrukt messen, sind nicht miteinander vergleichbar. In einem solchen Fall muss die Summe der Rohpunkte je nach Testumfang unterschiedlich interpretiert werden. Unter anderem aus diesen Kritikpunkten heraus entwickelte Georg Rasch 1960 die erste probabilistische Testtheorie (Rasch, 1960), deren Verwendung Rost (2004) speziell für dichotome Daten empfiehlt. Probabilistische Testtheorien, unter besonderer Betrachtung des dichotomen Rasch-Modells, werden im folgenden Abschnitt beschrieben.

7.1.2 Probabilistische Testtheorien

Trotz der „revolutionären Bedeutung“, die Kubinger (2000, S. 2) den probabilistischen Testtheorien für die Diagnostik latenter Merkmale zuspricht, hat es lange gedauert, bis diese sich in ihren Anwendungsfeldern etabliert haben (Rost, 2004). Die Modellierung von Kompetenzen mittels probabilistischer Testmodelle hat sich insbesondere durch die internationalen Vergleichsstudien

TIMSS (Baumert, Bos & Lehmann, 2000) und PISA (Klieme, 2010) als Standard in der empirischen Bildungsforschung etabliert.

Im Gegensatz zur klassischen Testtheorie nehmen probabilistische Testtheorien das Antwortverhalten von Personen je Item in den Blick (Bühner, 2011). Dabei werden Modelle unterschiedlicher Komplexität aufgestellt, die die Wahrscheinlichkeit einer Itemlösung vorhersagen. Mögliche Parameter, die die Lösungswahrscheinlichkeit einer Person für ein bestimmtes Item beeinflussen können, sind die Fähigkeit einer Person (θ, Theta), die Itemschwierigkeit (β, Beta), die Ratewahrscheinlichkeit (γ, Gamma) und die Itemtrennschärfe (σ, Sigma) (Bühner, 2011). Der Begriff probabilistische Testtheorien beschreibt nicht eine singuläre Theorie, sondern eine Gruppe von Testmodellen. Das einfachste und bekannteste Modell ist das Rasch-Modell (1-Parameter-Modell). Es gilt für dichotome Daten, die häufig bei Leistungstests vorliegen. Komplexere Modelle für dichotome Daten, in denen mehr Parameter als der oben angegebene Schwierigkeitsparameter geschätzt werden, sind das Birnbaum-Modell (2-Parameter-Modell) und das 3-Parameter-Modell (Rost, 2004). Darüber hinaus gibt es Modelle für partial-credit Daten (Masters, 1982) und Daten aus Ratingskalen (Andrich, 2011). Mehrparametrische Modelle benötigen sehr große Stichproben für eine stabile Parameterschätzung. Deshalb wird in der vorliegenden Arbeit in Anlehnung an die Empfehlung von Bühner (2011) das 1-Parameter-Modell vorgezogen. Im Folgenden wird das Rasch-Modell vorgestellt, da dieses zur Auswertung der Pilotierungsdaten herangezogen wird und zudem eine gute Basis darstellt, um die Grundgedanken der probabilistischen Testtheorien zu erklären.

7.1.2.1 Grundüberlegungen des Rasch-Modells

Das Rasch-Modell sagt voraus, dass mit steigender Personenfähigkeit die Wahrscheinlichkeit einer Itemlösung zunimmt (Bühner, 2011). Die Beziehung zwischen Fähigkeit und Leistung bei einer Testaufgabe wird nicht als deterministisch angesehen, sondern durch eine Wahrscheinlichkeitsfunktion modelliert. Daher rührt der Begriff der probabilistischen Testtheorien. Im einfachen dichotomen Rasch-Modell hängt die Lösungswahrscheinlichkeit für ein bestimmtes Item zum einen von der Personenfähigkeit ab, zum andern von der Aufgabenschwierigkeit (Bühner, 2011). Damit wird der Tatsache Rechnung getragen, dass eine Person mit geringer Fähigkeitsausprägung mit einer geringen Wahrscheinlichkeit Aufgaben lösen kann, die über ihrem eigentlichen Fähigkeitsniveau liegen. Ebenso, wie es einer sehr fähigen Person in seltenen Fällen nicht gelingt, eine Aufgabe zu lösen, die unter ihrem Fähigkeitsniveau liegt (Bühner, 2011). Das Rasch-Modell ist ein einfaches Modell, da es nur zwei Pa-

rameter heranzieht, um die Lösungswahrscheinlichkeit für ein Item vorherzusagen.

Eine Modellgleichung, die den oben genannten Überlegungen Rechnung trägt, sollte folgende Eigenschaften aufweisen (Strobl, 2012, S. 7):

1. Die Fähigkeit der Person (θ_i) soll berücksichtigt werden.
2. Die Schwierigkeit der Aufgabe (β_j) soll berücksichtigt werden.
3. Je fähiger die Person, desto höher soll ihre Lösungswahrscheinlichkeit sein.
4. Da es sich um eine Wahrscheinlichkeit handelt, muss sie sich zwischen

den Grenzen 0 und 1 bewegen.

Eine Funktion, die diesen Anforderungen entspricht, ist die S-förmige logistische Funktion. Die unten dargestellte Formel beschreibt die Wahrscheinlichkeit, dass eine Aufgabe gelöst wird als logistische Funktion der Differenz zwischen Personenfähigkeit und Aufgabenschwierigkeit.

$$P(X_{ij} = 1 | \theta_i, \beta_j) = \frac{e^{\theta_i - \beta_j}}{1 + e^{\theta_i - \beta_j}} \qquad \text{(Formel 2)}$$

An der Formel ist zu erkennen, dass die Wahrscheinlichkeit, ein Item richtig zu lösen, davon abhängt, wie fähig eine Person im Verhältnis zur Schwierigkeit einer Ausgabe ist.

Die Eigenschaften der logistischen Funktion werden in ihrer grafischen Darstellung deutlich (vgl. Abbildung 9). Je mehr die Fähigkeit einer Person die Aufgabenschwierigkeit übersteigt, umso wahrscheinlicher ist es, dass sie die Aufgabe richtig löst. Die Lösungswahrscheinlichkeit liegt zwischen 0 und 1 und nähert sich diesen Werten asymptotisch an. Damit wird der Tatsache Rechnung getragen, dass auch bei einer Person, die so fähig ist, dass sie eine Aufgabe mit großer Wahrscheinlichkeit lösen sollte, nie eine 100 %ige Lösungswahrscheinlichkeit erreicht werden kann. Ebenso, wie man nicht davon ausgehen kann, dass eine leistungsschwache Person eine Aufgabe niemals lösen wird.

Die logistische Funktion wird im Rasch-Modell genutzt, um die Beziehung zwischen Personenfähigkeit und Lösungsschwierigkeit pro Item darzustellen. Die so entstehenden Grafen werden Item Characteristic Curves (ICCs) genannt (Bühner, 2011) (vgl. Abbildung 9). Die Schwierigkeit eines Items liegt per Definition an der Stelle auf dem Fähigkeitskontinuum, an der es mit einer 50 %igen Wahrscheinlichkeit gelöst wird. In Abbildung 9 sind drei Items mit unterschiedlichen Schwierigkeiten abgetragen. Das Item ganz links auf der x-

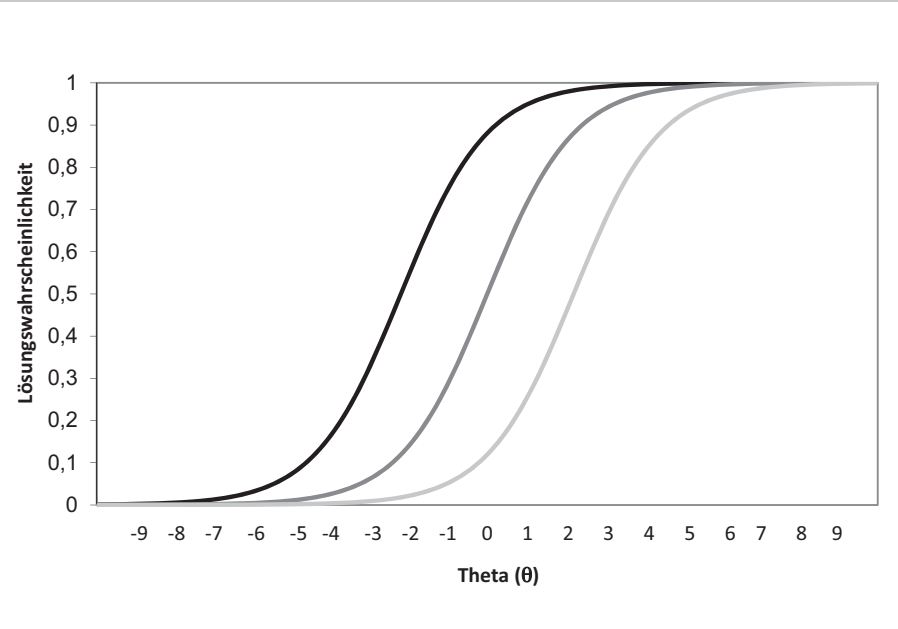


Abb. 9: Grafische Darstellung Kategorienfunktionen dichotomer Items nach dem Rasch-Modell *(Quelle: Bühner, 2011, S. 498)*

Achse (in schwarz) ist das leichteste, das Item ganz rechts auf der x-Achse (in hellgrau) ist das schwerste von den drei dargestellten Items. Da im Rasch Modell ein Item nur über den Schwierigkeitsparameter beschrieben wird, verlaufen alle ICCs eines Tests parallel zueinander. Das heißt, sie sind auf der x-Achse verschoben, unterscheiden sich jedoch nicht in der Steigung. Modelle, die neben der Schwierigkeit eines Items auch die Steigung der ICC miteinbeziehen, werden 2-PL Modelle genannt und hier nicht behandelt (siehe dafür Rost, 2004).

Damit die Daten eines Fragebogens mit dem Rasch-Modell ausgewertet werden können, müssen einige Vorannahmen erfüllt sein, die im Folgenden erläutert werden.

7.1.2.2 Modellgeltung und dessen Überprüfung

Das Rasch-Modell wurde aus bestimmten theoretischen Annahmen abgeleitet, die bei der Anwendung der oben vorgestellten Modellgleichung zur Testentwicklung und Auswertung überprüft werden können. Die Frage, ob ein Test-

modell auf die Daten passt, kann nicht deterministisch beantwortet werden (Rost, 2004). Vielmehr wird anhand zuvor per Konvention festgelegter Grenzen bestimmt, ob die Passung zwischen Modellannahmen und Daten ausreichend ist. Im Rahmen von Modelltests, in denen Signifikanzen ermittelt werden, muss berücksichtigt werden, dass in diesen die Nullhypothese die Forschungshypothese (H1) ist (Bühner, 2011). Das heißt, mit dem gängigen α-Niveu von 5 % ist damit zu rechnen, dass der β-Fehler unkontrolliert hoch ist. So kann man aufgrund des Nicht-Auffindens von Modellverletzungen nicht automatisch davon ausgehen, dass das Modell gilt (Kubinger, 2000). Vielmehr entspricht ein nicht-signifikanter Modelltest einer notwendigen Bedingung für die Modellpassung, jedoch nicht einer hinreichenden Bedingung. Trotz dieser grundsätzlichen Kritik an Modelltests sollte von einer probabilistischen Modellierung der Testdaten Abstand genommen werden, wenn eine Modellverletzung nachgewiesen wurde (Kubinger, 2000), zumal die Durchführung mehrer Modelltests den Bewährungsgrad eines Modells im Popperschen Sinne (Popper, 2009) erhöht (Kubinger, 2000).

Obwohl die Parameterschätzung den Modelltests zugrunde liegt (Rost, 2004), wird diese erst im Anschluss an die Darstellung der Modelltests behandelt, um zu verdeutlichen, dass die Prüfung der Modellannahmen der Interpretation der Parameter vorausgehen sollte.

Es gibt keinen etablierten Kanon der Modelltestung (Rost, 2004). Jedoch lassen sich aus der aktuellen Literatur (Bühner, 2011; Rost, 2004; Strobl, 2012) Empfehlungen ableiten, welche Tests durchgeführt werden sollten, bevor Schätzungen mit dem Rasch-Modell vorgenommen werden. Die Modellannahmen und Möglichkeiten deren empirischen Überprüfungen werden im Folgenden dargestellt.

Suffiziente Schätzer

Eine Statistik ist dann ein suffizienter Schätzer, wenn die gesamten Informationen, die die Daten über eine Person enthalten, in der Statistik enthalten sind (Strobl, 2012). Im Rasch-Modell werden für die Schätzung des Personenparameters die Zeilenrandsummen herangezogen und für die Schätzung der Aufgabenschwierigkeiten die Spaltenrandsummen. Das bedeutet, um zu einer Schätzung der Personenfähigkeit zu gelangen, muss nicht das individuelle Antwortmuster einer jeden Person betrachtet werden, sondern lediglich die Randsummen. Die Annahme, dass die jeweiligen Randsummen suffiziente Schätzer der Fähigkeit und Aufgabenschwierigkeit darstellen, lässt sich inhaltlich gut begründen: Eine fähige Person wird in einem Rasch-Konformen Test immer mehr Aufgaben lösen als eine weniger fähige Person. Genauso verhält es sich für die Anzahl der gelösten Aufgaben (Strobl, 2012).

Lokale stochastische Unabhängigkeit und Eindimensionalität

Lokale stochastische Unabhängigkeit setzt voraus, dass die Antworten auf die Items eines Rasch-homogenen Tests unbeeinflusst voneinander zustande kommen. Unter Kontrolle der Personenfähigkeit müssen alle Itemantworten unabhängig voneinander sein (Bühner, 2011). Das heißt, die Lösungswahrscheinlichkeit eines Items darf nicht die Lösungswahrscheinlichkeit eines anderen Items verändern (Strobl, 2012). Diese Bedingung wäre zum Beispiel verletzt, wenn die Lösung einer Aufgabe nur möglich ist, wenn eine vorhergehende Aufgabe richtig gelöst wurde (Strobl, 2012). Für die Aufgabenkonstruktion bedeutet das, jede Aufgabe muss unabhängig von den anderen Aufgaben im Test lösbar sein. Diese Annahme ist ebenso verletzt, wenn die Antwortwahrscheinlichkeiten durch mehr als nur eine latente Fähigkeitsdimension beeinflusst werden. In diesem Fall ist der Test nicht eindimensional und das Rasch-Modell ist nicht gültig (Bühner, 2011) und es sollten mehrdimensionale Modelle in Betracht gezogen werden (Strobl, 2012). Die Problematik der Eindimensionalität von situativen Aufgaben wurde bereits in Kapitel 5 diskutiert und in Kapitel 6 bei der Testkonstruktion berücksichtigt.

Auch für die Personen sind lokale statistische Unabhängigkeiten notwendig. Das heißt, dass die Wahrscheinlichkeit, mit der eine bestimme Person eine Aufgabe lösen kann, nicht systematisch davon abhängen darf, ob eine andere Person diese Aufgabe lösen kann (Strobl, 2012). Für die Testdurchführung bedeutet das, dass insbesondere darauf geachtet werden muss, dass wärend der Testung nicht abgeschrieben wird, da sonst die Annahme der lokalen statistischen Unanbhängigkeit verletzt wäre (Strobl, 2012).

Spezifische Objektivität

Mit dem Rasch-Modell, in dem die ICCs parallel verlaufen, also die Trennschärfen aller Items gleich sind, ist eine weitere messtheoretische Forderungen erfüllt, nämlich die spezifische Objektivität (Strobl, 2012).

Spezifische Objektivität bedeutet, dass unabhängig davon, welche Items vorgegeben werden, die Schätzungen der Fähigkeitsparameter zweier Personen immer die gleiche Differenz aufweisen. Ebenso sollten die Differenzen der Itemparameterschätzungen unabhängig von der Eigenschaftsausprägung der der Schätzung zugrunde gelegten Personengruppe sein. Diese Vorannahmen können durch Modelltests geprüft werden, jedoch gibt es keinen eindeutigen Ablauf der Modelltests (Rost, 2004). Die vorliegende Arbeit orientiert sich an Bühner (2011), der folgendes Vorgehen empfiehlt:

Im ersten Schritt der Modellgeltungskontrolle sollte der Andersen-Test (Andersen, 1973) durchgeführt werden. Dieser gehört zu den globalen Modellgel-

tungstests (Rost, 2004) und überprüft, ob spezifische Objektivität vorliegt. Die Stichprobe wird dafür in Teilstichproben untergliedert. In der Regel wird als Teilungskriterium der Median der Testrohwerte herangezogen (Bühner, 2011). Für die Teilstichproben und die Gesamtstichprobe wird die Likelihood errechnet, die mittels des Likelihoodquotienten an der χ^2-Verteilung auf Signifikanz geprüft wird. Überschreitet die Prüfstatistik den kritischen Wert, sollte das Rasch-Modell verworfen werden (Bühner, 2011).

Im Anschluss an die Prüfung der Modellpassung mittels Andersen-Test wird empfohlen, den Martin-Löf-Test durchzuführen, der eine Alternative zum Andersen-Test darstellt. Beim Martin-Löf-Test werden die Items nach einem oder mehreren Kriterien in zwei Hälften geteilt und ein modifizierter Likelihood-Quotienten-Test durchgeführt. Ein signifikantes Ergebnis weist darauf hin, dass die Items nicht homogen sind und somit eine grundlegende Annahme des Rasch-Modells verletzt ist. Alternativ oder ergänzend zum Martin-Löf-Test können explorative und konfirmatorische Faktoranalysen Hinweise auf Verletzungen der Eindimensionalität geben (Bühner, 2011).

Ein weiterer Test zur Überprüfung der spezifischen Objektivität ist der Wald-Test. Der Wald-Test basiert ebenfalls auf der Idee, dass sich die geschätzen Aufgabenparameter nicht stichprobenabhängig unterscheiden dürfen. Im Rahmen dieses Tests werden die Schätzungen der Aufgabenparameter direkt miteinander verglichen. Dabei können im Vergleich zum Andersen-Test immer nur die Werte von zwei Substichproben miteinander verglichen werden (Strobel, 2012). Die Teilung der Stichprobe erfogt in der Regel anhand der Kriterien „hoher vs. niedriger Rohwert", „jung vs. alt", „männlich vs. weiblich" sowie anhand weiterer inhaltlich relvanter Kriterien (Kubinger, 2000).

Der aufgabenspezifische Wald-Test prüft für jedes einzelne Item die Abweichung der Parameter, der globale Wald-Test prüft alle Aufgaben gleichzeitig auf Gruppenunterschiede. Über die bereits beschriebenen Tests hinaus gibt es die Möglichkeit, die Personenhomogenität der Parameterschätzung durch grafische Modelltests vorzunehmen (Bühner, 2011). Das Verfahren macht sich zunutze, dass Aufgabenparameter, die aus unterschiedlichen Substichproben geschätzt werden, bei vorliegender Personenhomogenität grafisch betrachtet auf einer Geraden auf der Winkelhalbierenden liegen müssen. Wie weit die Parameter von der Winkelhalbierenden abweichen dürfen, damit eine Modellverletzung angenommen wird, kann mit Unterstützung von zweidimensionalen Konfidenz-Regionen ermittelt werden (Strobl, 2012).

7.1.2.3 Parameterschätzung

Sind die Bedingungen für das Rasch-Modell erfüllt, kann die Parameterschätzung und insbesondere deren Interpretation stattfinden. Die Schätzung der un-

bekannten Parameter im Modell kann durch unterschiedliche Methoden vorgenommen werden. Eine häufige Schätzmethode beruht auf dem Maximum-Likelihood-Prinzip (Strobl, 2012). Die Maximum-Likelihood (ML)-Schätzung dient dazu, aus beobachteten Daten auf die unbekannten Parameter eines Modells zu schließen (Strobl, 2012). Die Ausgangsgleichung für die ML-Schätzverfahren im Rasch-Modell ist die zusammengefasste Darstellung der Wahrscheinlichkeit, dass eine Aufgabe gelöst wird und der Gegenwahrscheinlichkeit.

Für eine Person *i* und das Item *j* ist die Likelihood:

$$L_{uij}(\theta_i,\ \beta_j) = P(u_{ij}|\theta_i,\beta_j) = \frac{e^{u_{ij}*(\theta_i-\beta_j)}}{1+e^{\theta_i-\beta_j}} \qquad \text{(Formel 3)}$$

Um daraus die Likelihood für einer Person *i* über alle Aufgaben *j* hinweg zu erhalten, wird das Produkt über alle Aufgaben gebildet.

$$L_{ui}(\theta_i,\beta) = \prod_{j=1}^{m} \frac{e^{u_{ij}*(\theta_i-\beta_j)}}{1+e^{\theta_i-\beta_j}} \qquad \text{(Formel 4)}$$

Da der Nullpunkt der Skala im Rasch-Modell frei verschiebar ist, muss eine willkürliche Normierungsbedingung festgestellt werden (Strobl, 2012). In der Statistiksoftware ConQuest, die in der vorliegenden Arbeit zur Rasch-Skalierung herangezogen wird, wird der Schwierigkeitsparameter des letzten Items so festgelegt, dass die Summe aller Itemparameter Null ergibt (Wu, Adams & Wilson, 1998). Die Parameterschätzung in ConQuest wird über eine marginale ML-Schätzung vorgenommen (Wu et al., 1998), das bedeutet, dass im ersten Schritt der Schätzung die Personenparameter über die Annahme einer marginalen Verteilung der Personenparameter „rausgerechnet" werden (Strobl, 2012).

Neben der ML-Schätzung liegen noch weitere Schätzverfahren vor, wie zum Beispiel die Weighted Likelihood Schätzung (WLE), die speziell für Daten mit Extremwerten empfohlen wird. Da es sich bei der Maximum-Likelihood-Schätzung um ein etabliertes Schätzverfahren handelt, wird dieses in den Berechnungen dieser Arbeit verwendet.

Die Raschskalierbarkeit eines Tests ist ein wichtiges Merkmal seiner psychometrischen Qualität (Bühner, Ziegler, Krumm & Schmidt-Atzert, 2006; Kubinger, 2000). Darüber hinaus werden Kriterien der klassischen Testtheorie zur Beurteilung der Güte von Tests herangezogen. Probabilitische und klassische Testtheorien sollten bei der Beurteilung von Tests nicht als konkurrierend angesehen werden, sondern als komplementär (Kubinger, 2000; Rost, 2004).

Neben den oben beschriebenen Modellen liegen zahlreiche weitere Modelle zur Testauswertung vor. In latenten Klassenanalysen werden Personen entsprechend bestimmter Antwortmuster einer latenten Klasse zugeordnet (Rost, 2004). Die Klassenzuordnung erfolgt jedoch nicht deterministisch, sondern probabilistisch. Eine Weiterentwicklung ist das mixed Rasch-Modell in dem Daten sowohl klassifiziert als auch quantifiziert werden (Rost, 2004). Die Quantifizierung wird innerhalb der Klassen mittels Rasch-Modell vorgenommen. Die möglichen Einsatzgebiete von mixed Rasch-Modellen in der Bildungsforschung werden im Rahmen der Diskussion in Abschnitt 10.2 erläutert. Für eine Beschreibung der Verfahren wird auf Rost (2004) Kapitel drei verwiesen.

7.2 Kriterien der Beurteilung der Güte von Tests

In der Pädagogischen Psychologie spielen Gütekriterien sowohl bei der Entwicklung neuer Tests als auch bei der Bewertung bestehender Tests eine tragende Rolle. Die zentrale Bedeutung, die Gütekriterien zuteilwird, spiegelt sich in der Präsenz der Thematik in Standardlehrwerken (z. B. Bühner, 2011; Ingenkamp, 1997; Rost, 2004) sowie durch die feste Verankerung der Gütekriterien in den Standards psychologischer und pädagogischer Diagnostik (AERA et al., 2008) wider. Wie in der Einleitung dargestellt (vgl. Abschnitt 1.2), ist die Frage nach den Gütekriterien des entwickelten Tests eine zentrale Fragestellung, die in der vorliegenden Arbeit beantwortet werden soll. Die drei wichtigsten Gütekriterien Objektivität, Reliabilität und Validität werden im Folgenden dargestellt. Darüber hinaus wird im letzten Abschnitt die Testfairness als wichtiges Nebengütekriterium betrachtet.

7.2.1 Objektivität

Eine grundlegende Anforderung an diagnostische Tests mit Anspruch auf Wissenschaftlichkeit ist, dass sie objektiv sind. Objektivität beschreibt den Grad, in dem die Ergebnisse eines Testes unabhängig von der Person des Testleiters sind (Bühner, 2011, S. 58). Die Objektivität eines Tests kann nur begrenzt durch statistische Kennwerte repräsentiert werden. In der Regel erfolgt die Prüfung und Einschätzung der Objektivität anhand von sachlogischen Überlegungen, die im Folgenden ausgeführt werden.

Um einen hohen Grad an Objektivität zu gewährleisten, muss sowohl während der Durchführung eines Tests als auch während der Auswertung auf bestimmte Vorgaben geachtet werden. Um die sogenannte *Durchführungsobjektivität* (Moosbrugger & Kelava, 2012a) zu gewährleisten, muss vor der Testdurchführung genau definiert werden, unter welchen Bedingungen die Testung stattfinden soll. Zu den zentralen Bedingungen einer Erhebung gehören der Zeitpunkt

der Durchführung und die Dauer und die Art der Hilfestellungen, die während der Erhebung zur Verfügung stehen. Um einen hohen Grad an Objektivität zu erreichen, sollte der Einfluss der anwesenden Testleitung minimiert werden. Ein wichtiges Werkzeug zur Schaffung einer objektiven Testsituation sind detailliert ausgearbeitete Instruktionen für die Testteilnehmer. Durch standardisierte Instruktionen wird sichergestellt, dass allen Testteilnehmern dieselben Informationen zur Verfügung stehen und keine Verzerrungen entstehen, weil zum Beispiel ein Testleiter besonders hilfsbereit ist und bestimmten Testteilnehmern einen ungewollten Vorteil verschafft.

Ein weiterer Aspekt der Bewertung von Tests ist die *Auswertungsobjektivität* (Moosbrugger & Kelava, 2012a). Um diese zu gewährleisten, sollten klare Regeln bestehen, wie Aufgabenlösungen bewertet werden. Weitgehend unproblematisch ist die Auswertungsobjektivität bei geschlossenen Aufgabentypen. Bei diesem Aufgabentyp stehen unterschiedliche Antwortalternativen zur Verfügung, die durch Ankreuzen gekennzeichnet werden. Es ist für die Testauswertung leicht, einen eindeutigen Auswertungsschlüssel zu erstellen, den alle an der Auswertung beteiligten Personen nutzen können. Bei konsequenter Anwendung desselben Schlüssels müssten unterschiedliche Testauswerter bei ein und demselben Test immer auf identische Testresultate kommen. Beeinträchtigungen der Auswertungsobjektivität können in diesem Fall lediglich durch Flüchtigkeitsfehler des Testauswerters entstehen. Diese lassen sich durch übersichtliche Auswertungsschablonen und/oder Korrekturen durch einen anderen Testauswerter leicht reduzieren. Kritischer ist die Einhaltung der Auswertungsobjektivität bei sogenannten offenen Antwortformaten. Zu dieser Gruppe von Aufgabentypen gehören zum Beispiel Kurzantworten oder Aufsätze. Bei dieser Art von Aufgabe haben die Testteilnehmer mehr Freiheiten, ihre Antwort auf eine Frage auszugestalten. Für die Testauswertung ist die einheitliche Bewertung der heterogenen Antworten eine große Herausforderung.

Die Auswertungsobjektivität des vorliegenden Tests ist als hoch einzustufen, da alle Aufgaben mit geschlossenem Antwortformat gestellt wurden und der Kodierschlüssel zuvor unter Mitwirkung von verschiedenen Experten festgelegt wurde (siehe Itementwicklungsprozess in Anhang B, Abbildung B-1).

Sobald Testwerte herangezogen werden, um daraus weiterführende Schlüsse über die Testperson zu ziehen, muss eine weitere Dimension der Objektivität betrachtet werden: die *Interpretationsobjektivität* (Moosbrugger & Kelava, 2012a). Zur Einhaltung der Interpretationsobjektivität muss sichergestellt werden, dass gleiche Punktzahl in einem Test zu den gleichen Beurteilungskonsequenzen führt.

7.2.2 Reliabilität

Die Zuverlässigkeit oder *Reliabilität* wird als die Verlässlichkeit eines Tests beschrieben. Sie gibt an, wie genau ein Test misst (Bühner, 2011, S. 60). Ein Test liefert ungenaue Messungen, wenn Störeinflüsse das eigentliche Ergebnis einer Testung überlagern. Störeinflüsse können ganz unterschiedlicher Natur sein, zum Beispiel missverständlich formulierte Aufgaben oder erschwerte Lesbarkeit wichtiger Aufgabenelemente (z. B. durch schlechte Kopierqualität der Aufgabenblätter). Die Beispiele zeigen, dass im Vorfeld einer Testung Maßnahmen getroffen werden können, die eine reliable Testung begünstigen. Die tatsächliche Reliabilität eines Tests lässt sich jedoch erst bestimmen, wenn eine Mindestanzahl an Personen den Test ausgefüllt hat. Es gibt unterschiedlicher Verfahren der Reliabilitätsbestimmung: das Verfahren der Retest-Reliabilitätsbestimmung, die Paralleltestmethode, die Testhalbierungsmethode und die Bestimmung der internen Konsistenz. Allen Verfahren ist gemein, dass sie den Anteil der wahren Varianz an der beobachtbaren Varianz eines Tests schätzen (Bortz, 1999). Im Bereich der berufs- und wirtschaftspädagogischen empirischen Forschung hat sich die Berechnung der internen Konsistenz als Methode der Reliabilitätsbestimmung etabliert.

Die interne Konsistenz ist in den Grundüberlegungen eine Weiterentwicklung der Reliabilitätsbestimmung durch Testhalbierung. Der Test wird bei diesem Verfahren in so viele Teilstücke unterteilt, wie der Test Items hat. Für diese Items wird dann das Zusammenhangsmaß Cronbachs Alpha (Cronbach, 1951) ermittelt, welches die Messgenauigkeit des Tests zum Zeitpunkt der Erhebung der Daten angibt (Bühner, 2011, S. 61). Wie bereits erwähnt, wird Cronbachs Alpha häufig als Kennwert für die Reliabilität von Tests oder Subskalen publiziert. Die Interpretation der Kennwerte unterliegt jedoch keinen strengen Regeln, sondern Konventionen, die in der Forschungspraxis unterschiedlich ausgelegt werden. Gemeinhin gilt als Richtwert, dass Skalen mit einem Alpha > .70 als reliabel gelten (Peterson, 1994). Für Tests in ihrer Entwicklungsphase gelten teilweise Reliabilitäten ab .60 als akzeptabel (Peterson, 1994). Eine Interpretationshilfe für den Kennwert Cronbachs Alpha ist in Tabelle 11 dargestellt.

Tab. 11: Interpretationshilfe für den Kennwert Cronbachs Alpha

α	**Bedeutung**
> 0.9	exzellent
> 0.8	gut
> 0.7	akzeptabel
> 0.6	fragwürdig

(Fortsetzung Tab. 11)

α	**Bedeutung**
> 0.5	schlecht
≤ 0.5	inakzeptabel

(Quelle: George & Mallery, 2000 S. 231)

Zu beachten sind bei der Interpretation von Cronbachs Alpha zwei zentrale Punkte, (1) Cronbachs Alpha ist kein angemessenes Maß für die *Eindimensionalität* einer Skala, (2) steigt die Höhe des Kennwertes mit der Anzahl homogener Items. Das bedeutet, Cronbachs Alpha ist nur dann ein verlässliches Maß für die Reliabilität eines Tests oder einer Skala, wenn zuvor die Eindimensionalität der Itemgruppe abgesichert wurde und eine angemessene Anzahl von Items zur Verfügung steht. Dieser Zusammenhang ist der Grund dafür, dass eine Verlängerung einer Skala durch mehrere inhaltlich homogene Items die Reliabilität eines Tests erhöht (Bühner, 2011).

In den in Abschnitt 7.1.2 vorgestellten probabilistischen Testtheorien spielt das Konzept der Messgenauigkeit im Sinne der internen Konsistenz eine untergeordnete Rolle. Es hat sich jedoch etabliert, als Maß der Präzision der Personenparameterschätzung einen EAP/PV Koeffizienten anzugeben (z. B. Eggert & Bögeholz, 2010; Schöps, Senkbeil & Schütte, 2009). Der EAP/PV Koeffizient wird von der Software ConQuest (Wu et al., 1998) ausgegeben und ergibt sich aus dem Quotienten der durch das Modell erklärten Varianz und der Personenvarianz (Draney & Wilson, 2008). Seine Interpretation orientiert sich an den Richtwerten für Cronbachs Alpha (vgl. Tabelle 11).

Die Messgenauigkeit eines Tests ist die Grundlage für die im folgenden Abschnitt beschriebene Validität. Einige Autoren beschreiben die Validität als das wichtigste der Gütekriterien, da ihr eine wichtige Rolle bei der Interpretation von Forschungsergebnissen zukommt. Beachtet werden sollte, dass valide Ergebnisse nur mit objektiven und reliablen Testinstrumenten erzielt werden können und somit keines der Gütekriterien vernachlässigt werden kann. Die Validierung des entwickelten Tests ist ein zentrales Ziel der vorliegenden Arbeit (vgl. Abschnitt 1.1), deshalb wird das Konzept der Validität und Möglichkeiten deren Prüfung im folgenden Abschnitt ausführlich erläutert.

7.2.3 Validität

Die Gültigkeit oder *Validität* eines Tests gibt an, ob ein Test tatsächlich das misst, was er zu messen vorgibt. Die Validität gilt aus Sicht vieler Forscher als das wichtigste methodische Kriterium für Testverfahren (Ingenkamp, 1997). Ohne eine hinreichende Absicherung der Validität von Forschungsergebnissen

kann es zu fundamentalen Fehlinterpretationen von Testergebnissen kommen. Trotz dieser Wichtigkeit gibt es keinen einfachen und eindeutigen Weg der Validitätsfeststellung. Vielmehr gibt es unterschiedliche Arten der Validität, die jeweils unterschiedlich ermittelt werden. Generell gilt, dass die Absicherung der Validität durch die Prüfung möglichst vieler verschiedener Validitätsarten stattfinden sollte.

7.2.3.1 Augenscheinvalidität

Die *Augenscheinvalidität* (Moosbrugger & Kelava, 2012a) ist eine der intuitiv plausibelsten Validitätsarten. Sie bezeichnet, inwiefern Laien und Fachmänner per Augenschein, ohne tiefergehende Analysen, die Messintention eines Testes erkennen können und ob sich der subjektive Eindruck der Items mit dem intendierten Konstrukt deckt. Die Augenscheinvalidität ist naturgemäß bei manifesten und klar umrissenen Konstrukten besonders hoch. Je komplexer das zu messende Konstrukt ist, umso weniger Bedeutung kann der Augenscheinvalidität als Gütekriterium zugesprochen werden. Bei umfassenden Intelligenztests ist es zum Beispiel für den Laien deutlich schwerer zu erkennen, welche Intelligenzfacette gemessen werden soll. In einigen Persönlichkeitstests soll sich der Testteilnehmer bestenfalls nicht im Detail darüber bewusst sein, was gemessen wird. Dies ist zum Beispiel der Fall bei Integritätsfragebögen, die bei der Personalauswahl eingesetzt werden (Marcus, Funke & Schuler, 1997). Im Bereich der Diagnostik vom Lernergebnissen und Lernprozessen an Hochschulen ist es jedoch in den meisten Fällen wichtig, eine hohe Augenscheinvalidität des Tests zu gewährleisten. Diese erlaubt allen an einer Untersuchung direkt oder indirekt beteiligten Akteuren (Studierende, Dozenten und Professoren) zu erkennen, was wie getestet wird. In den meisten Fällen führt eine solche Transparenz zu einer größeren Testakzeptanz und bei den Testteilnehmern ist dies eine wichtige Grundlage für die Testmotivation (Kersting, 2008). Die Augenscheinvalidität von situativen Aufgaben wird in der Regel als hoch bewertet, da der direkte Anwendungsbezug deutlich wird (Möller, 2010). Augenscheinvalidität wird häufig mit erhöhter Testakzeptanz in Verbindung gebracht. Weitere Facetten der Akzeptanz von Beurteilungsverfahren sind Belastungsfreiheit, Messqualität und Kontrollierbarkeit (Kersting, 2008).

Ein wichtiges Kriterium der Güte von Tests, das mit der Augenscheinvalidität zusammenhängt, ist die im folgenden Abschnitt beschriebene Inhaltsvalidität (Bühner, 2011).

7.2.3.2 Inhaltsvalidität und curriculare Validität

Die *Inhaltsvalidität* (Moosbrugger & Kelava, 2012a) ist ein schwer quantitativ zu erfassendes Kriterium (Ingenkamp, 1997). Es wird von inhaltlicher Validität

gesprochen, wenn ein Test inklusive aller seiner Items das zu messende Merkmal präzise erfasst. Dabei ist nicht die Messgenauigkeit im Sinne der statistischen Reliabilität gemeint, sondern dass die Items das interessierende Merkmal im Sinne einer inhaltlichen Repräsentanz abbilden. Die Items sollen also für merkmalstypische Verhaltensweisen repräsentativ sein. Da ein Merkmal in den meisten Fällen über nahezu unendlich viele Items abbildbar ist, ist es besonders schwierig zu beurteilen, ob ein Test eine angemessen repräsentative Itemmenge enthält (Bühner, 2011). Obwohl es für die meisten Merkmale wahrscheinlich nie möglich ist, eine perfekte inhaltliche Validität herzustellen, sollte bei der Testentwicklung immer zumindest der Versuch unternommen werden, dem Anspruch der Inhaltsvalidität gerecht zu werden. Hilfreich sind dabei die Beschreibung der Inhalte des Konstrukts, die Zuordnung von Items zu entsprechenden Inhaltsbereichen und ein Vergleich der empirischen Faktorstruktur mit der Struktur des Tests (Murphy & Davidshofer, 2001 S. 150 zit. nach Bühner, 2011). Zudem ist es üblich, für die Beurteilung der Inhaltsvalidität Expertenmeinungen heranzuziehen.

Ein Spezialfall der Inhaltsvalidität ist die *curriculare Validität* (Hartig et al. 2012), die insbesondere für fachspezifische Leistungstests und somit auch für die Diagnostik beruflicher Lernprozesse eine hervorgehobene Rolle spielt. Bei der curricularen Validität geht es darum, dass Aufgaben in fachspezifischen Leistungstests die Inhalte des Curriculums bestmöglich abbilden sollen. Hier ist es sinnvoll, Experten zu Rate zu ziehen, die beurteilen können, welche Aufgaben sich tatsächlich auf das Curriculum beziehen und welche kognitiven Anforderungen mit den Aufgaben einhergehen. Aber auch Experten kommen diesbezüglich selten zu einem eindeutigen Konsens. In jüngerer Zeit wird bei der Erstellung von schulischen Curricula darauf geachtet, dass genaue Angaben über die Inhalte und Lernziele eines Fachs gemacht werden (Ingenkamp, 1997). Die curriculare Validität wurde in der vorliegenden Arbeit sowohl durch curriculare Analysen (vgl. Kapitel 6) als auch durch die Beurteilung von Experten und Studierenden abgesichert (vgl. Abbildung B-1 zum Itementwicklungsprozess in Anhang B).

7.2.3.3 Konstruktvalidität

Häufiger als die inhaltliche Validität wird die *Konstruktvalidität* (Moosbrugger & Kelava, 2012a) als Gütemerkmal zur Bewertung von Tests herangezogen. Ein Test ist konstruktvalide, wenn er tatsächlich das Konstrukt misst, das er zu messen vorgibt. Sofern ein theoretisches Konstrukt des gemessenen Merkmals vorhanden ist, wird die Konstruktvalidität nach der Erhebung der Daten geprüft, im Gegensatz zur inhaltlichen Validität, die schon während der Phase der Test- und Itementwicklung durch detaillierte Recherchen und Exper-

tenurteile geprüft wird. Die Überlegungen der Konstruktvalidierung basieren auf vorab theoretisch begründeten Vermutungen, die über den Zusammenhang des vorliegenden Testes mit konstruktverwandten (konvergenten) und konstruktfremden (diskriminanten oder divergenten) Tests formuliert werden (Hartig, Frey & Jude, 2012). Diese Vermutungen lassen sich über die Errechnung von Korrelationskoeffizienten prüfen. Tests, die ähnliche Konstrukte erfassen, sollten hoch miteinander korrelieren. Tests, die unterschiedliche Konstrukte erfassen, sollten möglichst gering miteinander korrelieren. Trifft die erste Anforderung zu, spricht man von *konvergenter Validität*. Trifft die zweite Anforderung zu, wird von *diskriminanter* oder *divergenter Validität* gesprochen (Moosbrugger & Kelava, 2012a). Eine Möglichkeit, Ergebnisse der konvergenten und diskriminanten Validierungsbemühungen zu systematisieren, ist der Multitrait-Multimethod-Ansatz in dem mehrere Konstrukte mit unterschiedlichen Methoden gemessn und systematisch miteinander verglichen werden (Campbell & Fiske, 1959). Darüber hinaus besteht ein faktoranalytischer Zugang zur Ermittlung der Konstruktvalidität, der von Bühner (2011) beschrieben wird und in der vorliegenden Arbeit Anwendung findet (vgl. Abschnitt 9.3.2).

7.2.3.4 Kriteriumsvalidität

Ähnlich wie bei der Konstruktvalidierung wird bei der kriteriumsbezogenen Validierung die korrelative Beziehung zu bedeutungsvollen Außenkriterien geprüft. Schulnoten könnten zum Beispiel als Kriterium für zentrale Schulleistungstests herangezogen werden oder anders herum. Es wird, je nach Beschaffenheit des Kriteriums, zwischen verschiedenen Arten der Kriteriumsvalidität differenziert. Die *Vorhersagevalidität* (auch *prognostische* oder *prädiktive Validität*) (Moosbrugger & Kelava, 2012a) prüft, ob ein Test Vorhersagen über zukünftige Entwicklungen der Testperson erlaubt (Hartig et al., 2012), zum Beispiel, ob die Abiturnote die Note des Studienabschlusses vorhersagt. Prognostische Validität ist insbesondere dann von großer Bedeutung, wenn auf Grundlage der Testwerte Auswahlentscheidungen getroffen werden, wie zum Beispiel in der Personalauswahldiagnostik. Der Begriff *Übereinstimmungsvalidität* oder *konkurrente Validität* bezieht sich auf den Zeitpunkt, zu dem das Kriterium erhoben wird (Moosbrugger & Kelava, 2012a). Werden Test und Kriterium zur gleichen Zeit erhoben, spricht man von konkurrenter Validität. Liegt das Kriterium zeitlich vor dem eigentlichen Test, so spricht man von *retrospektiver Validität* (Moosbrugger & Kelava, 2012a). *Inkrementelle Validität* (Hartig et al., 2012) bezeichnet die Forderung, dass Tests, über bestehende Tests hinaus, einen Beitrag zur Vorhersage bestimmter Kriterien leisten sollen. In der vorliegenden Arbeit werden zeitgleich mit der Erhebung des Wissenskonstruk-

tes Kriterien zur Validierung erhoben, die zuvor in Kapitel 3 als relevant identifiziert wurden und zu denen in Abschnitt 9.1.2 Hypothesen abgeleitet werden.

Drei zentrale Gütekriterien sozialwissenschaftlicher Messungen wurden im vorangegangenen Abschnitt einführend beschrieben. Neben diesen drei Hauptkriterien gibt es zahlreiche Nebengütekriterien, die je nach diagnostischem Anwendungsfall unterschiedlich wichtig sind. Die gängigsten zwei Nebengütekriterien werden im folgenden Abschnitt vorgestellt.

7.2.4 Testfairness und Testökonomie

Neben den etablierten Gütekriterien (Objektivität, Reliabilität und Validität) rücken sogenannte „Nebengütekriterien" immer mehr in das Licht der Aufmerksamkeit der pädagogisch-psychologischen Forschung. In der pädagogisch-psychologischen Diagnostik spielt die *Testfairness* (Moosbrugger & Kelava, 2012a) eine tragende Rolle. Eine grundlegende Forderung an einen fairen Test ist das Fehlen von systematischem Item Bias. Item Bias liegt dann vor, wenn identifizierbare Subgruppen in einem Test oder in spezifischen Aufgaben eines Tests unabhängig von Ihrer Fähigkeit besser oder schlechter abschneiden als die Vergleichsgruppe (AERA et al., 2008; Osterlind, 1989). Damit ist nicht gemeint, dass Subgruppen einer Stichprobe zwangsläufig immer die gleichen Ergebnisse haben müssen. Vielmehr darf die Wahrscheinlichkeit, ein Item zu lösen oder nicht zu lösen, nicht von gruppenspezifischen Merkmalen abhängen. Probabilistische Testtheorien bieten die Möglichkeit, über die Identifikation von Differential Item Function (DIF) aufzudecken, welche Items subgruppenspezifische item characteristic curves (ICCs) aufweisen (Osterlind & Everson, 2009). In Untersuchungen zu DIF werden soziale Gruppenunterschiede im Bereich Geschlecht, ethnische Zugehörigkeit und Klassenzugehörigkeit analysiert (Osterlind & Everson, 2009). Weist ein Test bedenkliches DIF auf, so kann dieser nicht für die betreffenden Subgruppen angewendet werden (AERA et al., 2008). Laut Osterlind und Everson (2009) gibt es kaum Tests, die kein DIF aufweisen, was eine eindeutige Bewertung von DIF im Rahmen der Testfairness erschwert. Neben der Vermeidung von DIF ist im Rahmen der Testfairness sicherzustellen, dass der Testprozess für alle Testteilnehmer identisch ist (AERA et al., 2008). In der vorliegenden Arbeit werden unter dem Rach-Modell differenzierte DIF-Analysen durchgeführt (vgl. Abschnitt 9.3.4.1) sowie Facetten der Testakzeptanz betrachtet (vgl. Abschnitt 9.3.4.2).

Die *Testökonomie* bezieht sich auf die Frage, ob ein Testinstrument ein relevantes Merkmal sparsam, ohne große zusätzliche Kosten (wie Zeit, Geld oder andere Ressourcen) erfasst (Fiege, Reuther & Nachtigall, 2011). Die Reduzierung zeitlichen und finanziellen Aufwands darf jedoch nicht zu Lasten der Hauptgütekriterien gehen (Moosbrugger & Kelava, 2012a). In der vorliegenden

Arbeit wird zum Beispiel unter Abwägen von Aspekten der Testökonomie vorerst auf eine computerbasierte Umsetzung der situativen Items verzichtet (vgl. Abschnitt 5.2).

Nachdem die Standards der pädagogisch-psychologischen Diagnostik aufgezeigt wurden, folgt die Beschreibung der Pilotstudie und deren Konsequenzen für die Haupterhebung.

8 Testpilotierung

Die 45 Items, die nach den in Kapitel 5 beschriebenen Prinzipien konstruiert wurden, wurden vor ihrem Einsatz in der Haupterhebung einer Pilotierung unterzogen. Eine Pilotierung dient vornehmlich der Absicherung der wissenschaftlichen Güte neu entwickelter Items. Ziel einer Pilotstudie ist es, diejenigen Items zu identifizieren, die nicht den Standards pädagogisch-psychologischer Diagnostik entsprechen (vgl. Kapitel 7), und über die Selektion dieser Items einen messgenauen und validen Test zusammenzustellen. Items, die den empirischen Kriterien nicht entsprechen, werden entweder aussortiert oder überarbeitet. Für die Auswertung von Testdaten können zwei unterschiedliche testtheoretische Zugänge gewählt werden (vgl. Kapitel 7):

(1) die klassische Testtheorie, auf der ein Großteil der in Deutschland veröffentlichten standardisierten Fragebögen beruht (Rost, 2004), und

(2) die Gruppe der probabilistischen Testtheorien, die die aktuelle Literatur im Bereich Psychometrie und Testtheorie dominieren (Rost, 2004).

Seit dem erfolgreichen Einsatz von probabilistischen Testmodellen im Rahmen der TIMS-Studien (Martin, Gregory, Stemler & Foy, 2000) und der PISA-Studien (Klieme, 2010) ist deren Verwendung für die Entwicklung und Auswertung neuer und bestehender Tests im Bereich der empirischen Bildungsforschung zum Standard geworden (Hartig & Frey, 2012). In der vorliegenden Arbeit werden beide testtheoretischen Zugänge im Sinne Rosts (2006) nicht als konkurrierende Ansätze, sondern als komplementäre Methoden betrachtet. Dementsprechend werden im Rahmen der Auswertung dieser Pilotierung Kennwerte aus beiden Testtheorien herangezogen, um die Items und ihre Eignung für diagnostische Zwecke umfassend zu beurteilen. Zu Beginn des Kapitels wird der Aufbau der Studie beschrieben (Abschnitt 8.1). Es folgt die Darstellung der empirischen Ergebnisse (Abschnitt 8.3). Das Kapitel schließt mit einer Zusammenfassung der Konsequenzen, die sich aus den Ergebnissen der Pilotierung für die Hauptuntersuchung ergeben.

8.1 Zielsetzung der Pilotierung

Die Pilotierung dient in erster Linie der Testoptimierung durch Itemselektion sowie der Überprüfung der Gütekriterien (vgl. Kapitel 7). Über die Reliabilität des Tests wird keine Hypothese aufgestellt. Es gilt jedoch sicherzustellen, dass der Test den in der empirischen Bildungsforschung etablierten Kennwert von Cronbachs Alpha = .7 (Bühner, 2006) erreicht.

Die Validität eines Tests lässt sich nicht durch einen einfachen Kennwert abbilden (Hartig & Frey, 2012). Vielmehr setzt sich die Bewertung der Validität aus unterschiedlichen Überlegungen und empirischen Herangehensweisen zusammen. Diese Arbeit orientiert sich am Vorgehen von Förster et al. (2012) und untersucht im Rahmen erster Validierungsuntersuchungen die Zusammenhänge zwischen der Testleistung und einer Auswahl von Außenkriterien, die mit Aspekten der Validität assoziiert sind (vgl. Kapitel 3). Im Rahmen der Pilotierung werden keine statistischen Hypothesen aufgestellt, jedoch Vermutungen über die jeweiligen Zusammenhänge formuliert. Eine detailierte Ausarbeitung der vermuteten Zusammenhänge ist unter Jähnig (2013) dokumetiert. In der Hauptuntersuchung werden die Hypothesen zur Testvalidität aus dem bisherigen Forschungsstand abgeleitet und statistisch geprüft.

8.2 Testdurchführung, Testmaterial und Stichprobenbeschreibung

Die Datenerhebung für die Pilotierung fand im Sommersemester 2012 an der Georg-August-Universität in Göttingen statt. Die Rekrutierung der Testteilnehmer für die Pilotierung erfolgte über die Dozenten wirtschaftswissenschaftlicher und wirtschaftspädagogischer Lehrveranstaltungen. Es wurden gezielt Lehrveranstaltungen ausgewählt, die von Studierenden am Ende des Bachelorstudiums und zu Beginn des Masterstudiums besucht werden. Die Erhebungen wurden nach Vorankündigung in den entsprechenden Lehrveranstaltungen durchgeführt. Während der gesamten Erhebung war mindestens eine Person mit der Testleitung betraut. Die Testleitung wurde immer von derselben Person durchgeführt, die keine inhaltlichen Hilfestellungen gewährte und im Rahmen der Möglichkeiten Kommunikation zwischen den Studierenden während der Testung unterband. Als technisches Hilfsmittel wurde allen Teilnehmern die Nutzung eines Taschenrechners gewährt. Zu Beginn der Erhebung wurden standardisierte Kurzinstruktionen vorgelesen, die in erster Linie der Aufklärung der Studierenden über den Untersuchungszweck galten. In zweiter Linie wurde beabsichtigt, die Testmotivation der Testteilnehmer durch eine persönliche Ansprache zu steigern. In einer Erhebungsgruppe mit ca. 20 Testteilnehmern wurde die persönliche Ansprache aus organisatorischen Gründen per Video übermittelt. Die Testzeit war auf den Umfang der Lehrveranstaltung begrenzt. Die durchschnittliche Bearbeitungszeit eines Testheftes betrug 60 Minuten.

Der Test wurde als Papierversion umgesetzt. Jedes Testheft wurde zusammen mit einer schriftlichen Instruktion an die Testteilnehmer ausgeteilt. Die Instruktion der Pilotierung wurde mit minimalen Änderungen in die Haupterhebung

übernommen und ist in Anhang C dokumentiert. Eine verkürzte mündliche Instruktion wurde bei jeder Erhebung durch die Testleitung vorgetragen. Das Testheft war in vier Abschnitte unterteilt und umfasste 16 doppeltbedruckte Seiten. Zuerst wurden 45 situative Aufgaben dargeboten (für einen Überblick über die Iteminhalte siehe Tabelle 10). Anschließend wurden Informationen zur Person abgefragt (Alter, Geschlecht, Muttersprache und elterlicher Bildungshintergrund). Es folgte ein Abschnitt mit Fragen zum Bildungsweg, in dessen Rahmen Bildungserfahrungen vor Beginn des Hochschulstudiums erfragt wurden (Art und Abschlussnote der Hochschulzugangsberechtigung, Fragen zu einer eventuellen kaufmännischen Berufsausbildung und kaufmännischen Praktika). Der letzte Fragenblock bezog sich auf das aktuelle Studium. Er enthielt Fragen zum Studiengang, Studienabschnitt, Studienortwechsel, bereits absolvierten betriebswirtschaftlichen Modulen, Studieninteresse und verschiedenen kaufmännischen Nebentätigkeiten während des Studiums.

An der Pilotierung nahmen 154 Studierende der Universität Göttingen teil. Die Testteilnehmer waren zum Zeitpunkt der Erhebung im Mittel 24 Jahre alt (*SD* = 2.64). 65 der Testteilnehmer waren männlich und 82 weiblich (sieben Personen gaben ihr Geschlecht nicht an). 89 Studierende gaben an, in einem betriebswirtschaftlichen Studium eingeschrieben zu sein, 56 im Studiengang Wirtschaftspädagogik. Um ein möglichst heterogenes Leistungsspektrum abzubilden, wurden neben Bachelorstudierenden auch Studierende am Anfang ihres Masterstudiums rekrutiert. Da es sich um eine Gelegenheitsstichprobe handelte, sind die Zellenbesetzungen zwischen den Studiengängen und den Studienabschnitten nicht ausgeglichen. In Tabelle 12 wird die Zusammensetzung der Stichprobe beschrieben.

Tab. 12: Beschreibung der Pilotierungsstichprobe nach Studiengang und Studienabschnitt (N = 154)

Studiengang	**Studienabschnitt**	
	Bachelor	**Master**
BWL	72	14
Wirtschaftspädagogik	45	11
VWL	3	0
Gesamt[10]	145	

Eine Analyse fehlender Werte für die 45 situativen Aufgaben ergab, dass die Testhefte weitgehend vollständig ausgefüllt wurden. Der mittlere Anteil fehlender Werte pro Aufgabe lag bei 2.6 % (*SD* = 2,08), wobei keine Aufgabe

10 9 Testteilnehmer machten keine Angaben zu ihrem Studiengang oder Studienabschnitt.

mehr als 10 % fehlende Werte generierte. 112 Testteilnehmer beantworteten alle Fragen. Alle weiteren Personen haben mind. 70 % der Fragen beantwortet. Fehlende Werte in dieser Größenordnung wurden als Indikator für das Nicht-Wissen der Lösung einer Aufgabe interpretiert und dementsprechend für weitere Berechnungen als falsch kodiert. Die Weiterverarbeitung der Daten erfolgte mit den Statistikprogrammen Statistical Package for the Social Sciences (SPSS), ConQuest (Wu et al., 1998) und R (R Core Team, 2012) und wird im folgenden Abschnitt beschrieben.

8.3 Ergebnisse der Pilotierung

Ein zentrales Gütekriterium für einen wissenschaftlichen pädagogisch-psychologischen Test ist dessen Rasch-Skalierbarkeit (vgl. Kapitel 7). Dabei wird geprüft, ob die erhobenen empirischen Daten mit den Annahmen des Rasch-Modells konform sind (vgl. Absatz 7.1.2). Werden keine signifikanten Verletzungen der Modellpassung identifiziert, können die geschätzten Parameter als Indikator für die latente Fähigkeit einer Person interpretiert werden und für Berechnungen im Rahmen der Validierung herangezogen werden. Im ersten Abschnitt dieses Unterkapitels wird die Itemselektion beschrieben; im zweiten Abschnitt werden erste Indikatoren der Testvalidität regressionsanalytisch ermittelt.

8.3.1 Itemselektion und Auswertung nach Rasch-Modell

Im ersten Schritt der Itemselektion wurden auf Grundlage der Pilotierungsdaten diejenigen Items ausgeschlossen, die eine negative (Bühner, 2011) oder sehr geringe Trennschärfe (Moosbrugger & Kelava, 2012b) aufwiesen. Die Trennschärfe drückt aus, wie gut ein Item zwischen den gemessenen Eigenschaftsausprägungen der Testpersonen unterscheidet (Rost, 2004). Sie wird über die Korrelation der Messwerte eines Items mit der Summe der Rohpunkte des gesamten Tests ermittelt (Bühner, 2011; Moosbrugger & Kelava, 2012b). Mit einem Item mit hoher Trennschärfe können Unterschiede zwischen Testpersonen leicht identifiziert werden, ein wenig trennscharfes Item eignet sich dafür weniger (Rost, 2004). Eine hohe Trennschärfe wird entsprechend als ein positives Merkmal für die Itemselektion herangezogen. Es muss beachtet werden, dass die Trennschärfe verteilungsabhängig ist (Rost, 2004) und insbesondere mit Blick auf die Rasch-Skalierung der Daten nicht als einziges Itemselektionskriterium herangezogen werden sollte.

Für die Auswahl der Items auf Grundlage der Pilotierungsdaten wurde eine Trennschärfe > .20 angestrebt. Um eine ausgewogene Verteilung der Items über die Themengebiete zu gewährleisten, wurden aus inhaltlichen Gründen in

Einzelfällen auch Items mit Trennschärfen zwischen .15 und .20 für die Haupterhebung zugelassen. Als Indikator für die lokale Itempassung unter dem Rasch-Modell wurde der gewichtete mittlere quadrierte Fehler (wMNSQ) herangezogen, da der ungewichtete MNSQ sehr sensitiv auf Abweichungen vom Modell reagiert (Segerer, Marx & Marx, 2012). Ein unter dem Rasch-Modell perfekt passendes Item hat einen Fitwert von 1 (Segerer et al. 2012). Werte kleiner als 1 stehen für eine Item-charakteristische Funktion, die flacher verläuft als im Modell vorgesehen, während Werte größer als 1 Funktionen beschreiben, die steiler als erwartet sind. Grenzwerte für die Bewertung des wMNSQ werden per Konvention festgelegt. Konservativ betrachtet wird von einer schlechten Passung ausgegangen, wenn der wMNSQ kleiner als .80 und größer als 1.20 ist (Bond & Fox, 2013). Zusätzlich gibt ConQuest (Wu et al., 1998) einen T-Wert aus, anhand dessen die Abweichung von der perfekten Passung auf Signifikanz geprüft werden kann. Alle in der Pilotierung untersuchten Variablen wiesen einen sehr guten Passungswert ohne signifikante Abweichung auf (vgl. Anhang B, Tabelle B-1).

Als weiteres Kriterium zur Bewertung der Items wurde geschlechtsspezifisches Differential Item Functioning (DIF) betrachtet. Die DIF-Statistik aus ConQuest gibt an, ob ein Item für ein Geschlecht leichter oder schwerer zu lösen ist als für das andere (Wu et al., 1998). Ist die Abweichung zwischen den Geschlechtern signifikant, sollte das Item auf geschlechtsdiskriminierende Elemente überprüft oder, bei erheblichen Abweichungen, aus dem Itempool ausgeschlossen werden. Von den 24 Items, die auf Basis der Trennschärfe für die Haupterhebung ausgewählt wurden, weisen 3 Items signifikantes geschlechtsspezifisches DIF auf. Die Aufgabe zur Deckungsbeitragsrechnung (Item 2.1) und die Aufgabe zur Einzelpreisbestimmung (Item 4.13) waren für Frauen schwieriger als für Männer. Die Aufgabe zur kriteriengeleiteten Bewertung von Outsourcing-Optionen wies für weibliche Testteilnehmer einen geringeren Schwierigkeitsgrad auf als für männliche Testteilnehmer (Item 2.6) (vgl. Anhang B, Tabelle B-1). Ein χ^2-Test auf Parametergleichheit aller 24 ausgewählter Items deutete ebenfalls auf geschlechtsspezifisches DIF hin ($\chi^2(23, N = 147) = 38.06, p < .05$). Weibliche Testteilnehmer erzielten im Mittel 0.149 Logits weniger auf der Personenparameterskala. Die Ergebnisse wiesen darauf hin, dass der Test insgesamt weibliche Testteilnehmer benachteiligte. Alle Itemkennwerte sind in im Anhang in Tabelle B-1 aufgelistet. Aufgrund der relativ geringen Fallzahl der Pilotierungsstichprobe, die möglicherweise zu Verzerrungen bei der Bewertung von DIF führte, blieben die 3 kritischen Items vorerst im Itempool für die Haupterhebung enthalten, wurden aber vor der Haupterhebung sprachlich so überarbeitet, dass beide Geschlechter angesprochen wurden. Zum Beispiel wurde die Formulierung „Sie sind Controller in einem Unternehmen“ in „Sie sind in einem Unternehmen im Bereich Controlling tätig“ abgeändert.

Bevor die Verteilung der Item- und Personenparameterschätzung betrachtet werden kann, wurde ein Likelihood Ratio (LR) Test von Andersen durchgeführt (Andersen, 1973). Ein signifikanter LR-Test ist ein Indikator für Modellverletzungen im Sinne der Abweichung von der Annahme spezifischer Objektivität (Strobl, 2012). Der Andersen LR-Test vergleicht die Passung zwischen den Parameterschätzungen und den Daten zum einen für den Fall einer Parameterschätzung auf Grundlage der gesamten Stichprobe und zum anderen für den Fall einer getrennten Schätzung nach Subgruppen. Ergeben sich für beide Schätzungen ähnliche Wahrscheinlichkeiten, geht der Likelihoodquotient gegen 1, was dafür spricht, dass die Daten nicht besser durch eine getrennte Schätzung beschrieben werden (Strobl, 2012). Als Teilungskriterium der Stichprobe wurde für die Pilotierung der Median der Rohwerte herangezogen. In der Pilotierung wich der Quotient der beiden Likelihood-Schätzungen nicht signifikant von 1 ab ($\chi^2(23, N = 154) = 25.879, p = .307$), womit ein Minimalkriterium der Modellgeltung erfüllt war (vgl. Abschnitt 7.1.2.2). In einem abschließenden Schritt wurden die Besetzungen der Distraktoren analysiert. Als Distraktoren bezeichnet man die falschen Antwortalternativen in einem Multiple-Choice-Test. Distraktoren, die von weniger als 10 % der Probanden gewählt wurden, wurden für die Endversion des Tests überarbeitet. Zudem wurde mittels Distraktoranalysen sichergestellt, dass die richtige Antwortalternative im Vergleich zu den Distraktoren den höchsten Zusammenhang mit der Testleistung aufwies. Würde ein Distraktor häufig, auch von leistungsstarken Testteilnehmern, gewählt, so wäre das ein Hinweis darauf, dass die Antwortalternativen missverständlich sind und umformuliert werden müssten. Da der Test vor der Pilotierung bereits durch 24 Masterstudierende evaluiert wurde, lagen nach der Itemselektion keine auffälligen Muster für den Zusammenhang zwischen Beantwortung eines Distraktors und der Gesamtpunktzahl im Test vor. In allen Fällen war das das Wählen der richtigen Antwortalternative signifikant positiv und das Wählen der falschen Alternativen negativ oder geringfügig positiv mit der Gesamtpunktzahl korreliert.

Nachdem die Auswahl der Items vorgenommen wurde und die 24 selektierten Items auf Modellverletzungen untersucht wurden, wurde die Verteilung der Personen und Items auf einem gemeinsamen Schwierigkeits-/Fähigkeitskontinuum betrachtet. Als Darstellungsform wird die Wright-Map gewählt (siehe Abbildung 10), auf deren linker Seite die Verteilung der Personenparameter und auf deren rechter Seite die Items mit Nummer entsprechend ihrer Schwierigkeit angeordnet sind. Eine Person, die sich auf dem Fähigkeitskontinuum exakt an der Position eines Items befindet, hat eine 50 %ige Wahrscheinlichkeit, dieses Item zu lösen (Wu et al., 1998). Alle Items, die unterhalb der Fähigkeit der Person liegen, werden mit einer größeren Wahrscheinlichkeit gelöst,

alle Items, die darüber liegen, mit einer geringeren Wahrscheinlichkeit (Wu et al., 1998). Die Wright-Map der Pilotierung ist in Abbildung 10 dargestellt.

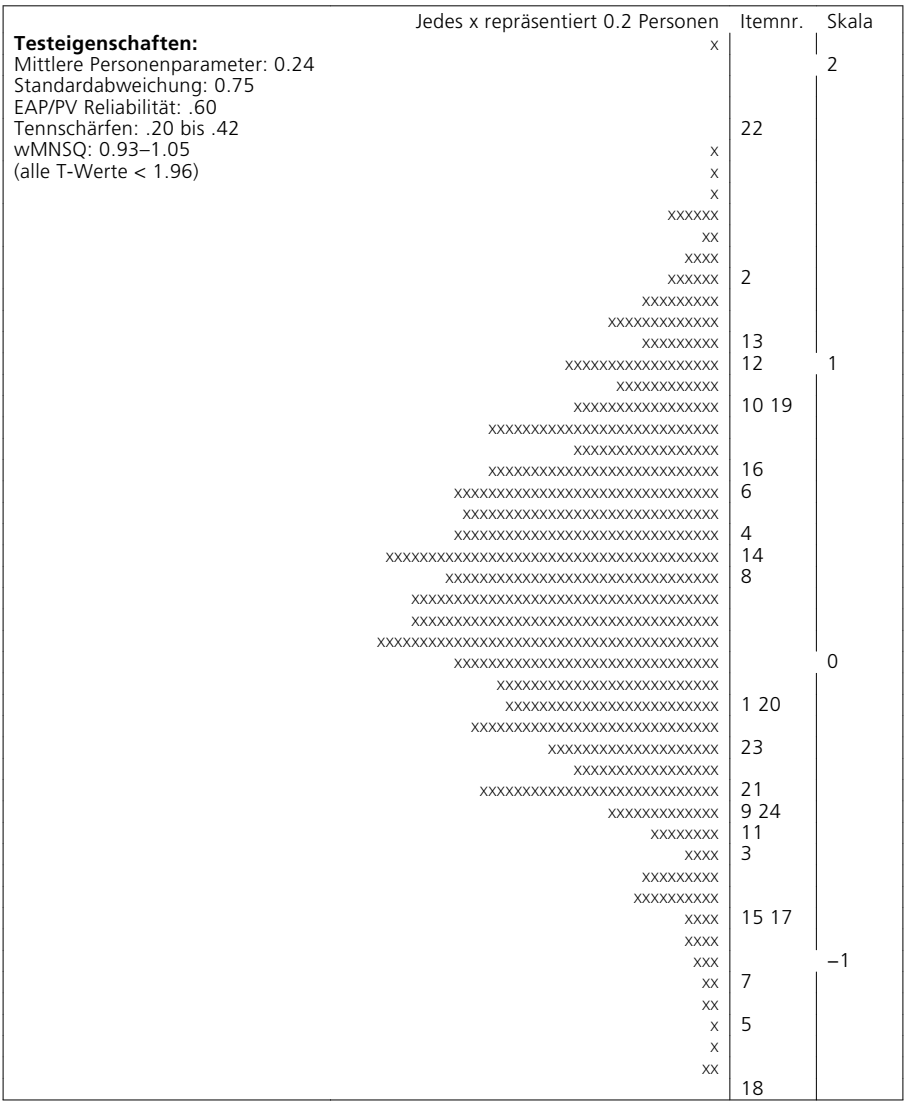


Abb. 10: Darstellung der Pilotierungsergebnisse in Form einer Wright-Map (rechts in der Abbildung) und zentraler Testeigenschaften (links in der Abbildung)

(Quelle: Jähnig, 2013, S. 53)

Die ausgewählten Items verteilten sich angemessen über das Schwierigkeitskontinuum. Lediglich im mittleren Schwierigkeitsbereich sollte die Itemlücke geschlossen werden. Insgesamt zeigte die leichte Verschiebung der Personenparameter in Richtung oberhalb des Nullpunktes der Skala, dass der Test tendenziell zu leicht war. Eine Zuordnung der Itemnummern zu einer detaillierten Beschreibung der Items liegt in der Tabelle B-1 im Anhang B vor.

Die EAP/PV-Reliabilität spiegelt den Anteil der Personenvarianz an der vom Modell geschätzten Gesamtvarianz wider und kann ähnlich wie Cronbachs Alpha interpretiert werden (Rost, 2004) (vgl. Abschnitt 7.2.2.). Sie wird von ConQuest (Wu et al., 1998) ausgegeben und betrug in der Pilotierung .60 (vgl. Abbildung 10).

Nach der oben beschriebenen Prüfung der Items und den Voraussetzungen für die Rasch-Skalierung wurden die Personenparameter mit ConQuest (Wu et al., 1998) nach dem marginalen Maximum-Likelihood-Prinzip geschätzt (Strobl, 2012) und im Rahmen der Validierung weiter verwendet.

8.3.2 Erste Indikatoren der Testvalidität

Die Validität eines Tests kann vor dem Hintergrund unterschiedlicher empirischer und theoretischer Überlegungen beurteilt werden (vgl. Abschnitt 7.2.3).

Im Rahmen der Pilotierung beschränkt sich die Betrachtung der Validität auf explorative Analysen. Komplexe Untersuchungen zur Konstruktvalidität anhand mehrdimensionaler Modellierungen werden aufgrund der geringen Stichprobengröße nicht vorgenommen. Diese werden in der Hauptuntersuchung ausführlich behandelt (siehe Abschnitt 9.3.2.). Für eine erste Kriteriumsvalidierung wurde der Zusammenhang zwischen den geschätzten Personenparametern und theoretisch sowie empirisch relevanten Außenkriterien betrachtet. Dabei wurde geprüft, ob die geschätzten Personenparameter im Mittel erwartungskonform mit den Kriterien korrelierten. Die Ergebnisse der Validierung im Rahmen der Pilotierung wurden bereits von Jähnig (2013) veröffentlicht und werden deshalb an dieser Stelle nur zusammenfassend dargestellt. Für eine detaillierte Darstellung und Diskussion der Ergebnisse wird auf Jähnig (2013) verwiesen.

Die im Folgenden zusammengefassten Befunde der Pilotierung beziehen sich auf die Ergebnisse einer Regressionsanalyse, die im Anhang B, Tabelle B-2 aufgeführt ist.

Eine absolvierte kaufmännische Ausbildung stand, wie erwartet (Förster et al., 2012), in einem positiven Zusammenhang mit der Testleistung, *stand.* $\beta = .18$, $t(122) = 2.00$, $p = .047$. Darüber hinaus bestätigte sich ein positiver Zusammenhang zwischen der Anzahl der absolvierten Module und der Testleistung

(*stand.* $\beta = .23$, $t(122) = 2.68$, $p = .008$), was als Indikator für die curriculare Validität des Tests und die Domänenspezifizität der Items interpretiert wurde. Der negative[11] Zusammenhang zwischen der durchschnittlichen Note der Hochschulzugangsberechtigung und der Testleistung (stand. $\beta = -.21$, $t(122) = -2.29$, $p = .024$) wurde als Gradmesser für den Einfluss allgemeiner akademischer Leistungsfähigkeit auf die Testleistung herangezogen. Die relativ geringe Stärke des Zusammenhangs wurde als positiver Indikator dahingehend interpretiert, dass der Test domänenspezifisches Wissen erfasst und die Varianz der Testleistung nicht hauptsächlich durch Faktoren allgemeiner kognitiver Leistungsfähigkeit aufgeklärt wird. Wie in der Mehrzahl der Studien zu Wirtschaftswissen bei Studierenden (vgl. Kapitel 3) wurde in der Pilotierung bestätigt, dass männliche Testteilnehmer unter Kontrolle anderer relevanter Faktoren bessere Testleistungen erzielen als weibliche (stand. $\beta = -.24$, $t(122) = -2.727$, $p = .007$).

Neben der Note der Hochschulzugangsberechtigung wurden auch die Abschlussnoten der absolvierten Ausbildung und die Noten der einzelnen BWL-Pflichtmodule abgefragt. Da die Fallzahlen für beide Variablen gering ausfielen, wurden an Stelle einer Regression einfache bivariate Korrelationen betrachtet. Für die Abschlussnote der Ausbildung wurde, analog zur Note der Hochschulzugangsberechtigung, ein negativer Zusammenhang erwartet, entsprechend wird das einseitige Signifikanzniveau angegeben. Dieser Zusammenhang wurde durch die Daten empirisch bestätigt ($r(47) = -.38$, $p = .004$). Die durchschnittliche Modulnote in den absolvierten BWL-Pflichtmodulen sollte ebenfalls negativ mit der Testleistung korreliert sein. Aufgrund der zeitlichen und thematischen Nähe des Tests zum Studium sollte die Korrelation höher ausfallen als die Korrelation zwischen Testleistung und dem Mittel der Abschlussnote der Berufsausbildung. Eine Spearman-Korrelation bestätigte einen signifikanten negativen Zusammenhang zwischen Testleistung und den mittleren Modulnoten, jedoch fiel der Zusammenhang entgegen den Erwartungen geringer aus als der Zusammenhang zwischen Testleistung und Abschlussnote der Ausbildung ($r(104) = -.22$, $p = .014$).

8.4 Konsequenzen aus der Pilotierung für die Haupterhebung

Zusammenfassend ergaben sich aus der Pilotierung vielversprechende Ergebnisse bezüglich der Rasch-Skalierbarkeit und der Validität des Tests. Die Vertei-

11 Negativ, weil bei Schulnoten eine 1 die beste Leistung repräsentiert, während eine 6 für die schlechteste Leistung steht.

lung der Items über das Schwierigkeitskontinuum war angemessen, und sowohl die lokale als auch die globale Itempassung sprachen für eine Schätzung der Itemparameter über das Rasch-Modell. Kritisch zu betrachten ist die geringe Reliabilität der 24 ausgewählten Items. Obwohl interne Konsistenzen zwischen .50 und .60 in frühen Phasen der Instrumentenentwicklung akzeptabel sind (Nunnally, 1967), kann sich ein Messinstrument erst als reliabel etablieren, wenn die interne Konsistenz den Wert .70 erreicht und bestenfalls überschreitet (Streiner, 2003). Als zentrale Konsequenz für die Haupterhebung ergab sich aus der Pilotierung die Zielsetzung, die Reliabilität des Tests zu erhöhen. Um reliabilitätsminderenden Ratetendenzen entgegenzuwirken, wurden für die Haupterhebung Distraktoren, die nur von wenigen Testteilnehmern gewählt wurden, umformuliert. Zudem wurden einige Items sprachlich vereinfacht. Um in der Haupterhebung Störeinflüsse durch gegenseitiges Abschreiben zu reduzieren, wurden in der Haupterhebung zwei Testversionen (Testheft A und Testheft B) mit unterschiedlichen Itemreihenfolgen eingesetzt. Da die geringe Reliabilität des Tests möglicherweise ein Hinweis auf bisher unerkannte Mehrdimensionalität des Tests war, sollen in der Hauptuntersuchung verschiedene mehrdimensionale Modelle gegeneinander getestet werden. Zudem wird eine verlängerte Version des Tests entwickelt. Mit Hilfe der Spearman-Brown-Formel (Schermelleh-Engel & Werner, 2012) wurde ein Verlängerungsfaktor von mind. 1.6 ermittelt, was ca. 14 Items entspricht. In Anbetracht der begrenzten Testzeit und der parallel initiierten alternativen Bemühungen um Reliabilitätssteigerung wurde eine Verlängerung um sechs Items vorgenommen.

Die Pilotierung lieferte neben der kritischen Reliabilität Hinweise, dass der Test möglicherweise weibliche Testteilnehmer benachteiligt. Als Konsequenz wurden alle ausgewählten Items auf geschlechtsspezifische Formulierungen (z. B. Formulierungen im generischen Maskulinum) geprüft und sprachlich entsprechend angepasst.

Die ersten Indikatoren der Testvalidität zeigten, dass der Test sich im Verhältnis zu Außenkriterien so verhielt, wie auf Grundlage der Forschung zum MFAT-B, WBT und BAKT zu erwarten war (vgl. Kapitel 3). Die verhältnismäßig geringe Varianzaufklärung der linearen Regressionsgleichung von 11 % (vgl. Anhang B, Tabelle B-2) wies jedoch darauf hin, dass es noch zahlreiche weitere Faktoren gibt, die mit der Testleistung zusammenhängen und im Rahmen einer weiterführenden Validierung herangezogen werden könnten. In der Hauptuntersuchung gilt es dementsprechend, elaboriertere Analysen zur Testvalidierung durchzuführen. Dazu gehörte auch eine Konstruktvalidierung in Abgrenzung zum BAKT (Bothe, 2013). Zudem mussten sich die Items über den universitären Standort Göttingen hinaus als curricular valide erweisen. Dafür wurden in der

Haupterhebung Testteilnehmer an weiteren universitären Standorten rekrutiert.

Die Anlage und die Ergebnisse der Hauptuntersuchung werden im folgenden Kapitel beschrieben.

9 Testvalidierung und weiterführende Analysen

Nachdem in der Pilotierung gute Indikatoren für die Validität und eine ausreichende, aber ausbaufähige Reliabilität ermittelt wurden, folgte die Hauptuntersuchung zur Analyse situativer Aufgaben für die Messung betriebswirtschaftlichen Wissens an deutschen Universitäten. Vorrangiges Ziel der Studie war die Validierung des Tests. Darüber hinaus sollten weiterführende Erkenntnisse darüber gewonnen werden, wie betriebswirtschaftliches Wissen organisiert ist und welche individuellen Faktoren Testleistungen in betriebswirtschaftlichen Wissenstests begünstigen oder reduzieren. Schlussendlich sollen die Berechnungen auch als Grundlage für weitere Forschung auf dem Gebiet der empirischen Bildungsforschung an Hochschulen dienen.

9.1 Ableitung der Hypothesen

Eine zentrale Fragestellung, die sowohl im Rahmen der Validierung eine Rolle spielte als auch von allgemeinem Erkenntnisinteresse war, ist die Frage nach der Binnenstruktur (bzw. Dimensionalität) betriebswirtschaftlichen Wissens auf Bachelorniveau. Bei der Prüfung der Dimensionalität geht es darum, empirisch zu erfassen, ob ein Merkmal (situatives betriebswirtschaftliches Wissen) über mehrere Unterfacetten (z. B. Wissen über Marketing und Wissen über Produktion) beschrieben werden kann oder ob es sich um ein eindimensionales relativ homogenes Konstrukt (situatives betriebswirtschaftliches Wissen auf Bachelorniveau) handelt. Zudem stellt sich die Frage, inwiefern situative und nicht-situative Testaufgaben voneinander trennbare Wissensdimensionen erfassen. Im Rahmen der Hauptuntersuchung wurde diesbezüglich zwei Fragen nachgegangen:

(1) Bilden die situativen Items eine latente Dimension oder lassen sich spezifische Subdimensionen betriebswirtschaftlichen Wissens auf Bachelorniveau bestätigen?

(2) Bilden deklarative Wissensitems und anwendungsorientierte situative Wissensitems zwei empirisch trennbare Dimensionen?

Neben der Untersuchung der Struktur betriebswirtschaftlichen Wissens wurde analysiert, welche individuellen Faktoren der Studierenden positiv oder negativ mit der Testleistung assoziiert sind. Die Auswahl der Validierungskriterien erfolgte auf der Basis theoretischer Überlegungen und nationaler sowie internationaler empirischer Forschungsbefunde (vgl. Kapitel 3). Ziel der Überprüfung der Determinanten war zum einen die Untermauerung der Testvalidität, zum

anderen sollen die Ergebnisse einen Beitrag dazu leisten, die Genese guter betriebswirtschaftlicher Studienleistungen aufzuklären.

Die Modellbildung und Ableitung der Hypothesen zur Struktur betriebswirtschaftlichen Wissens werden in Abschnitt 9.1.1 beschrieben. Die Ableitung der Hypothesen über die Determinanten erfolgt in Abschnitt 9.1.2.

9.1.1 Hypothesen zur Struktur betriebswirtschaftlichen Wissens

Zur Struktur betriebswirtschaftlichen Wissens auf Bachelorniveau liegen bisher nur wenige Forschungsbefunde vor (vgl. Kapitel 3). Grundlegend wird davon ausgegangen, dass betriebswirtschaftliches Wissen neben volkswirtschaftlichem Wissen eine eigene Dimension abbildet (Förster et al., 2012). Die Binnenstruktur betriebswirtschaftlichen Wissens wurde von Bothe (2003) und Ling (2012) faktoranalytisch untersucht. Beide Autoren ziehen eine eindimensionale Modellierung betriebswirtschaftlichen Wissens bis zum Bachelor vor. Die Ergebnisse von Bothe (2003) sind aufgrund geringer Itemzahlen nur mit Einschränkungen interpretierbar. Die Ergebnisse von Ling (2012) sind hingegen empirisch sehr gut untermauert (vgl. Abschnitt 3.5). Beide Autoren verweisen darauf, dass es sich um ein breites Konstrukt mit heterogenen Eigenschaften handelt (Bothe, 2003; Ling, 2012). Auf dieser Grundlage wurde für den hier konstruierten situativen Test die Hypothese aufgestellt, dass es sich bei betriebswirtschaftlichem Wissen auf Bachelorniveau um ein vielschichtiges eindimensionales Konstrukt handelt.

H1–1: Die 24 situativen Items lassen sich am besten über ein eindimensionales Modell abbilden.

Das eindimensionale Modell wurde mit folgenden alternativen Modellen verglichen: (1) einem an den Fachinhalten orientierten vierdimensionalen Modell und (2) einem in Anlehnung an Befunde von Seeber (2008) aufgestellten zweidimensionalen Modell, das zwischen Items mit Bezug zum Rechnungswesen und Items mit Bezug zu allgemeinen betriebswirtschaftlichen Organisations- und Leistungsprozessen unterscheidet (vgl. Kapitel 4.)

Im Sinne der divergenten und konvergenten Konstruktvalidität wurde die Hypothese aufgestellt, dass die situativen Items und die Items des BAKT (Bothe, 2003) eine hohe latente Korrelation aufweisen (da sie inhaltlich das gleiche Zielkonstrukt messen). Trotzdem wird vermutet, dass sich die Items beider Tests durch ein zweidimensionales Modell am besten beschreiben lassen, weil zu ihrer Lösung unterschiedliche Wissensarten benötigt werden (vgl. Abschnitt 2.3).

H1–2: Die situativen Items bilden neben den nicht-situativen Items des BAKT eine eigene Dimension ab; die latente Korrelation zwischen den Dimensionen ist hoch.

Das zweidimensionale Modell wurde mit einem eindimensionalen Modell verglichen, das alle Items beinhaltete, und mit einem vierdimensionalen Modell, in dem alle Items nach Modulinhaltsbereichen gruppiert wurden (Unternehmensführung, Finanz- und Rechnungswesen, Produktion und Marketing).

Auf Basis der Überlegungen zu schwierigkeitsbestimmenden Aufgabenmerkmalen und vorangegangenen empirischen Untersuchungen zum Einfluss der Wissensart auf die Aufgabenschwierigkeit (vgl. Abschnitt 4.2.2 und Abschnitt 4.3) wurde vorhergesagt, dass die situativen Aufgaben des hier entwickelten Tests, unter anderem aufgrund ihrer höheren Komplexität, schwerer sind als die deklarativen Aufgaben des BAKT (Bothe, 2003).

H1–3a: Die situativen Aufgaben haben im Mittel niedrigere Lösungsquoten als nicht-situative Aufgaben.

Unter Berücksichtigung der Forschungsbefunde zum positiven Einfluss von Vorerfahrung auf die Testleistung (vgl. Kapitel 3) sowie der Rolle des Bekanntheitsgrades mit einer Aufgabe für deren Aufgabenschwierigkeit (Gschwendtner, 2008) wird vorhergesagt, dass der Effekt vom Aufgabentyp (situativ/nicht-situativ) auf die Testleistung durch das Ausmaß der Bekanntheit mit situativen beruflichen Aufgabenstellungen beeinflusst wird.

H1–3b: Bezogen auf die Testleistung besteht ein Interaktionseffekt zwischen der Art der Aufgabenstellung (situativ vs. nicht-situativ) und dem Ausmaß der Bekanntheit mit situativen Aufgabenstellungen.

Der Bekanntheitsgrad mit situativen Aufgaben wurde darüber erfasst, ob die am Test teilnehmende Person vor dem Studium eine kaufmännische Ausbildung absolviert hat. Die Richtung des Effekts wurde dahingehend vorhergesagt, dass situative Aufgaben für Testteilnehmer mit kaufmännischer Ausbildung einfacher sein sollten als für Personen ohne kaufmännische Ausbildung. Diese Hypothese beruht auf zwei zentralen Annahmen: (1) weist der betriebliche Lernort in der kaufmännischen Ausbildung naturgemäß einen hohen Praxisbezug auf und (2) werden Auszubildende im Lernort Schule verstärkt auf problem- und praxisorientierte Aufgabenstellungen vorbereitet, da die Aufgaben der Abschlussprüfungen zumeist an realitätsnahen Geschäftsprozessen orientiert sind (Achtenhagen & Winther, 2011). Eine Bestätigung der Hypothesen H1–3a und H1–3b würde einen weiteren Baustein für die Konstruktvalidierung darstellen. Darüber hinaus liefern die Analysen interessante Hinweise da-

rüber, inwiefern sich Studierende mit berufpraktischer Vorerfahrung bezüglich ihrer Wissensstrukturen von Studierenden ohne Vorerfahrung unterscheiden.

9.1.2 Hypothesen zu den Determinanten betriebswirtschaftlichen Wissens

Die Hypothesen zu den Determinanten der Testleistung beruhen in erster Linie auf den Ergebnissen der Forschung zum WBT und zum MFAT-B (vgl. Kapitel 3).

Analog zur Pilotierung wurde im Sinne einer curricularen Validität ein positiver Zusammenhang zwischen der Anzahl der besuchten Lehrveranstaltungen (testrelevante Pflichtmodule) und der Testleistung vorhergesagt. Unabhängig von der Anzahl der besuchten Module sollte der Zusammenhang zwischen der mittleren Modulabschlussnote in den testrelevanten Pflichtmodulen und der Testleistung substanziell sein.

H2–1a: Die Anzahl der besuchten testrelevanten Pflichtmodule hängt positiv mit der Testleistung zusammen.

H2–1b: Die durchschnittliche Modulnote in den testrelevanten Pflichtmodulen hängt negativ mit der Testleistung zusammen.

Als Proxy für allgemeine kognitive Leistungsfähigkeit wurde die Note der Hochschulzugangsberechtigung herangezogen. Der Zusammenhang zwischen der Note der Hochschulzugangsberechtigung und der Testleistung wurde bereits in der Pilotierung bestätigt (vgl. Abschnitt 8.3.2). Die Durchschnittsnote der Hochschulzugangsberechtigung sollte jedoch nicht den Hauptteil der Varianz der Testleistung aufklären, da dies möglicherweise ein Indikator dafür wäre, dass der Test zu stark auf allgemeine akademische Leistungsfähigkeit abzielt und nicht ausreichend domänenspezifisch angelegt wurde.

H2–2: Die Durchschnittsnote der Hochschulzugangsberechtigung hängt negativ mit der Testleistung zusammen.

Der Zusammenhang wird in negativer Richtung vorhergesagt, weil niedrigzahlige Noten im deutschen Schul- und Hochschulsystem für bessere Leistungen stehen als hochzahlige Noten.

Aus lerntheoretischer Perspektive liegt nahe, dass domänenspezifisches Vorwissen positiv mit Wissensaufbau in derselben Domäne verbunden ist (Maier, 2012) und durch berufspraktische Erfahrungen Wissen aufgebaut werden kann (Mandl & Gerstenmaier, 2000). Der positive Effekt von domänenspezifischer Vorerfahrung auf Leistungen in Wissenstests wurde sowohl in nationalen Studien (vgl. Abschnitt 3.2) als auch in internationalen Studien (vgl. Abschnitt 3.5) mehrfach empirisch bestätigt. In Anlehnung an diese Befunde wur-

de für die situativen Testaufgaben ein positiver Zusammenhang zwischen einer vor dem Studium absolvierten kaufmännischen Berufsausbildung und der Testleistung vorhergesagt.

H2–3: Eine vor dem Studium absolvierte kaufmännische Berufsausbildung hängt positiv mit der Testleistung zusammen.

Zahlreiche empirische Studien und die Ergebnisse der Pilotierung untermauern die Überlegenheit männlicher Testteilnehmer in wirtschaftswissenschaftlichen Wissenstests gegenüber weiblichen (vgl. Kapitel 3). Konsequenterweise wurde auch in der Haupterhebung ein Leistungsvorsprung von männlichen Testteilnehmern vorhergesagt. Es ist zu beachten, dass ein Leistungsunterschied zwischen männlichen und weiblichen Testteilnehmern in diesem Fall für die Kriteriumsvalidität des Tests sprechen würde. In Bezug auf die Testfairness weist ein geschlechtsspezifischer Unterschied möglicherweise auf mangelhafte Fairness des Tests hin. Um diesen Überlegungen nachzugehen, wurden in Abschnitt 9.3.4 differenzierte Analysen zur Beurteilung der Testfairness durchgeführt.

H2–4: Männliche Testpersonen erzielen bessere Leistungen in dem Test als weibliche Testpersonen.

Die durch die Hypothesen H2–1 bis H2–4 abgedeckten Validierungskriterien haben sich bereits in verschiedenen Studien zur Testleistung in wirtschaftswissenschaftlichen Tests bewährt (z. B. Förster et al., 2012). Abgesehen von den Indikatoren der curricularen Validität (H1a und H1b), handelte es sich bei allen Kriterien um individuelle Eigenschaften der Testteilnehmer, die schon vor Beginn des Studiums feststehen. Sie sind somit durch die Universitäten nicht im Studienprozess veränderbar. Um Variablen zu identifizieren, die durch Prozesse innerhalb der Universität veränderbar wären, wurden vier weitere individuelle studienbezogene Faktoren zur Testvalidierung herangezogen: (1) das fachspezifische Studieninteresse, (2) die selbstberichteten Lernstrategien der Studierenden während des Studiums, (3) die Leistungs- und Wettbewerbsmotivation der Studierenden und (4) das akademische und mathematische Selbstkonzept der Studierenden. Alle vier Variablen wurden in vorangegangenen empirischen Studien positiv mit Studienleistung in Verbindung gebracht (z. B. Schiefele, Streblow, Ermgassen & Moschner, 2003; Zeegers, 2004). Ihre Rolle im Zusammenhang mit Leistungen in betriebswirtschaftlichen Wissenstests auf Bachelorniveau wurde jedoch noch nicht ausreichend untersucht.

Auf der Basis von Befunden zum positiven Zusammenhang zwischen fachspezifischem Interesse und schulischen Leistungen (Krapp, Schiefele & Schreyer, 1993; Köller, Trautwein, Lüdtke & Baumert, 2006) wurde für die Testleistung

ein positiver Zusammenhang mit Interesse an betriebswirtschaftlichen Fachinhalten vorhergesagt.

H2–5: Interesse an betriebswirtschaftlichen Fachinhalten hängt positiv mit der Testleistung zusammen.

Tiefenlernstrategien wurden in unterschiedlichen Studien mit guter Test- (Duff, 2004) und Studienleistung (Schiefele et al., 2003) in Verbindung gebracht (vgl. Absatz 3.5). Zu den Tiefenlernstrategien zählt zum Beispiel das Herstellen von Zusammenhängen (Schiefele et al., 2003; Wild & Schiefele, 1994). Tiefenlernstrategien werden häufig mit Oberflächenlernstrategien kontrastiert, zu denen beispielsweise das Lernen mit Hilfe von Wiederholungen zählt (Duff, 2004). In Anlehnung an die Überlegungen von Schiefele et al. (1993) und die empirischen Ergebnisse von Duff (2004) wurde folgende Hypothese aufgestellt.

H2–6: Studierende, die überwiegend Tiefenlernstrategien verwenden, weisen bessere Testleistungen auf als Studierende, die überwiegend Oberflächenlernstrategien verwenden.

Wettbewerbsmotivation und allgemeine Leistungsmotivation wurden in einer Studie von Schiefele et al. (2003) als positive Prädikatoren für die Vordiplomsnote in unterschiedlichen Fächern identifiziert. Für die Testleistung wurde ebenfalls ein positiver Zusammenhang postuliert, da vermutet wird, dass sich Leistungs- und Wettbewerbsmotivation nicht nur positiv auf die Lernleistung im Studium auswirken, sondern speziell während der Testung eine leistungsförderliche Rolle spielen.

H2–7: Leistungs- und Wettbewerbsmotivation hängen positiv mit der Testleistung zusammen.

Das akademische Selbstkonzept wird als Gesamtheit der kognitiven Repräsentationen eigener Fähigkeiten in akademischen Leistungssituationen definiert (Dickhäuser, Schöne, Spinath & Stiensmeier-Pelster, 2002). Es gilt als wichtige Einflussgröße leistungsthematischen Verhaltens in verschiedenen Bereichen (Möller & Köller, 2004). Entsprechend wurde ein positiver Zusammenhang zwischen akademischem Selbstkonzept und Testleistung vorhergesagt. Als Unterfacette des allgemeinen akademischen Selbstkonzepts wurde das mathematische Selbstkonzept erhoben und explorativ auf Zusammenhänge mit der Testleistung geprüft.

H2–8: Das akademische Selbstkonzept hängt positiv mit der Testleistung zusammen.

Bevor in Abschnitt 9.3 die Ergebnisse der Hypothesentestung vorgestellt werden, erfolgt im nächten Abschnitt die Beschreibung der Rahmenbedingungen der Datenerhebung.

9.2 Testdurchführung, Testmaterial und Stichprobenbeschreibung

Die Datenerhebung für die Haupterhebung fand im Wintersemester 2012/13 und im Sommersemester 2013 an der Georg-August-Universität Göttingen, der Universität Hamburg und der Universität Mannheim statt. Die drei Standorte wurden ausgewählt, da sie wichtige wirtschaftswissenschaftliche Studienstandorte in Deutschland sind und jeweils hohe Studierendenzahlen in den Fächern Wirtschaftspädagogik und Betriebswirtschaftslehre aufweisen (Henning & Henning, 2009). Deshalb waren die Modulbeschreibungen dieser drei Universitäten rahmengebend für die Testentwicklung (vgl. Anhang A).

Die Rekrutierung der Testteilnehmer für die Haupterhebung erfolgte über die Dozierenden wirtschaftswissenschaftlicher und wirtschaftspädagogischer Lehrveranstaltungen. An den Standorten Hamburg und Mannheim konnten aus organisatorischen Gründen nur Studierende der Wirtschaftspädagogik rekrutiert werden.

Die Rahmenbedingungen der Testdurchführung und die Instruktionen waren mit der Pilotierung weitgehend identisch (vgl. Kapitel 8).

Das Testmaterial bestand aus einem papierbasierten, neun Doppelseiten umfassenden Testheft. Die erste Seite beinhaltete die Instruktion, die folgenden Seiten waren in vier thematische Abschnitte unterteilt. Im ersten Abschnitt befanden sich die beiden Wissenstests: Der hier entwickelte situative Test (24 Items) und eine inhaltlich auf den situativen Test abgestimmte Auswahl von Items des BAKT (Bothe, 2003) (23 Items). Items des BAKT (2003) mit Bezug zu Steuern und Human Recources wurden ausgeschlossen (vgl. Tabelle 4). Die situativen Items wurden je nach Parallelversion des Testhefts entweder an erster oder an zweiter Position dargeboten. Im zweiten Abschnitt befanden sich Fragen zur Person, insbesondere zum schulischen Werdegang vor Beginn des Studiums. Im dritten Abschnitt wurden Fragen zum aktuellen Studium gestellt. Im letzten Abschnitt des Testhefts wurden individuelle studienbezogene Faktoren, wie zum Beispiel das Studieninteresse und die Wettbewerbs- und Leistungsmotivation, erfasst. Der Begleitfragebogen zu den Abschnitten zwei bis vier ist in Anhang C dokumentiert. Eine tabellarische Darstellung des gesamten Testshefts mit Angaben zu den verwendeten Skalen und den Quellen der Skalen ist

im Anhang C, Tabelle C-1 dargestellt. Die Dokumentation aller Skalen aus der Haupterhebung befindet sich im Anhang D.

An der Haupterhebung nahmen insgesamt 421 Studierende teil. Davon waren 211 Studierende männlich und 208 Studierende weiblich, zwei Personen gaben ihr Geschlecht nicht an. Zum Zeitpunkt der Erhebung waren die Testteilnehmer im Mittel zwischen 25 und 26 Jahre alt (*SD* = 3.29). Die Zusammensetzung der Stichprobe nach Studiengang und Studienabschnitt ist in Tabelle 13 aufgeführt.

Tab. 13: Darstellung der Stichprobenzusammensetzung für die gesamte Haupterhebung nach Studiengang und Studienabschnitt (N = 421)

Studiengang	**Studienabschnitt**		**Summe**
	Bachelor	**Master**	
BWL	124	46	170
Wirtschaftspädagogik	49	130	179
VWL	35	2	37
Anderer Studiengang	4	5	9
Summe	212	183	395[12]

Ein Großteil der Testteilnehmer wurde am Hochschulstandort Göttingen rekrutiert (*n* = 204). Weitere Erhebungsstandorte waren die Universität Hamburg mit 62 Testteilnehmern und die Universität Mannheim mit 64 Teilnehmern.

Die Mehrzahl der Testteilnehmer (65 %) füllte den situativen Test bis auf ein Item vollständig aus. Im Mittel wurden pro Person 1.7 (*SD* = 2.45) der 24 situativen Items nicht beantwortet. Nicht beantwortete Items wurden als falsch bewertet. Fragebögen, in denen mehr als 70 % der Aufgaben nicht beantwortet wurden, wurden aus der Analyse ausgeschlossen, da nicht davon auszugehen war, dass der Test ernsthaft bearbeitet wurde. Es waren jedoch nur drei Fragebögen von dieser Regel betroffen.

Eine Teilstichprobe von 70 Studierenden der Universität Göttingen erhielt einen abgeänderten Fragebogen. Zur Reliabilitätssteigerung wurde der situative Test um sechs Items (vorwiegend aus dem Inhaltsbereich Unternehmensführung und Finanz- und Rechnungswesen) erweitert. Zudem wurde anstelle der Fragen zu individuellen studienbezogenen Faktoren ein Fragebogen zur Testakzeptanz (Kersting, 2008) ausgegeben. In diesem Fragebogen sollten die Studierenden einschätzen, inwiefern sie den Test belastend empfanden und wie stark sie die Ergebnisse des Tests für kontrollierbar hielten (Kersting, 2008). Zudem

12 26 Testteilnehmer machten keine Angaben zum Studiengang oder Studienabschnitt.

wurde nach der subjektiven Einschätzung der Relevanz der Aufgaben für das spätere Berufsleben (Kersting, 2008) und für das Studium (eigene Items) gefragt. Als weitere Maßnahme der Reliabilitätssteigerung wurde die Testzeit im Vergleich zur Haupterhebung reduziert, indem jeweils entweder nur Items des BAKT (Testheft A, $n = 35$) oder nur situative Items (Testheft B, $n = 35$) dargeboten wurden. Ziel dieses Erhebungsdesigns war zum einen die Steigerung der Reliabilität, zum anderen sollte ein statistischer Vergleich zwischen situativen und nicht-situativen Items mit Bezug auf die Testakzeptanz von Seiten der Studierenden ermöglicht werden, da Testakzeptanz als ein Nebengütekriterium bei der Entwicklung und Bewertung pädagogisch-psychologischer Tests gilt (vgl. Abschnitt 7.2.3.1).

Der Aufbau der beiden Studien, die im Rahmen der Haupterhebung durchgeführt wurden, ist in der Tabelle 14 aufgeführt.

Tab. 14: Beschreibung des Aufbaus der beiden Studien der Haupterhebung

	Studien der Haupterhebung	
	Validierungsstudie	**Studie zur Reliabilitätssteigerung und Testakzeptanz**
Stichprobengröße	N = 351	N = 70
Erhebungsstandorte	Hamburg, Göttingen, Mannheim	Göttingen
Studentische Zielgruppe	BWL, Wipäd, VWL, Andere	BWL
Testmaterial	1. Situative Items **und** ausgewählte Items des deklarativen BAKT 2. Fragen zur Person 3. Fragen zum Studium 4. Fragen zu individuellen Studienbezogenen Faktoren	1. Um sechs Items verlängerte Version des situativen Tests ($n = 35$) (Testheft A) **oder** ausgewählte Items aus dem BAKT ($n = 35$) (Testheft B) 2. Fragen zur Person 3. Fragen zum Studium 4. Fragen zur Testakzeptanz und zur Anstrengungsbereitschaft
Ziele	Konstrukt- und Kriteriumsvalidierung sowie Vergleich von (Extrem-) Gruppen innerhalb der Stichprobe	Vergleich der beiden Aufgabentypen bezüglich Facetten der studentischen Testakzeptanz sowie Steigerung der Reliabilität durch Verkürzung der Gesamt- testzeit und Verlängerung des SJTs um sechs Items

Die Ergebnisse der Haupterhebung und der Zusatzstudie zur Testakzeptanz werden in den folgenden Abschnitten dargestellt. Die Interpretation der Ergebnisse erfolgt in Kapitel 10.

9.3 Empirische Ergebnisse

Die Auswertung der Ergebnisse des in dieser Arbeit entwickelten Tests erfolgte über das Rasch-Modell. In Abschnitt 9.3.1 werden im Rahmen der Prüfung der Hypothese H1–1 zur Binnenstruktur situativen betriebswirtschaftlichen Wissens unterschiedlich-dimensionale Modelle miteinander verglichen. Zudem werden zentrale Testeigenschaften, wie die Reliabilität, aufgeführt. In Abschnitt 9.3.2 werden die Ergebnisse der dimensionalen Vergleiche der situativen Items mit den Items des BAKT (Bothe, 2003) dargestellt. Darüber hinaus werden die Testwerte von unterschiedlichen (Extrem-)Gruppen der Stichprobe miteinander verglichen und hinsichtlich ihrer Erwartungskonformität geprüft. In Abschnitt 9.3.3 erfolgt der Ergebnisbericht über die Prüfung der Hypothesen zur Kriteriumsvalidität (vgl. Abschnitt 9.1.2) mittels korrelativer und regressionsanalytischer Berechnungen. Im letzten Abschnitt (9.3.4) werden als wichtige Nebengütekriterien die Testfairness und die Testakzeptanz betrachtet.

9.3.1 Auswertungen nach Rasch-Modell und Prüfung der Dimensionalität

Bevor die über das Rasch-Modell geschätzten Parameter dargestellt und interpretiert werden, muss die lokale und globale Passung der Items betrachtet werden (vgl. Vorgehen in der Pilotierung Abschnitt 8.3.1). Aufgrund geringer Trennschärfe und schlechter lokaler Modellpassung mussten die Items zwei bis fünf aus dem Inhaltsbereich „Unternehmensführung" ausgeschlossen werden. Nach Ausschluss dieser Items zeigte der LR-Test (Andersen, 1973) mit dem Teilungskriterium „Mittelwert" keine signifikante Verletzung der spezifischen Objektivität an (LR-Test: $\chi^2(19, N = 351) = 21.923, p = 0.23$).

Die Verteilung der Personenfähigkeiten und Aufgabenschwierigkeiten ist in Abbildung 11 in Form einer Wright-Map dargestellt. Die im Test verbliebenen 20 Items verteilten sich gleichmäßig über das Schwierigkeitskontinuum. Die Verteilung der Personenparameter ähnelte einer Normalverteilung. Lediglich im sehr leichten Schwierigkeitsbereich wäre eine dichtere Abdeckung durch Items wünschenswert. Insgesamt ist es im Vergleich zur Pilotierung gelungen, das Schwierigkeitsniveau des Tests leicht anzuheben.

	Jedes x repräsentiert 0.6 Personen	Itemnr.	Skala
Testeigenschaften:			2
Mittlerer Personenparameter: -0.02			
Standardabweichung: 0.85		20	
EAP/PV Reliabilität: .60	x		
Trennschärfen: .21 bis .47	x		
Nach Verlängerung (+6 Items) α = .71	xxxx		
wMNSQ: .96–1.05	xx		
(alle T-Werte < 1.96)	xxxxxx	12	
	xxxxxx		
	xxxxx	8	
	xxxxxxx	9	1
	xxxxxxxxx		
	xxxxxxxxxxxx		
	xxxxxxxxxxxx		
	xxxxxxxxxxxxxxxxxxxx	6 15	
	xxxxxxxxxxxxxxxxxxxxxx		
	xxxxxxxxxxxxxxxxxxxx		
	xxxxxxxxxxxxxxx	4	
	xxxxxxxxxxxxxxxx	10	
	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx	2	
	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx		
	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx		
	xxxxxxxxxxxxxxxxxxxxxxxxxxxx		0
	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx		
	xxxxxxxxxxxxxxxxxxxxxxxxxxxx	5 18	
	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx	17	
	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx	1	
	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx		
	xxxxxxxxxxxxxxxxxxxxxxxxxxx		
	xxxxxxxxxxxxxxxxxxxxxxxxxxx	16 19	
	xxxxxxxxxxxxxxxxxxx		
	xxxxxxxxxxxxxx		
	xxxxxxxxxxxxxx		
	xxxxxxxxxxx	13	
	xxxxxxxxx	7	–1
	xxxxxxx	11	
	xxxx		
	xx	3	
	xxxx		
	xx		
	xxxx		
	x		
	x	14	
	xx		
	x		
			–2

Abb. 11: Darstellung der Ergebnisse der Haupterhebung in Form einer Wright-Map (rechts in der Abbildung) und zentraler Testeigenschaften (links in der Abbildung)

Die EAP/PV-Reliabilität der Testversion mit 20 Items fiel mit .60, wie in der Pilotierung, zu gering aus. Im Entwicklungsstadium einer Skala gelten Werte um . 60 zwar als akzeptabel (Nunnally, 1967). Um jedoch den Standards wissenschaftlicher pädagogisch-psychologischer Diagnostik zu entsprechen, wird ein Wert von mindestens .70 gefordert (AERA et al., 2008). Eine um sechs Items verlängerte Testversion des situativen Tests wurde in einer Zusatzstudie zur

Testakzeptanz von 35 Testteilnehmern bearbeitet. Die verlängerte Testversion (26 Items nach oben beschriebenem Itemausschluss), erreichte ein Cronbachs Alpha von .71, womit der verlängerte Test als ausreichend messgenau eingestuft werden konnte (George & Mallery, 2000). Eine detaillierte Darstellung der Itemkennwerte aller Items befindet sich im Anhang E, Tabelle E-1. Neben dem einfachen Rasch-Modell, in das nur dichotome Antwortmuster eingehen, wurde geprüft, ob die Vergabe von Teilpunkten für das Ankreuzen der teilrichtigen Antwortoption zu einer höheren Messgenauigkeit führt. Der Lösungsschlüssel für richtige und teilrichtige Antworten ist in Anhang E, Tabelle E-2 hinterlegt. Eine Auswertung über das für diese Art von Daten vorgesehene Partial-Credit Modell von Masters (1982) ergab zwar mit .65 eine etwas bessere EAP/PV Reliabilität, zeigte aber auf Itemebene schlechte Modellpassungen, so dass es verworfen wurde.

Die relativ geringe Reliabilität war möglicherweise die Folge von unkontrollierten Störeinflüssen während der Testung. Sie könnte aber auch die Folge heterogener Leistungsanforderungen zwischen den Items (Abele et al., 2012) sein und somit auf bisher unentdeckte Mehrdimensionalität der Items hinweisen. Die Prüfung auf Mehrdimensionalität wird im Folgenden beschrieben.

Auf der Basis empirischer Analysen von Ling (2012) und Bothe (2003) wurde davon ausgegangen, dass sich die Testdaten des situativen Tests am besten über ein eindimensionales Modell erklären lassen. Um die Hypothese H1–1 zur eindimensionalen Struktur des Tests zu prüfen, wurde das eindimensionale Modell mit zwei Modellen verglichen, die im Rahmen der Forschung zu kaufmännischen Kompetenzen entwickelt wurden (vgl. Kapitel 4). Für die direkte Vergleichbarkeit der Modelle müssen diese so gewählt werden, dass sie sich hierarchisch aufeinander beziehen lassen. Zur Prüfung der Hypothese H1–1 wurde das vorhergesagte eindimensionale Modell mit folgenden Modellen verglichen:

(1) Mit einem Modell mit zwei interkorrelierten Faktoren, das zwischen Items mit Inhalten zu Finanz- und Rechnungswesen und Aufgaben, die sich auf betriebswirtschaftliche Organisations- und Leistungsprozesse beziehen, unterscheidet (Seeber, 2008).

(2) Mit einem vierfaktoriellen Modell, in dem jeder Inhaltsbereich (Unternehmensführung, Finanz- und Rechnungswesen, Produktion und Marketing) jeweils eine interkorrelierte Dimension bildet.

Der Modellierung der mehrdimensionalen Modelle wurde das Multidimensional Random Coefficients Multinomial Logit Model (Adams, Wilson & Wang, 1997) zugrunde gelegt. Das Modell ist in ConQuest (Wu et al., 1998) implementiert und stellt eine mehrdimensionale Erweiterung des Rasch-Modells dar.

Unter dem Modell werden Variablen mit mehreren latenten Dimensionen untersucht und können mit einem eindimensionalen Modell, das auf den gleichen Daten und gleicher Itemzahl aufbaut, verglichen werden (Wu, Adams & Wilson, 2001). Eine formale Beschreibung des Modells befindet sich in Anhang F, Formel 5.

In das eindimensionale Modell gingen die 20 situativen Items ein, die einer einzigen Dimension zugeordnet wurden. Die Anzahl der geschätzten Parameter ergibt sich aus der Anzahl der geschätzten Schwierigkeitsparameter sowie der Schätzung eines Populationsmittelwertes und der Varianz. Im zweidimensionalen Modell wurden sieben Items der Skala „Finanz- und Rechnungswesen" zugeordnet, die verbleibenden 13 Items bildeten die Skala „betriebswirtschaftliche Organisations- und Leistungsprozesse". Zusätzlich zu den Itemparametern wurden im zweidimensionalen Modell zwei Populationsmittelwerte, zwei Varianzschätzer und ein Kovarianzschätzer ermittelt. Im vierdimensionalen Modell wurden zwei Items der Skala „Unternehmensführung", sieben Items der Skala „Finanz- und Rechnungswesen", vier Items der Skala „Produktion" und sieben Items der Skala „Marketing" zugeordnet. Die Zahl der geschätzten Parameter ergibt sich wie im zweifaktoriellen Modell aus der Anzahl der Itemparameterschätzung plus Populations-, Varianz- und Kovarianzschätzer pro Dimension. Die Deviance oder Diskrepanz (Bühner, 2011) ist eine Statistik, die angibt, wie gut das Modell auf die Daten passt. Die Modellpassung verschiedener Modelle kann damit relativ verglichen werden (Wu et al., 1998). Sie wird über die Multiplikation der log-Likelihood mit dem Faktor -2 ermittelt (Winther, 2010). Generell gilt: Eine geringe Deviance weist auf eine relativ bessere Repräsentation der Daten anhand des Modells hin. Die Ergebnisse des Modellvergleichs werden in Tabelle 15 dargestellt.

Tab. 15: Vergleich der Passung des eindimensionalen Modells mit zwei alternativen Modellen (N = 351)

	Eindimensionales Modell	**Zweidimensionales Modell**	**Vierdimensionales Modell**
Deviance (–2*log-Likelihood)	8577.34137	8577.07914	8577.54751
Anzahl der geschätzten Parameter	21	23	30
Differenz zum eindimensionalen Mode		0.26223 (Δ *df* = 2)	0.20614 (Δ *df* = 9)
Signifikanz		*p* = n. s.	*p* = n. s.

Auf deskriptiver Ebene zeigte das eindimensionale Modell eine minimal schlechtere Passung als die beiden alternativen Modelle. Ein statistischer Ver-

gleich zwischen den Modellen ist auf Basis der χ^2-Statistik möglich, indem die Differenz der Deviance mit dem kritischen Wert der χ^2-Verteilung verglichen wird. Die Anzahl der Freiheitsgrade ergibt sich aus der Differenz der Anzahl der geschätzten Parameter. Beide alternativen Modelle wiesen keine signifikant ($p \sim .90$) bessere Passung als das vorhergesagte eindimensionale Modell auf. Somit wurde nach dem Prinzip der Einfachheit von Modellen (Bühner, 2011) das eindimensionale Modell den mehrdimensionalen Modellen vorgezogen.

Für die H1–1 wird die Nullhypothese beibehalten. Keines der getesteten alternativen Modelle bildet die Daten besser ab als das eindimensionale Modell.

Es ist anzumerken, dass es sich um eine relative Bewertung der Modellpassung handelte und somit nicht auszuschließen ist, dass es alternative Modelle mit besserer Passung gibt (Bühner, 2011). Um die Annahme der Eindimensionalität der Testitems zu untermauern, wurden weitere Berechnungen durchgeführt. Ein etablierter Test zur Prüfung der Itemhomogenität ist der Martin-Löf-Test (Rost, 2004) (vgl. Abschnitt 7.1.2.2). Dieser prüft, ob die Items zweier Testhälften das gleiche Konstrukt messen. Die Teilung der Items erfolgt nach theoriegeleiteten Kriterien. Sinnvoll ist zum Beispiel die Teilung der Items in eine Gruppe leichter Items und eine Gruppe schwerer Items. Die χ^2-verteilte Prüfgröße des Modelltests ähnelt einem Likelihood-Quotienten-Test (Bühner, 2011). Sowohl die Teilung der Testitems anhand des Medians ($\chi^2(98, N = 351) = 80.701$, $p = 0.91$) als auch anhand des Mittelwerts ($\chi^2(98, N = 351) = 66.908$, $p = 0.99$) wies nicht darauf hin, dass die Items unterschiedliche Konstrukte messen. Eine einfaktorielle konfirmatorische Faktoranalyse für dichotome Daten über alle 20 situativen Items wies eine gute Modellpassung auf und untermauerte die Hypothese der Eindimensionalität ($\chi^2(170, N = 351) = 185.267$, $p = .20$, CFI = .95, RMSEA = .02).

Für eine eindimensionale Interpretation betriebswirtschaftlichen Wissens auf Bachelorniveau sprach ebenfalls, dass die Modulabschlussnoten der Fächer Unternehmensführung, Finanz- und Rechnungswesen, Produktion und Marketing in einer explorativen Faktorenanalyse auf einen Faktor luden (siehe Ergebnisse der Hauptkomponentenanalyse in Tabelle 16). Das heißt, die Leistungen in allen genannten Fächern korrelierten hoch miteinander und wurden mit großer Wahrscheinlichkeit durch eine latente Fähigkeit (betriebswirtschaftliches Wissen auf Bachelorniveau) beeinflusst.

Tab. 16: Ergebnisse der explorativen Faktoranalyse (Hauptkomponentenanalyse) über die Modulnoten in den testrelevanten Modulen für metrische und ordinale Daten

Modulnoten in den Themengebieten:	**Faktorladung (Modulnoten als metrisch verrechnet)**	**Faktorladung (Modulnoten als ordinal verrechnet)**
	Faktor 1	Faktor1
Unternehmensführung	.63	.65
Rechnungswesen	.69	.55
Finanzwirtschaft	.65	.58
Produktion	.76	.70
Marketing	.82	.86
	Eigenwert 1. Faktor 2.55 Anteil der Gesamtvarianz 51 %	Eigenwert 1. Faktor 2.29 Anteil der Gesamtvarianz 46 %

An dieser Stelle ist anzumerken, dass Hochschulnoten streng genommen nicht intervallskaliert sind. Da es sich jedoch etabliert hat, für Schulnoten Intervallskalenniveau anzunehmen (Ingenkamp, 1997), wurden im Rahmen der vorliegenden Arbeit Abschlussnoten der einzelnen Module ebenfalls wie intervallskalierte Merkmale behandelt. Die Ergebnisse der Hauptkomponentenanalyse in Tabelle 16 wiesen darauf hin, dass die Mittelwertbildung über alle Modulnoten hinweg ohne Gewichtung gerechtfertigt war. Die geringfügigen Abweichungen zwischen den beiden Faktoranalysen für metrische und ordinale Daten wiesen zusätzlich darauf hin, dass eine Verrechnung der Noten als metrische Daten nur zu geringen Diskrepanzen in der Auswertung führt (vgl. Tabelle 16).

Nachdem die Binnenstruktur der situativen Items untersucht wurde, werden im Folgenden die Ergebnisse der weiterführenden Konstruktvalidierung dargestellt.

9.3.2 Konstruktvalidierung

Im Rahmen der Konstruktvalidierung wird geprüft, inwiefern sich ein Konstrukt theoriekonform zu anderen Konstrukten verhält (Hartig et al., 2012). Die Prüfung der Konstruktvalidität auf Itemebene wurde mittels eines konfirmatorischen faktoranalytischen Zugangs (Hartig et al., 2012) vorgenommen (Abschnitt 9.3.2.1). In Abschnitt 9.3.2.2 wird beschrieben, wie die Konstruktvalidität über den Vergleich von unterschiedlichen Subgruppen der Stichprobe (Hartig et al., 2012; Jacobs, Heubrock, Petermann, Kubinger & Wurst, 2003) vorgenommen wurde. Als weitere Möglichkeit, um zu prüfen, ob das zu messende Konstrukt durch die Items gut abgebildet wurde, wurde das Verhältnis von schwierigkeitsbestimmenden Aufgabenmerkmalen mit den empirischen

Schwierigkeiten betrachtet (vgl. Abschnitt 4.2.2). Die Ergebnisse dieser Analysen werden in Abschnitt 9.3.2.3 vorgestellt.

9.3.2.1 Analyse der faktoriellen Struktur

Die in der vorliegenden Arbeit entwickelten Items sollten im Sinne der Konstruktvalidität zwei Anforderungen erfüllen:

(1) Die Items sollten neben den Items des BAKT (Bothe, 2003) eine eigene Dimension abbilden.

(2) Die latente Korrelation zwischen den beiden Faktoren (situativer Test und nicht-situativer) sollte substanziell ausfallen.

Die erste Forderung zielte darauf ab, den Mehrwert der neu entwickelten Items als Ergänzung zu den bereits bestehenden Items zu untermauern. Zudem wurde die Bestätigung einer zweidimensionalen Struktur im Sinne der diskriminanten Konstruktvalidierung (Moosbrugger & Kelava, 2012a) als Indikator dafür gesehen, dass die situativen Items ein anderes Wissenskonstrukt erfassen als die Items des deklarativen BAKTs (Bothe, 2003). Trotzdem sollten beide latenten Wissenskonstrukte stark miteinander zusammenhängen, da beide Tests betriebswirtschaftliches Wissen auf Bachelorniveau erfassen. Im Modellvergleich wurde das vorhergesagte zweidimensionale Modell mit korrelierten Gruppenfaktoren mit zwei alternativen Modellen verglichen: (1) mit einem eindimensionalen Generalfaktormodell und (2) mit einem Modell, in dem über situative und nicht-situative Items themenbezogene (Unternehmensführung, Finanz- und Rechnungswesen, Produktion und Marketing) Dimensionen gebildet werden.

Der BAKT (Bothe, 2003) wurde nicht unter der Annahme des Rasch-Modells entwickelt. Entsprechend zeigt der Andersen-Test (Andersen, 1973), dass die spezifische Objektivität (Teilungskriterium Mittelwert) bei den Items des BAKT (Bothe, 2003) verletzt ist (LR-Test: χ^2 (22, $N = 351$) = 62.214, $p < .001$). Als Konsequenz der Modellverletzung musste zur Prüfung der Dimensionalität der situativen Items im Verhältnis zu den Items des deklarativen BAKTs (Bothe, 2003) auf probabilistische Modellvergleiche verzichtet werden. Alternativ wurde der Modellvergleich auf Basis von konfirmatorischen Faktoranalysen für dichotome Daten durchgeführt (Muthén & Muthén, 2010).

In einer konfirmatorischen Faktorenanalyse wird vorab ein Modell darüber spezifiziert, welche manifesten Variablen (Items) auf welchen latenten Faktor laden und wie die latenten Faktoren miteinander zusammenhängen (Bühner, 2011). Ziel der konfirmatorischen Faktorenanalyse ist es, Parameter so zu schätzen, dass sie in einer linearen Gleichung die empirische Varianz-/Kovari-

anzmatrix bestmöglich reproduzieren (Bühner, 2011). Je ähnlicher die modellimplizierte Kovarianzmatrix der empirischen Kovarianzmatrix ist, umso besser beschreibt das Modell die Daten (Bühner, 2011). Die Bewertung der Güte der Modelle erfolgt anhand etablierter Grenzwerte (Hu & Bentler, 1999). Mittels Likelihood-Ratio-Test wird die Nullhypothese geprüft, dass die empirische Kovarianzmatrix der modelltheoretischen Kovarianzmatrix entspricht (Backhaus et al., 2003). Der χ^2-Wert sollte im Verhältnis zu seinen Freiheitsgraden möglichst klein sein, und die Wahrscheinlichkeit (p), dass die Ablehnung der Nullhypothese eine Fehlentscheidung ist, sollte bestenfalls größer .10 (Backhaus et al., 2003) und mindestens gleich .05 (Marsh, Hau & Weng, 2004) sein. Da der Likelihood-Ratio-Test vielfach kritisiert wurde (unter anderem von Bühner (2011) für seine Abhängigkeit von der Stichprobengröße), werden zur Bewertung der Modellpassung standardmäßig zwei weitere Kennwerte herangezogen: (1) der Comparative Fit Index (CFI) und (2) der Root Mean Square Error of Approximation (RMSEA) (Rigdon, 1996). Der CFI vergleicht das spezifizierte Modell mit einem Basismodell und zieht dabei die Zahl der Freiheitsgrade in Betracht (Backhaus et al., 2003). CFI-Werte werden je nach Auffassung ab .90 als gut (Backhaus et al., 2003) oder lediglich als akzeptabel betrachtet (Hu & Bentler, 1999). Der RMSEA gibt an, ob das Modell die Realität hinreichend abbildet (Backhaus et al., 2003), Werte < .08 gelten insbesondere bei kleinen Stichproben (Bühner, 2011) als akzeptabel, und Werte < .05 weisen auf eine gute Modellpassung hin (Backhaus et al., 2003). Die miteinander verglichenen Modelle und ihre Modellpassungsstatistiken sind in Tabelle 17 aufgeführt.

Tab. 17: Vergleichende Darstellung der Modellpassung unterschiedlicher Modelle über die Items des BAKT (Bothe, 2003) und des situativen Tests (N = 351)

Modell	*df*	χ^2	*p*	CFI	RMSEA
Ein Generalfaktor	860	927.617	.05	.903	.015
Zwei korrelierte Gruppenfaktoren	859	919.117	.08	.914	.014
Vier korrelierte Gruppenfaktoren	852	913.946	.07	.910	.014

Die Parameterschätzung im Modell wurde über einen wLSMV[13]-Schätzer vorgenommen. Dieser schätzt auf der Grundlage gewichteter kleinster Quadrate mit adjustiertem Mittelwert und adjustierter Varianz. Dieser Schätzalgorithmus liefert für kleine Datensätze mit kategorialen Daten robuste Schätzungen (Urban, 2004). Der Vergleich der Modellpassung über die Differenzen der χ^2-Werte ist jedoch bei diesem Schätzverfahren nicht zulässig, da die Differenz nicht

13 weighted least squares means and variance adjusted

χ^2-verteilt ist (Muthén & Muthén, 2010). Stattdessen müssen als alternative Indizes der CFI und der RMSEA betrachtet werden.

Unter Betrachtung des CFI und des RMSEA wiesen alle Modelle eine akzeptable Passung auf. Theoretisch wäre es zu rechtfertigen, die Items sowohl über ein einfaktorielles Modell als auch über ein zweifaktorielles Modell zu beschreiben. In der vorliegenden Arbeit wurde das zweifaktorielle Modell den alternativen Modellen aufgrund des besseren CFI, des geringeren RMSEA und des Sparsamkeitsprinzips (Bühner, 2011) vorgezogen.

Der latente Zusammenhang zwischen den beiden Faktoren betrug $cov = 0.83$ ($SD = 0.05$) und war somit, wie vorhergesagt, als hoch einzustufen. Die Prüfung eines zweifaktoriellen Modells mit einem Faktor 2. Ordnung konnte aufgrund von Schätzproblemen nicht durchgeführt werden.

Für die H1–2 zur zweidimensionalen Struktur situativer und nichtsituativer Items wird die H0 (mit Einschränkungen) verworfen.

Die Einschränkung besteht darin, dass die relativ nah beieinander liegenden Modellpassungsstatistiken auch andere Interpretationen zulassen und somit zwar in dieser Arbeit als Arbeitsgrundlage von einer zweifaktoriellen Struktur ausgegangen wurde, die Hypothese H1–2 darüber hinaus jedoch anhand größerer Item- und Stichprobenzahlen geprüft werden sollte.

9.3.2.2 Subgruppenvergleiche

Eine weitere Methode, um die Konstruktvalidität eines Tests abzusichern, ist der Vergleich von Extremgruppen bezüglich der Ausprägung des gemessenen Merkmals. In der Regel werden vor einem Extremgruppenvergleich Hypothesen darüber aufgestellt, wie sich das gemessene Merkmal bei bestimmten Extremgruppen unterscheidet (Hartig et al., 2012). Aufgrund der begrenzten empirischen und theoretischen Vorarbeiten in Bezug auf betriebswirtschaftliches Wissen bei Studierenden wurden im Rahmen dieser Analysen keine statistischen Hypothesen aufgestellt, sondern lediglich begründete Vorannahmen getroffen. Dieser explorative Charakter der Auswertungen muss bei der Interpretation der inferenzstatistischen Kennwerte einschränkend berücksichtigt werden.

Die erste zu prüfenden Vermutung bezieht sich auf den Unterschied in der Testleistung zwischen den Studiengängen. Ist es gelungen, das Konstrukt domänenspezifisch zu operationalisieren, sollten Studierende der Hauptfächer, für die der Test entwickelt wurde (Betriebswirtschaftslehre und Wirtschaftspädagogik), besser abschneiden als Studierende der Volkswirtschaft oder anderer Studiengänge. Für den Vergleich zwischen Studierenden der Betriebswirt-

schaftslehre und Studierenden der Wirtschaftspädagogik wurden geringfügig bessere Leistungen als für Studierende der Betriebswirtschaftslehre vermutet. Studierende mit dem Hauptfach Betriebswirtschaftslehre verfügen in der Regel über weiterführende Lerngelegenheiten im Fach BWL (z. B. durch Vertiefungsmodule) als Studierende der Wirtschaftspädagogik. Um zu prüfen, ob sich die mittlere Testleistung zwischen den vier Studierendengruppen (Betriebswirtschaftslehre, Wirtschaftspädagogik, Volkswirtschaftslehre und anderer Studiengang) signifikant voneinander unterscheidet, wurde eine Varianzanalyse durchgeführt (zur Beschreibung des Verfahrens siehe Backhaus et al., 2003). Auf deskriptiver Ebene zeigte sich, dass Studierende der Betriebswirtschaftslehre (n = 109; M = 0.13; SD = 0.89) und Studierende der Wirtschaftspädagogik (n = 171; M = 0.08; SD = 0.70) deutlich besser abschnitten als Studierende der Volkswirtschaft (n = 34; M = –0.76; SD = 1.07) oder anderer Studiengänge (n = 5; M = –.89; SD = 1.34). Zwischen Studierenden der Betriebswirtschaftslehre und Studierenden der Wirtschaftspädagogik lagen nur minimale Unterschiede in der Testleistung vor. Um den Einfluss von studiengangspezifischen kognitiven Leistungsunterschieden bei der Analyse der Testleistungen gering zu halten, wurde die Note der Hochschulzugangsberechtigung kontrolliert (F(1,319) = 5.14, p = .02, η_p^2 = .016). Der signifikante Haupteffekt der univariaten Varianzanalyse (ANOVA) bestätigte die Vermutung, dass sich die Leistungen der Studierenden entsprechend ihrem Studiengang unter Kontrolle der Note der Hochschulzugangsberechtigung unterscheiden (F(3,319) = 11.91, p < .001, η_p^2 = .10). Die über die ANOVA geschätzten Mittelwerte und die jeweiligen 95 %- Konfidenzintervalle sind in Abbildung 12 abgetragen.

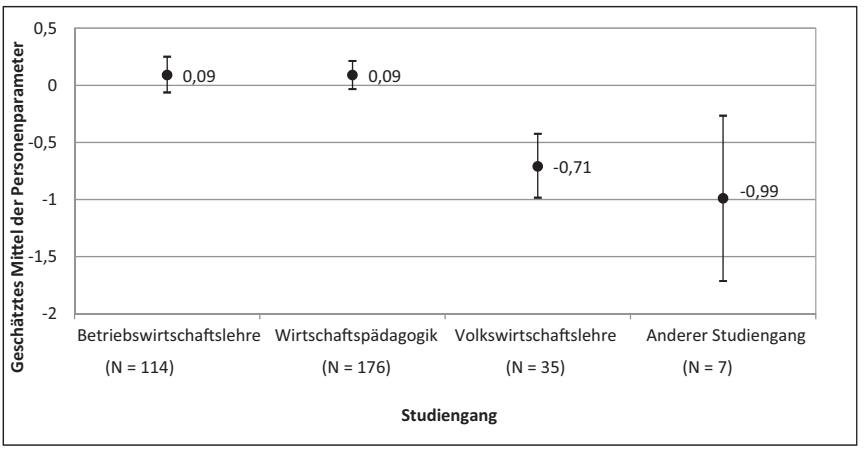


Abb. 12: Unterschiede in der Testleistung in Abhängigkeit des Studiengangs unter Kontrolle der Note der Hochschulzugangsberechtigung (Note = 2.42) mit 95 %-Konfidenzintervall

An den Konfidenzintervallen ist zu erkennen, dass sich die Testergebnisse der Zielgruppen (Betriebswirtschaftslehre und Wirtschaftspädagogik) signifikant von den Testergebnissen von Studierenden anderer Studiengänge unterschieden. Zwischen Studierenden der Betriebswirtschaftslehre und der Wirtschaftspädagogik lagen keine signifikanten Unterschiede vor.

Über Leistungsunterschiede zwischen Studierenden im Bachelor und Studierenden Master konnten keine eindeutigen Vorhersagen getroffen werden. Da der Test spezifisch auf Inhalte des Bachelorstudiums abzielte, könnte angenommen werden, dass Bachelorstudierende besser abschneiden, weil der Inhalt des Tests für diese Gruppe Studierender aktueller ist. Eine gegenläufige Argumentation wäre, dass Studierende im Master über mindestens genauso viel „Bachelorwissen" verfügen wie Bachelorstudierende und darüber hinaus eine höhere Zahl an Modulen besucht haben, was zu besseren Testleistungen führen könnte. Die empirischen Ergebnisse bestätigten weder die eine noch die andere Vermutung. Der Mittelwert der Masterstudierenden (n = 192; M = –0.1; SD = 0.79) lag genau bei dem Mittelwert der Bachelorstudierenden (n = 148; M = –0.1; SD = 0.92).

Ebenso wie die Leistungsunterschiede nach Studienabschnitt wurden die Leistungsunterschiede zwischen Studierenden unterschiedlicher Standorte explorativ betrachtet. Nach Kontrolle der Note der Hochschulzugangsberechtigung (F(1,171) = 5.104, p = .025, η_p^2 = .030) zeigten sich keine signifikanten Unterschiede bezüglich der Testleistung zwischen den Standorten Hamburg (n = 61; M = 0.06; SD = .66), Göttingen (n = 46; M = 0.04; SD = 0.78) und Mannheim (n = 64; M = 0.12; SD = 0.68), (F(2,171) = 0.163, p =.850, η_p^2 = .002). Dieses Ergebnis kann als Hinweis dahingehend interpretiert werden, dass der Test für den Einsatz an verschiedenen Standorten geeignet ist. Einschränkend muss angemerkt werden, dass der Vergleich der standortspezifischen Leistung nur für Studierende der Wirtschaftspädagogik und für relativ geringe, nicht-repräsentative Fallzahlen vorgenommen wurde.

Zuletzt wurde geprüft, inwiefern die Testleistungen davon abhängen, an welcher Position im Testheft die Items dargeboten wurden. Der Vergleich der Leistungsmittelwerte des Testhefts A (situative Aufgaben vor nicht-situativen Aufgaben) und Testfeft B (nicht-situative Aufgaben vor situativen Aufgaben) zeigte auf deskriptiver Ebene einen Trend zu besseren Testleistungen, wenn die situativen Items an erster Position im Testheft bearbeitet wurden (n = 173; M = 0.04; SD = 0.85). Die schlechteren Leistungen, wenn die situativen Items nach den nicht-situativen Items dargeboten wurden (n = 173; M = –0.10; SD = 0.86), sind möglicherweise durch nachlassende Konzentration und Ermüdungseffekte bei den Testteilnehmern zu erklären (Auspurg, Hinz, Liebig &

Sauer, 2009). Der Haupteffekt „Itemposition" war jedoch nicht signifikant ($t(343) = 1.564$, $p = .119$).

9.3.2.3 Analyse der Aufgabenschwierigkeiten

Eine weitere Möglichkeit, um zu überprüfen, ob das zu messende Konstrukt im Rahmen der Testentwicklung richtig operationalisiert wurde, bietet die Betrachtung schwierigkeitsbestimmender Merkmale (vgl. 4.2.2). Dabei geht es darum festzustellen, ob Items, die aufgrund zuvor definierter Merkmale in der Testentwicklungsphase als schwer oder leicht eingestuft wurden, sich tatsächlich empirisch als schwieriger oder leichter erweisen. Bestätigen sich die bei der Konstruktion eines Tests intendierten Schwierigkeiten, spricht das insgesamt für eine gute Passung zwischen den theoretischen Überlegungen zum Konstrukt und dessen Operationalisierung. Darüber hinaus können vordefinierte schwierigkeitsbestimmende Aufgabenmerkmale genutzt werden, um ein Niveaumodell zu entwickeln (Schumann & Eberle, 2011). Von einer Niveaumodellierung wurde in der vorliegenden Arbeit aufgrund der geringen Itemzahl und des frühen Entwicklungsstadiums des Tests abgesehen. Es wurde jedoch geprüft, ob sich die für die Testkonstruktion gewählten schwierigkeitsbestimmenden Aufgabenmerkmale kognitives Anforderungsniveau, Komplexität und mathematische Modellierung (vgl. 4.2.2 und Tabelle 1) empirisch in entsprechenden Aufgabenschwierigkeiten niederschlagen.

Das kognitive Anforderungsniveau der Aufgaben wurde in Anlehnung an bestehende Konzepte aus der berufs- und wirtschaftspädagogischen Forschung in die Stufen Reproduzieren, Anwenden und Reflektieren unterteilt (Hofmeister, 2005; Schumann & Eberle, 2011). Es wurde vermutet, dass die Stufen eine Schwierigkeitshierarchie abbilden. Die Betrachtung der mittleren Schwierigkeitsparameter je Anforderungsdimension bestätigte diese Vermutung nicht. Aufgaben auf dem Niveau „Anwenden" erwiesen sich im Mittel als leichter ($n = 9$, $M = -0.23$, $SD = 1.16$) als Aufgaben mit dem Niveau „Reproduzieren" ($n = 6$, $M = 0.02$, $SD = 0.95$) oder „Reflektieren" ($n = 5$, $M = 0.03$, $SD = 0.60$). Die Unterschiede zwischen den Anforderungsniveaus waren nicht signifikant ($F(2,20) = 0.160$, $p = .854$, $\eta_p^2 = .018$). Abbildung 13 verdeutlicht, dass die Items auf dem Niveau „Anwenden" stärker über das Schwierigkeitskontinuum streuten als Items der anderen Anforderungsniveaus. Zudem wurde die Gruppe der Items auf dem Niveau „Anwenden" im Vergleich zu den anderen beiden Anforderungsniveaus durch eine höhere Itemzahl repräsentiert.

Diese Ergebnisse decken sich mit Befunden, die Witt (2006) zu den Taxonomiestufen des WBT zusammengetragen hat. Auch hier zeigte sich, dass „Anwenden" empirisch die leichteste Prozessdimension war, und entgegen der Annahme, dass Anwenden nur auf der Basis von Verstehen stattfinden kann,

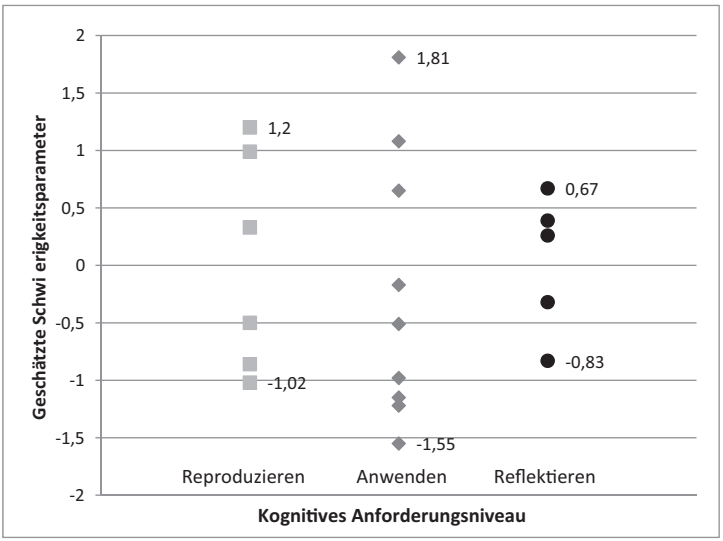


Abb. 13: Schwierigkeitsindizes der Testaufgaben nach kognitiven Anforderungen (jeweils minimale und maximale Schwierigkeiten sind angegeben)

waren Aufgaben auf der Stufe „Verstehen" für die Testteilnehmer deutlich leichter (Witt, 2006). Die bei der Testentwicklung beabsichtigte Varianz in der Komplexität der Aufgaben schlug sich auf deskriptiver Ebene in erwartungskonform variierenden Schwierigkeitsparametern nieder. Geringfügig komplexe Aufgaben, die nur eine zu prüfende Bedingung im Aufgabenstamm enthielten, waren leichter (n = 13, M = –0.16, SD = .60) als Aufgaben, die zwei zu prüfende Bedingungen im Aufgabenstamm enthielten (n = 7, M = 0.04, SD = 1.09). In Abbildung 14 sind die Schwierigkeitsindizes der Testaufgaben je nach bei der Itementwicklung intendierter Komplexität abgetragen. Es zeichnete sich ein schwacher linearer Trend dahingehend ab, dass komplexe Aufgaben empirisch schwieriger sind als weniger komplexe Aufgaben.

Ein Signifikanztest des Mittelwertunterschiedes zwischen hoch komplexen und geringfügig komplexen Aufgaben anhand eines zweiseitigen t-Tests für unabhängige Stichproben klassifizierte den Unterschied jedoch als nicht signifikant (t(18) = –0.438, p = .667).

Aufgaben, die eine mathematische Modellierungsleistung erforderten, waren tendenziell schwerer (n = 6, M =0.28, SD = 1.09) als Aufgaben, in denen keine solche Modellierungsleistung vorgenommen werden musste (n = 14, M = –0.25, SD = 0.87). Der Unterschied war jedoch ebenfalls nicht signifikant (t(18) = 1.156, p = .263). In Abbildung 15 sind die Schwierigkeitsindizes ge-

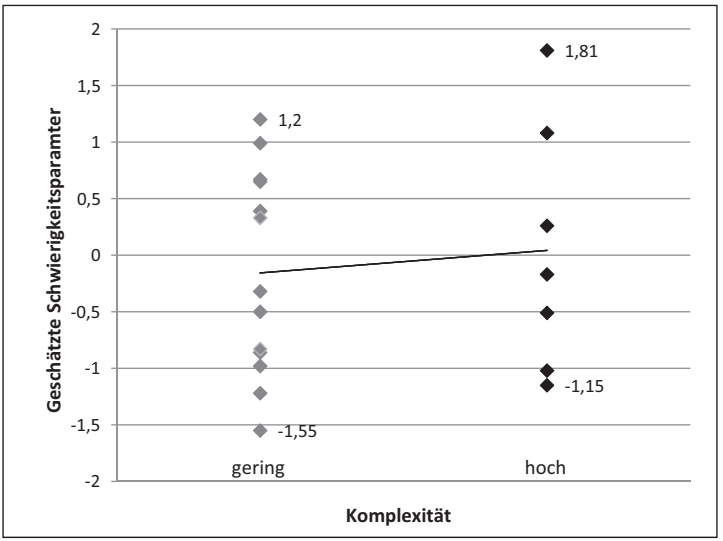


Abb. 14: Schwierigkeitsindizes der Testaufgaben nach Komplexität (die schwarze Linie repräsentiert einen linearen Trend, jeweils minimale und maximale Schwierigkeiten sind angegeben)

trennt nach Items mit und ohne mathematische Modellierungsanforderung abgetragen. Die schwarze Linie repräsentiert den linearen Trend, dass Aufgaben, die eine mathematische Modellierungsleistung erfordern, tendenziell schwieriger waren als Aufgaben, in denen keine mathematische Modellierung vorgenommen werden musste.

Die letzten beiden berichteten Trends für den Zusammenhang zwischen angestrebten Aufgabenschwierigkeiten und empirischen Aufgabenschwierigkeiten gingen in die bei der Testentwicklung intendierte Richtung, was für eine gelungene Operationalisierung der Schwierigkeit innerhalb des Konstrukts spricht. Lediglich die Kategorisierung anhand des kognitiven Anforderungsniveaus gab keinen systematischen Aufschluss über die Aufgabenschwierigkeiten. Ähnlich unbefriedigende Ergebnisse im Zusammenhang mit der Kategorisierung nach kognitivem Anforderungsniveau wurden im beruflichen Fachtest für den Ausbildungsberuf „Kaufmann/Kauffrau im Einzelhandel" in ULME III (Lehmann & Seeber, 2007) gefunden. Eine Diskussion der Befunde unter Einbezug vergleichbarer Studien wird in Kapitel 10 vorgenommen. Vorerst sei einschränkend vermerkt, dass aufgrund der geringen Itemzahl aus den Ergebnissen keine generalisierbaren Aussagen getroffen werden konnten. Darüber hinaus muss beachtet werden, dass die gewählten schwierigkeitsbestimmenden Aufgabenmerkmale möglicherweise nicht unabhängig voneinander sind und somit

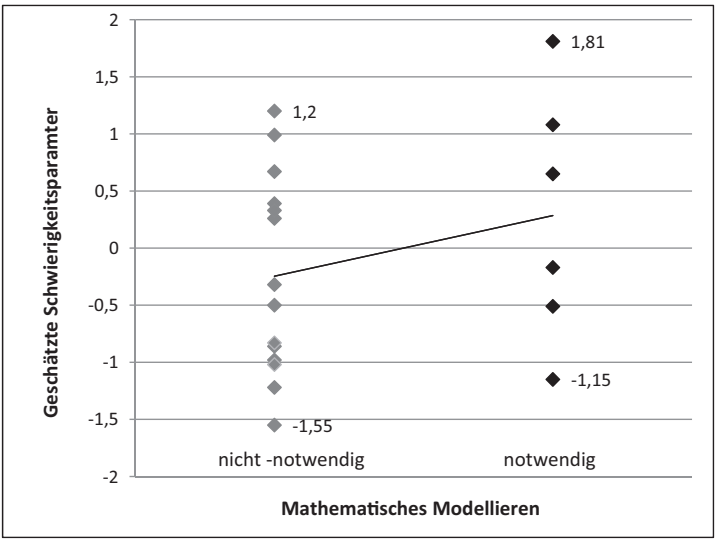


Abb. 15: Schwierigkeitsindizes der Testaufgaben nach mathematischer Modellierungsanforderung (die schwarze Linie repräsentiert einen linearen Trend, jeweils minimale und maximale Schwierigkeiten sind angegeben)

Interaktionseffekte vorliegen könnten, die es mit einer größeren Itemanzahl zu prüfen gilt.

Neben der Betrachtung der Aufgabenschwierigkeiten innerhalb des situativen Tests ist es ebenso interessant zu prüfen, ob situative Aufgaben im Vergleich zu nicht-situativen Aufgaben tatsächlich schwieriger sind, wie in H1–3a vorhergesagt. Zur Testung dieser Hypothese wurde als Schwierigkeitsindikator die prozentuale Lösungsquote der Aufgaben herangezogen. Ein Vergleich auf Basis der über das Rasch-Modell geschätzten Aufgabenparameter war aufgrund der Modellverletzungen der Items des BAKT (Bothe, 2003) nicht möglich (vgl. Abschnitt 9.3.2.1).

In Abbildung 16 wurden die prozeduralen Lösungshäufigkeiten getrennt nach Aufgabentyp abgetragen. Es wurde deutlich, dass die Lösungsquoten der deklarativen Aufgaben breiter streuen als die Lösungsquoten der situativen Aufgaben. Tendenziell waren die situativen Aufgaben schwerer ($n = 20$, $M = 49.74$, $SD = 18.67$) als die deklarativen Aufgaben ($n = 23$, $M = 59.32$, $SD = 23.62$). Die Mittelwerte wichen jedoch nicht statistisch signifikant voneinander ab, $t(41) = 1.461$, $p = .152$.

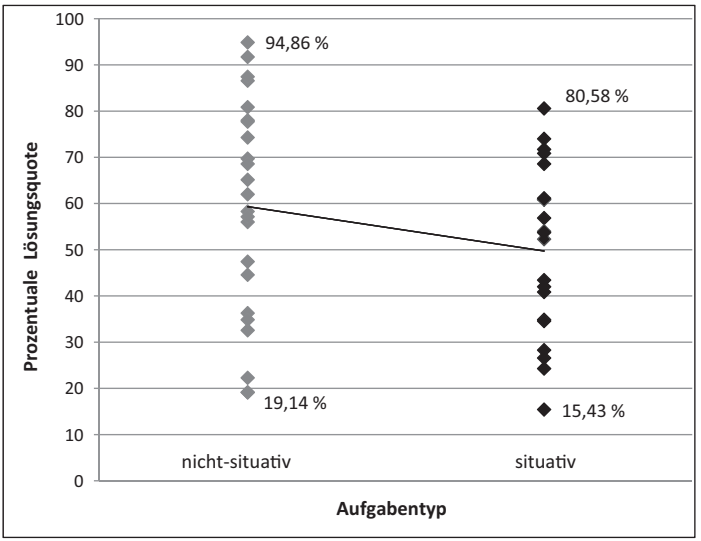


Abb. 16: Prozentuale Lösungshäufigkeit der einzelnen Testaufgaben nach Aufgabentyp (die schwarze Linie repräsentiert einen linearen Trend, jeweils minimale und maximale Lösungsquoten sind angegeben)

Für die Hypothese H1–3a wird die Nullhypothese beibehalten; es zeichnet sich jedoch ein vorhersagekonformer Trend ab.

Um zu prüfen, wie sich die Aufgabentypen auf die mittlere Testleistung auswirken, und ob es eine Wechselwirkung zwischen Aufgabentyp und Bekanntheit mit der Art der Aufgabenstellung gab, wurde zur Testung der Hypothese H1–3b eineANOVA mit wiederholten Messungen gerechnet. Als abhängige Variable ging jeweils der Prozentsatz gelöster Aufgaben pro Aufgabentyp in die Berechnung ein. Als unabhängige Variable wurde eine absolvierte kaufmännische Ausbildung vor dem Studium als Indikator für den Bekanntheitsgrad mit situativen Aufgaben herangezogen.

Es zeigte sich, dass Testteilnehmer mit absolvierter kaufmännischer Ausbildung im Mittel bei beiden Testtypen mehr Aufgaben lösten als Testteilnehmer ohne kaufmännische Ausbildung, $F(1,351) = 30.668$, $p < .001$, $\eta_p^2 = .086$. Ebenso erzielten beide Gruppen bei den situativen Aufgaben weniger Prozentpunkte als bei den nicht-situativen Aufgaben, $F(1, 351) = 36.243$, $p < .001$, $\eta_p^2 = .10$. Besonders hervorzuheben ist der Interaktionseffekt zwischen Aufgabentyp und absolvierter kaufmännischer Ausbildung. Die Schere zwischen der Testleistung im situativen Test und der Testleistung im nicht-situativen Test war für Personen ohne kaufmännische Ausbildung signifikant größer als für Personen mit

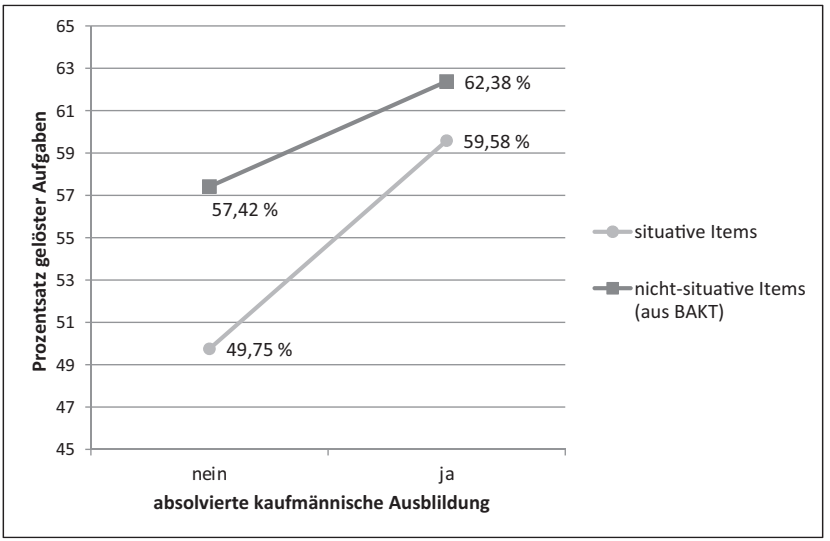


Abb. 17: Prozentsatz gelöster Aufgaben nach Aufgabentyp und absolvierter kaufmännischer Ausbildung

kaufmännischer Ausbildung, $F(1,351) = 7.887$, $p = .005$, $\eta_p^2 = .024$. Dieser Effekt wird dahingehend interpretiert, dass Personen mit einer kaufmännischen Ausbildung vertrauter mit situativen Aufgabenstellungen sind und ihre berufspraktische Erfahrung die Lösung der situativen Aufgaben begünstigt.

Die Nullhypothese H1–3b zum Interaktionseffekt von Aufgabentyp und kaufmännischer Vorerfahrung wird verworfen.

Nachdem mehrere Berechnungen zur Etablierung der Konstruktvalidität erfolgreich durchgeführt wurden, folgt im nächsten Abschnitt die empirische Analyse der Kriteriumsvalidität.

9.3.3 Kriteriumsvalidierung

Im Rahmen der Kriteriumsvalidierung wird geprüft, ob die Testwerte in vorhersagekonformen Zusammenhängen mit relevanten Außenkriterien stehen. Dabei soll der Zusammenhang mehrerer aus Theorie und Empirie abgeleiteter Prädiktoren und der Testleistung ermittelt werden. Alle Hypothesen zur kriterialen Validität wurden über eine multiple Ordinary Least Square (OLS)-Regression getestet (zum Verfahren: Backhaus, Erichson, Plinke & Weiber, 2003; Fahrmeir, Kneib & Lang, 2009). Die Ergebnisse wurden mit einer OLS-Regression auf die Testergebnisse des BAKTs (Bothe, 2003) kontrastiert, um auf explorativer Ebe-

ne Gemeinsamkeiten und Unterschiede zwischen den Vorhersagemustern beider Tests zu identifizieren.

In einer multiplen Regression wird durch eine Linearkombination mehrerer Prädiktoren eine abhängige Variable vorhergesagt (Bühner & Ziegler, 2009). Ein Hauptziel der Regressionsanalyse besteht darin, den Einfluss erklärender Variablen auf den Mittelwert der Zielgröße zu untersuchen (Fahrmeir, Kneib & Lang, 2009). Ein Vorteil der Regressionsanalyse im Vergleich zur Betrachtung bivariater Korrelationen ist, dass der Einfluss jedes Prädiktors auf die abhängige Variable unter Kontrolle der anderen Prädiktoren ermittelt werden kann (Fahrmeir, Kneib & Lang, 2009). Die Regressionsanalyse wurde im Rahmen dieser Untersuchung nicht genutzt, um eine möglichst genaue Vorhersage der Testwerte zu erzielen. Vielmehr wurden die Regressionsgewichte genutzt, um zu ermitteln, ob und mit welchem Gewicht die ausgewählten Prädiktoren mit der Testleistung zusammenhingen. Dabei wurde implizit eine kausale Beziehung zwischen Prädiktor und Testleistung unterstellt. Die Regressionsanalyse war im Rahmen des vorliegenden Forschungsdesigns mit nur einem Messzeitpunkt jedoch kein geeignetes Instrument, um Kausalität zu bestätigen. Entsprechend sollten alle in der Regression gefundenen Zusammenhänge unter kritischer Betrachtung der Wirkrichtung interpretiert werden. Auf diese Problematik wird in der Diskussion (Kapitel 10) der Arbeit erneut eingegangen.

Mittels der Regressionsanalyse wurden die Hypothesen H2–1 bis H2–8 (vgl. Abschnitt 9.1.2) geprüft. Als Indikatoren für die Testvalidität, also als Prädiktorvariablen, wurden analog zu dem Vorgehen von Förster et al. (2012) eine absolvierte kaufmännische Ausbildung (ja = 1, nein = 0), die Anzahl (1–5) und Durchschnittsnote der belegten BWL-Pflichtmodule, die durchschnittliche Note der Hochschulzugangsberechtigung und das Geschlecht (weiblich = 0, männlich = 1) herangezogen. Darüber hinaus wurden das Interesse an betriebswirtschaftlichen Studieninhalten (Skala von 1–4) (Schiefele et al., 1993), der Lerntyp (überwiegend Oberflächenstrategien = 0, überwiegend Tiefenstrategien = 1) (Wild & Schiefele, 1994 zit. nach Hußtegge, 2011), die Leistungs- und Wettbewerbsmotivation (1–6) (Wild et al., 2005 zit. nach Hußtegge, 2011) und eine Skala zum Selbstkonzept (1–6) (akademisch und mathematisch) (March & O'Neill, 1984 zit. nach Hußtegge, 2011) als mögliche Prädiktoren herangezogen. Die Skalen zu Studieninteresse, Leistungsmotivation und Selbstkonzept wurden als metrisch skaliert angenommen.

Im ersten Schritt der Auswertung wurden die korrelative Beziehung der ausgewählten Prädiktoren zur Testleistung und deren Interkorrelationen betrachtet (vgl.Tabelle 18). Prädiktoren, die keinen korrelativen Zusammenhang mit der abhängigen Variablen (Testleistung) aufwiesen oder zu hoch mit anderen Prädiktoren korreliert waren, wurden nicht in die Regressionsgleichung aufge-

nommen (Rumsey, 2008). Zu den Variablen, die nicht weiter betrachtet wurden, gehörten Lerntyp, Leistungs- und Wettbewerbsmotiv und mathematisches Selbstkonzept, denn sie standen in keinem Zusammenhang mit der Testleistung. Die selbstberichtete Verwendungshäufigkeit der Lernstrategien stand auch ohne Dichotomisierung in die zwei Gruppen „überwiegend Verwendung von Tiefenstratgien" und „überwiegend Verwendung von Oberflächenstrategien" in keinem signifikanten Zusammenhang mit der Testleistung (vgl. Anhang F, Tabelle F- 1). Mögliche Begründungen für diese nicht-vorhersagekonformen Ergebnisse werden in Abschnitt 10.2 aufgeführt und diskutiert. Das akademische Selbstkonzept war, wie vorhergesagt, positiv mit der Testleistung korreliert, wies aber zudem relative hohe Korrelationen mit anderen Prädiktoren, wie zum Beispiel der Note der Hochschulzugangsberechtigung, auf. Da Prädiktoren in einer Regressionsrechnung möglichst unabhängig voneinander sein sollten (Rumsey, 2008), wurde das Selbstkonzept nicht in die Regressionsgleichung aufgenommen. Da weder das Studienfach noch der Studienstandort signifikant mit der Testleistung in Verbindung standen (vgl. 9.3.2.2), wurden diese beiden Variablen nicht in die Regression aufgenommen.

Tab. 18: Interkorrelationsmatrix potenzieller Prädiktoren zur Vorhersage der Testleistung (Korrelationen mit der Tesleistung sind grau markiert)

		1	2	3	4	5	6	7	8	9	10	11	12
[1] Leistung im situativen Test	r	1											
	N	351											
[2] Geschlecht weiblich	r	–,078*	1										
	N	350	418										
[3] Anzahl relevanter Module	r	,164**	,230**	1									
	N	325	325	325									
[4] Note in relevanten Modulen	r	–,151**	–,021	–,017	1								
	N	302	301	300	302								
[5] Note Hochschulzugangsberechtigung	r	–,141**	,003	–,003	,161**	1							
	N	333	393	308	286	393							
[6] Absolvierte kaufm. Ausbildung	r	,287**	,078	,257**	–,034	,201**	1						
	N	329	329	311	288	316	329						
[7] Interesse an BWL	r	,145**	–,090*	,010	–,129*	–,052	–,159**	1					
	N	335	334	319	296	319	322	335					
[8] Lerntyp Tiefenstrategien	r	–,003	–,224**	–,138**	–,041	,018	–,065	,109*	1				
	N	319	318	304	284	305	306	319	319				
[9] Wettbewerbsmotivation	r	,034	–,036	–,067	–,169**	–,124*	–,074	,264**	,102	1			
	N	266	265	252	239	256	258	266	252	266			
[10] Leistungsmotivation	r	,076	,203**	,027	–,133*	–,107*	,006	,176**	,023	,545**	1		
	N	268	267	254	241	258	259	268	254	266	268		
[11] Akademisches Selbstkonzept	r	,189**	–,155**	,005	–,232**	–,328**	,015	,190**	,173**	,255**	,273**	1	
	N	263	262	249	236	253	254	263	253	260	261	263	
[12] Mathematisches Selbstkonzept	r	,031	–,182**	–,099	–,137*	–162**	–,201**	,172**	,106*	,238**	,170**	,272**	1
	N	263	262	249	236	253	254	263	253	260	261	263	263

*Die Korrelation ist auf dem Niveau von 0,05 (1-seitig) signifikant. **Die Korrelation ist auf dem Niveau von 0,01 (1-seitig) signifikant.

Die verbleibenden Prädiktoren wurden in einem zweiten Schritt genutzt, um die Testleistung für die situativen Items und die Testleistung für die Items des deklarativen Wissenstest vorherzusagen. Die Ergebnisse der multiplen OLS-Regression sind in Tabelle 19 abgetragen.

Tab. 19: Multiple OLS-Regression zur Vorhersage der geschätzten Personenparameter im situativen betriebswirtschaftlichen Wissenstest im Vergleich zu der Vorhersage der Leistung in Items des BAKT (Bothe, 2003) von N = 272 Studierenden

	Leistung im situativen BWL-Test (geschätzte Personenparameter)			**Leistung im BAKT (Summe der Rohpunkte)**		
Variable	***ß***	**stand. *ß***	***SE***	***ß***	**stand. *ß***	***SE***
Konstante	–.07			15.5		
Anzahl der besuchten Pflichtmodule	.09*	.13*	.04	.00	.00	.14
Note in den Pflichtmodulen	–.17*	–.12*	.08	–.31	–.06	.29
Note Hochschulzugangs-berechtigung	–.27**	–.17**	.09	–1.27***	–.23***	.32
Kaufmännische Ausbildung	.55***	.32***	.10	1.49***	.25***	.36
Weibliches Geschlecht	–.23*	–.13*	.10	–1.56***	–.26***	.34
Interesse an BWL-Inhalten	.27**	.18**	.08	.92**	.18**	.29
korr. R^2	.19			.17		
F	11.41			10.36		

* p < .05 ** p < .01 ***p < .001
ß-Gewichte zwischen den beiden Regressionsmodellen dürfen aufgrund der unterschiedlichen Metrik nicht direkt miteinander verglichen werden

Im Sinne der curricularen Validität stand die Anzahl der besuchten Pflichtmodule in einem positiven Zusammenhang mit der Testleistung im situativen Test. Das heißt, je mehr testrelevante BWL-Module die Testteilnehmenden belegt hatten, umso besser schnitten sie im Test ab. Ebenfalls im Sinne der curricularen Validität ist der Zusammenhang zwischen der Durchschnittsnote in den besuchten relevanten Bachelormodulen und der Testleistung zu interpretieren. Je besser die Noten in den testrelevanten Modulen, umso besser ist auch die Testleistung. Das negative *ß*-Gewicht kam dadurch zustande, dass im deutschen Hochschulsystem kleine Zahlen für bessere Noten stehen als große Zahlen. Die Note der Hochschulzugangsberechtigung, als Indikator der allgemeinen kognitiven Leistungsfähigkeit, hing, wie in H2–2 vorgesagt, negativ mit der Testleistung im situativen Test zusammen. Das heißt, Studierende mit guten Noten in

ihrer Hochschulzugangsberechtigung schnitten auch im Test gut ab. Das standardisierte β von –.17 wies im Vergleich zu den Regressionsgewichten der anderen Prädiktoren auf einen Einfluss mittlerer Stärke hin.

Im Einklang mit früheren empirischen Ergebnissen (z. B. Förster et al., 2012) schnitten Personen, die eine kaufmännische Ausbildung vor dem Studium absolviert hatten, in dem Test besser ab als Studierende ohne eine kaufmännische Ausbildung. Weibliche Testteilnehmer wiesen hingegen signifikant schlechtere Testleistungen auf als männliche. Dies war sowohl für die situativen Aufgaben als auch für die deklarativen Aufgaben der stärkste Effekt. Fachspezifisches Interesse stand in einem positiven Zusammenhang mit der Testleistung. Die Ergebnisse der Regressionsanalyse in Bezug auf die aufgestellten Hypothesen (vgl. Abschnitt 9.1.2) sind in Tabelle 20 dargestellt.

Tab. 20: Interpretation der Ergebnisse der Regressionsanalyse mit Bezug auf die aufgestellten Hypothesen

Hypothese	Vorhergesagter Zusammenhang	H0
H2–1a H2–1b	positiver Zusammenhang zwischen Anzahl der besuchten BWL-Module und Testleistung sowie negativer Zusammenhang zwischen Note in den BWL-Modulen und Testleistung (curriculare Validität)	**abgelehnt**
H2–2	negativer Zusammenhang zwischen der Note Hochschulzugangsberechtigung und Testleistung	**abgelehnt**
H2–3	positiver Zusammenhang zwischen absolvierter kaufm. Ausbildung und Testleistung	**abgelehnt**
H2–4	negativer Zusammenhang zwischen weiblichen Geschlecht und Testleistung	**abgelehnt**
H2–5	positiver Zusammenhang zwischen Interesse und Testleistung	**abgelehnt**
H2–6	positiver Zusammenhang zwischen Tiefenlernstrategien und Testleistung	**beibehalten**
H2–7	positiver Zusammenhang zwischen Leistungs- und Wettbewerbsmotivation und Testleistung	**beibehalten**
H2–8	positiver Zusammenhang zwischen akademischem und mathematischem Selbstkonzept und Testleistung	**teilweise abgelehnt**

Die Regressionsgewichte zwischen der Regression auf die situativen Items und der Regression auf die deklarativen Wissensitems durften aufgrund der unterschiedlichen Metrik der abhängigen Variablen nicht direkt miteinander verglichen werden. Das Vorhersagemuster und die Signifikanzen wurden jedoch deskriptiv verglichen. Die Regressionsgewichte auf die Summenwerte des BAKT (Bothe, 2003) und auf die situativen Items wiesen auf ähnliche Beziehungen der beiden Tests zu den untersuchten Außenkriterien hin. Lediglich die Anzahl

der besuchten Pflichtmodule und die mittlere Note in den Pflichtmodulen standen in keinem signifikanten Zusammenhang mit der Testleistung im BAKT (Bothe, 2003). Dieser Hinweis auf nicht ausreichende curriculare Validität ist möglicherweise auf die Entstehung der BAKT-Items auf Basis der alten Diplomstudiengänge (Größler et al., 2002) zurückzuführen. Möglicherweise bedeutet das aber auch, dass in den Modulabschlussprüfungen mehr als nur deklaratives Wissen erfasst wird.

9.3.4 Analyse der Testfairness und Testakzeptanz

Neben den klassischen Gütekriterien haben sich sogenannte Nebengütekriterien für die Beurteilung von Tests etabliert. Die Beurteilung der Fairness des Tests nimmt dabei eine zunehmend wichtige Rolle ein (AERA et al., 2008). Auch die Akzeptanz des Tests von Seiten der Testteilnehmenden ist ein wichtiger Aspekt bei der Beurteilung der Güte von Tests (Kersting, 2008; Schmidt-Atzert, 2008). Insbesondere, weil eine hohe Testakzeptanz mit hoher Testmotivation assoziiert wird. Im ersten Abschnitt dieses Unterkapitels wird über die Ergebnisse der Untersuchungen zu Differential Item Functioning als ein Aspekt der Testfairness berichtet. Im zweiten Abschnitt werden unterschiedliche Facetten der Testakzeptanz der situativen Items dargestellt und mit den Ergebnissen der Einschätzung der nicht-situativen Items kontrastiert.

9.3.4.1 Differential Item Functioning

Differential Item Functioning (DIF) beschreibt unterschiedliche Itemfunktionsweisen in Abhängigkeit einer Stichprobensubgruppe (z. B. Gruppe der weiblichen vs. Gruppe der männlichen Testteilnehmer) (Osterlind & Everson, 2009). Es gibt zahlreiche statistische Methoden, die zur Identifikation von DIF auf Item- und auf Testebene herangezogen werden können (Osterlind & Everson, 2009; für einen ausführlichen Methodenüberblick siehe Penfield & Lam, 2000). Im Rahmen dieser Arbeit wurden DIF-Analysen auf der Basis des Item-Response-Modells durchgeführt. Für die Betrachtung von DIF auf Item-Level wurde der Wald-Test herangezogen (Strobl, 2012). Im Wald-Test werden pro Subgruppe Itemparameter geschätzt, deren Abweichung voneinander über die T-Statistik auf Signifikanz geprüft wird. Weichen zwei Parameter signifikant voneinander ab, bedeutet dies, dass das entsprechende Item für die getesteten Subgruppen unterschiedliche Item Characteristic Curves aufweist (Strobl, 2012). Das heißt, die Items "funktionieren" in beiden Subgruppen unterschiedlich. DIF kann ein Indikator dafür sein, dass neben dem im Rahmen der Messung anvisierten Zielkonstrukt noch weitere unbekannte latente Konstrukte erfasst wurden, die systematisch zwischen den Subgruppen variieren (Cohen & Bolt, 2005). Liegt auf Itemebene DIF vor, ist das Item, bezogen auf das

Zielkonstrukt, kein fairer Bestandteil der Messung und sollte überarbeitet oder aus dem Test entfernt werden (Westers & Kelderman, 1992). Signifikantes DIF ist nicht nur ein Problem in Bezug auf die Testfairness, sondern stellt strenggenommen auch eine Verletzung der Modellannahmen spezifische Objektivität und Eindimensionalität im Rasch-Modell dar (vgl. Abschnitt 7.1.2.2).

Allerdings gibt es kaum einen Test, der kein DIF aufweist (Osterlind & Everson, 2009), und für die Beurteilung des Ausmaßes von DIF liegen noch keine allgemein etablierten Konventionen vor. Items wurden in dieser Arbeit als kritisch eingestuft, wenn die aus den Subgruppen geschätzten Parameter signifikant voneinander abwichen. Zur Testung der Signifikanz wurde die quadrierte Differenz der Parameterschätzungen zur Summe der Varianz der Schätzer in Beziehung gesetzt (Strobl, 2012). Die daraus entstehende Prüfgröße folgt einer Standard-Normalverteilung (Strobl, 2012). Anhand der z-Verteilung können die Überschreitungswahrscheinlichkeiten ermittelt werden. Um über die Signifikanztestung hinaus einen Eindruck davon zu bekommen, um wie viele Logits die Schwierigkeitsparameter von einer Subgruppe zur anderen Subgruppe abweichen, wurde zusätzlich der über ConQuest (Wu et al., 1998) geschätzte Interaktionsparameter zwischen dem Item und der jeweiligen Substichprobe (item*"Subgruppe"-Parameter) ermittelt.

Im Rahmen der vorliegenden Studie war der Vergleich von mehreren Subgruppen interessant. (1) wurden die Items auf geschlechtsspezifisches DIF geprüft. (2) wurde untersucht, ob sich Items standortspezifisch unterschieden, und (3) wurde auf DIF-Effekte zwischen den Studiengängen BWL und Wirtschaftspädagogik getestet. Die DIF-Analysen wurden in R unter Nutzung des Pakets eRm (Mair, Hatzinger & Maier, 2012) sowie in ConQuest (Wu et al., 1998) durchgeführt. Die Ergebnisse sind in Anhang F, Tabelle F-2 dargestellt. Aus den Kennwerten in Anhang F, Tabelle F-2 wurde ersichtlich, dass von insgesamt 20 situativen Items sechs bedenkliches geschlechtsspezifisches DIF aufwiesen (Itemnr: 7, 10, 11, 13, 19, 22). Dieser Befund kann durch einen grafischen Modelltest verdeutlicht werden (vgl. Anhang F, Abbildung F-1). Der grafische Modelltest basiert auf dem Vergleich der geschätzten Aufgabenparameter in zwei Personengruppen: den männlichen und den weiblichen Testteilnehmern. Unter der Geltung des Rasch-Modells und unter der Annahme, dass kein geschlechtsspezifisches DIF vorliegt, müssten die Schätzungen der beiden Gruppen, wenn sie grafisch abgebildet werden, auf einer Geraden liegen (Strobl, 2012). Ab welcher Entfernung von der Winkelhalbierenden eine bedeutsame Abweichung besteht, kann über zweidimensionale Konfidenzintervalle grafisch dargestellt werden (Strobl, 2012). Liegt ein Item so weit von der Winkelhalbierenden entfernt, dass die Konfidenz-Ellipse die Winkelhalbierende nicht schneidet, ist von einer signifikanten Abweichung auszugehen (Strobl, 2012). Problematisch ist

jedoch, wie bei vielen Modelltests, dass bei großen Stichproben auch kleinste Abweichungen signifikant werden. Deshalb ist es sinnvoll, zusätzlich den geschätzte Item*"Subgruppe"-Parameter aus ConQuest (Wu et al., 1998) zu betrachten. Dieser gibt an, um wie viele Logits die beiden Subruppen voneinander abweichen. Die entsprechenden Kennwerte für alle Items sind in Anhang F, Tabelle F-2 angegeben. Zwei Items (10 und 11) wurden aufgrund ihres Unterschieds in den geschätzten Parametern als bedenklich eingestuft, verblieben aufgrund des relativ geringen Effekts aber im Itempool (Wu et al., 1998). Zwei der bereits als problematisch identifizierten Items und ein weiteres Items wiesen standortspezifische DIF-Effekte auf (Itemnr: 8, 19 und 22). Studiengangspezifische Effekte lagen für diese drei Items ebenfalls vor (Itemnr: 8, 19 und 22). Die hohe Deckung zwischen standort- und studiengangspezifischem DIF hing wahrscheinlich damit zusammen, dass Studiengang und Standort in der vorliegenden Stichprobe nicht unabhängig voneinander waren. Zur Veranschaulichung des Geschlechtereffekts werden in Abbildung 18 die Item charakteristischen Kurven (ICCs) für das Item Nummer 19 nach Geschlecht dargestellt. Das Item war dem Bereich Marketing zugeordnet und erforderte zur Lösung die Wahl der richtigen Marktforschungsmethode. Weibliche Testteilnehmer (obere Line) hatten, trotz ansonsten gleicher Fähigkeit, eine höhere Wahrscheinlichkeit das Item zu lösen als männliche Teilnehmer (untere Linie). Der Unterschied in der Lösungswahrscheinlichkeit wird durch die Verschiebung der ICCs auf der y-Achse deutlich (vgl. Abbildung 18).

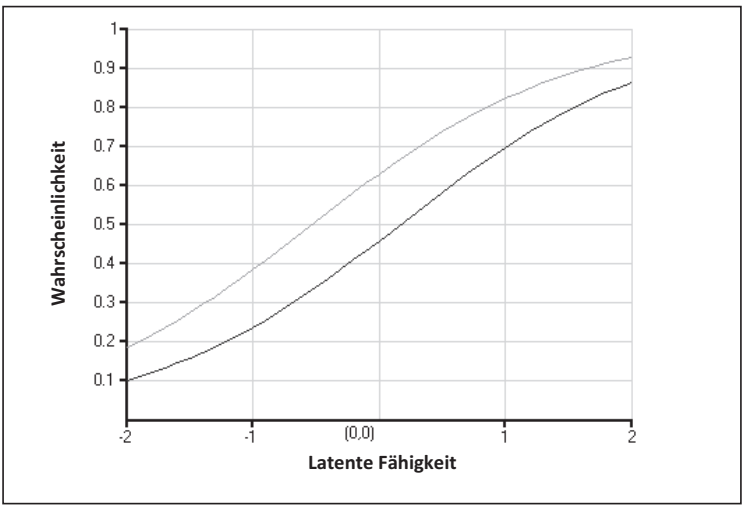


Abb. 18: Geschlechtsspezifische Item Characteristic Curves für das Item Nummer 19 (weibliche Testteilnehmer werden durch die obere Line, männliche durch die untere Linie repräsentiert)

Nach Entfernung der Items 7, 8, 13, 19 und 22 aus der Skalierung reduzierte sich das globale DIF bezogen auf den Standort (LR-Test: $\chi^2(14, N = 351) = 13.745$, $p = .47$) und bezogen auf den Studiengang (LR-Test: $\chi^2(14, N = 290) = 11.927$, $p = .66$) auf ein unbedenkliches Niveau. Auch das geschlechtsspezifische DIF reduzierte sich auf ein nicht-signifikantes Level (LR-Test: $\chi^2(14, N = 351) = 23.706$, $p = .05$), blieb aber auf Grund der nur knappen Überschreitung des Kriteriums $p < .05$ unter Aspekten der Fairness bedenklich. Die Analyse der Parameter, die auf Grundlage der fairnessorientierten Itemselektion geschätzt wurden, ergab, dass weibliche Testteilnehmer weiterhin schlechtere Testleistungen erzielen, der Effekt aber das Signifikanzniveau knapp überschreitet (stand. $\beta = .11$, $t(266) = 1.96$, $p = .051$). Die komplette Berechnung ist in Anhang F, Tabelle F-2 dargestellt. Die Ausprägung der Personen der Personenparameter über die Studiengänge hinweg blieb weitgehend unverändert (vgl. Abbildung 12 mit Anhang F, Abbildung F-2). Das heißt, der Ausschluss der als problematisch identifizierten Items führt nicht zu grundlegenden Veränderungen der in den vorangegangen Abschnitten beschriebenen Ergebnisse. Eine systematische Begründung für die DIF-Effekte konnte unter Post-hoc-Betrachtung der Items nicht abgeleitet werden. Eine kritische Diskussion und Interpretation der Ergebnisse erfolgt im letzten Teil der Arbeit (Kapitel 10).

9.3.4.2 Einschätzung der situativen Items durch Studierende

Wie in Abschnitt 7.2 beschrieben, gilt es, neben den Hauptgüterkriterien der pädagogisch-psychologischen Diagnostik auch Nebengütekriterien zu betrachten. Eines dieser Nebengütekriterien ist die Testakzeptanz. Testakzeptanz wird häufig mit Testmotivation in Verbindung gebracht (z. B. Chen & Hoshower, 2003). Deshalb ist es im Rahmen der Testentwicklung sinnvoll zu überprüfen, wie der Test von Testteilnehmenden wahrgenommen wird. In der Literatur zur Akzeptanz von Personalauswahlverfahren haben sich als Facetten der Testakzeptanz die Messqualität, die Augenscheinvalidität in Bezug auf spätere berufliche Tätigkeiten, die Kontrollierbarkeit und die Belastungsfreiheit etabliert (Kersting, 2008). Zur Abfrage dieser Facetten im Rahmen von Leistungstests wurde der Fragebogen AKZEPT!-L von Kersting (2008) entwickelt. Der Fragebogen umfasst pro Dimension vier Items, die Antwortmöglichkeiten auf einer Skala von 1–6 anbieten, wobei 1 für eine geringe und 6 für eine hohe Ausprägung auf der Dimension spricht. Um neben der Augenscheinvalidität in Bezug auf die spätere Tätigkeit auch die wahrgenommene curriculare Validität zu erfassen, wurden darüber hinaus im Rahmen dieser Arbeit sechs weitere Items entwickelt, die als fünfte Dimension die curriculare Validität des Tests aus Sicht der Studierenden erfassen sollten. Die Stichprobe für die Einschätzung der

Tests bestand insgesamt aus 70 Studierenden. In der Stichprobe waren nur Studierende der Betriebswirtschaftslehre und weiterer Wirtschaftswissenschaften vertreten. Studierende der Wirtschaftspädagogik wurden nicht berücksichtigt, da diese durch die berufliche Perspektive im Lehramt an berufsbildenden Schulen den Test bezüglich der beruflichen Augenscheinvalidität möglicherweise vor einem anderen beruflichen Erwartungsrahmen bewertet hätten. Jeweils 34 Studierende absolvierten und bewerteten entweder den situativen Test oder den deklarativen Test (Bothe, 2003). In jeder Gruppe füllte jeweils eine Person den Bewertungsfragebogen zur Testakzeptanz nicht aus und wurde aus den weiteren Berechnungen ausgeschlossen. Der Aufbau der Teilstudie zur Testvalidierung ist in Tabelle 14 dargestellt. Aus Mangel an theoretischen und empirischen Vorarbeiten wurden im Vorfeld der Untersuchung keine statistischen Hypothesen aufgestellt. Alle, im folgenden Abschnitt vorgestellten Signifikanztests weisen deshalb lediglich explorativen Charakter auf, was bei der Interpretation der Ergebnisse und der Einschätzung deren Reichweite berücksichtigt werden sollte.

Hinter dem Vergleich situativer und nicht-situativer Aufgaben stand jedoch die Vermutung, dass situative Aufgaben durch ihren anwendungsorientierten Charakter bei Studierenden auf größere Akzeptanz treffen als nicht-situative Aufgaben.

Um die Höhe der Ausprägung der Facetten der Testakzeptanz vergleichbar zu machen, wurden die Ergebnisse mit der Bewertung der nicht-situativen Items kontrastiert. Zur Prüfung des Unterschiedes zwischen den beiden Testformaten pro Akzeptanzfacette wurde eine multivariate ANOVA gerechnet. Tendenziell erwiesen sich die nicht-situativen Items gegenüber den situativen Items in vier der fünf ausgewählten Akzeptanzfacetten als überlegen ($F(5,68) = 4.039$, $p = .003$, $\eta_p^2 = .246$). Zwei dieser Unterschiede waren in der multivariaten ANOVA signifikant: Die Items des deklarativen Tests wurden von den Testteilnehmern als signifikant kontrollierbarer erlebt als die situativen Items ($F(1,34) = 12.001$, $p = .001$, $\eta_p^2 = .154$) und Studierenden empfanden die nicht-situativen Items als weniger belastend als die situativen Items, ($F(1,34) = 4.285$, $p = .042$, $\eta_p^2 = .61$). Darüber hinaus schätzten die Studierenden die Messqualität der nicht-situativen Items tendenziell höher ein als die der situativen Items. Zudem wurde den nicht-situativen Items eine geringfügig höhere curriculare Validität zugesprochen als den situativen Items. Einzig die Bewertung der beruflichen Augenscheinvalidität des Tests fiel für die situativen Items geringfügig positiver aus als für die nicht-situativen Items. Für beide Tests wurde die Relevanz für spätere berufliche Tätigkeiten jedoch vergleichsweise gering bewertet. Die curriculare Validität der Tests erhielt insgesamt die höchsten Bewertungen. Die große Abweichung zwischen der Bewertung der beruf-

lichen und der curricularen Validität ist möglicherweise ein Hinweis darauf, dass Studierende wenige Parallelen zwischen den universitären Anforderungen und den Anforderungen des Arbeitsmarktes sehen. Eine vergleichende Abbildung der Mittelwerte und 95 %-Konfidenzintervalle ist in Abbildung 19 dargestellt.

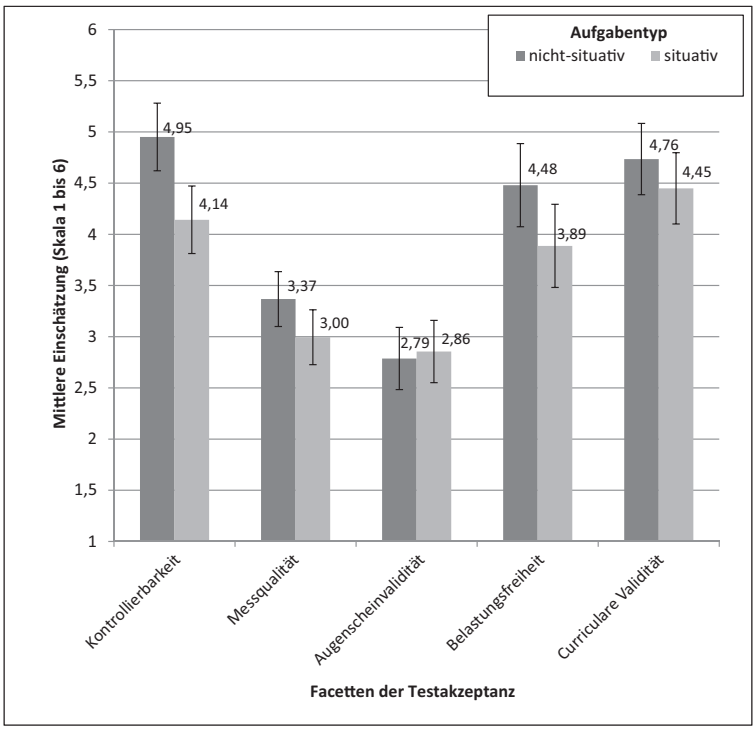


Abb. 19: Vergleichende Darstellung der Einschätzung situativer und nicht-situativer Items hinsichtlich unterschiedlicher Facetten der Testakzeptanz (Kersting, 2008) sowie der curricularen Validität, Mittelwerte und 95 %-Konfidenzintervall angegeben (n = 34 Studierende pro Test)

Da der situative Test für die Studierenden schwieriger war als der deklarative Test und dies möglicherweise die Bewertung von Seiten der Studierenden beeinflusste, wurde in einer zweiten Berechnung zum Vergleich der beiden Testtypen bezüglich der Testakzeptanz die Lösungshäufigkeit kontrolliert ($F(5,68) = 5.630$, $p < .001$, $\eta_p^2 = .319$). Nach Kontrolle der Lösungshäufigkeit war der Unterschied zwischen den beiden Testtypen auf der Facette der Belastungsfreiheit nichtmehr signifikant ($F(1,34) = 0.004$, $p = .951$, $\eta_p^2 = .00$). Alle weiteren Unterschiede blieben vergleichbar. Die Ergebnisse der multivariaten

ANOVA unter Kontrolle der Lösungshäufigkeit sind in Anhang F, Tabelle F-4 aufgeführt.

Die Ergebnisse bezüglich der Einschätzung der beruflichen Augenscheinvalidität warfen die Frage auf, inwiefern Studierende ohne berufspraktische Erfahrungen in betriebswirtschaftlichen Tätigkeitsfeldern in der Lage sind, die berufliche Augenscheinvalidität situativer Items angemessen zu beurteilen. Vergleiche der Einschätzungen der Studierenden mit betriebswirtschaftlicher Praxiserfahrung (erlangt durch kaufmännische Berufsausbildung, Praktikum oder Nebentätigkeit) ergaben, dass Studierende mit berufspraktischer Erfahrung den situativen Aufgaben eine etwas höhere Augenscheinvalidität zusprachen ($M = 3.15$; $SD = 1.03$) als Studierende ohne Praxiserfahrung ($M = 2.70$; $SD = 0.66$). Der Unterschied war jedoch statistisch nicht signifikant ($t(32) = -1.426$, $p = .17$). Die grafische Darstellung der Einschätzung der Augenscheinvalidität unter Berücksichtigung der betriebswirtschaftlichen Praxiserfahrung ist in Anhang F, Abbildung F-3 dokumentiert. Wie die Einschätzung der beruflichen Augenscheinvalidität von dem Erfahrungshintergrund der Testteilnehmer abhängt, sollte zukünftig mit größeren Stichproben und mit differenzierteren Abstufungen der Praxiserfahrung betrachtet werden.

Die Ergebnisse entsprachen nicht der Vermutung, dass situative Aufgaben attraktiver für studentische Testteilnehmer sind als nicht-situative Aufgaben. In Einklang mit den in Abbildung 19 dargestellten Befunden zur Testakzeptanz war die nach dem Test erhobene selbstberichtete Anstrengungsbereitschaft für nicht-situative Items signifikant höher ($M = 3.19$, $SD = 0.67$) als für situative Items ($M = 2.72$, $SD = 0.65$), ($t(65) = 2.975$, $p = .004$). Eine grafische Darstellung der Ergebnisse liegt in Abbildung 20 vor.

Die Dokumentation der verwendeten Skala „Anstrengungsbereitschaft" adaptiert nach Kunter (2002) liegt in Anhang F, Tabelle F-6 vor. Bei der Interpretation der Einschätzung von Seiten der Studierenden ist zu beachten, dass der Test mit den 30 situativen Items mehr Aufgaben enthielt als der Test mit den 23 nicht-situativen Items und somit mehr Bearbeitungszeit in Anspruch genommen hat, womit zum Beispiel das erhöhte Belastungsempfinden erklärbar wäre.

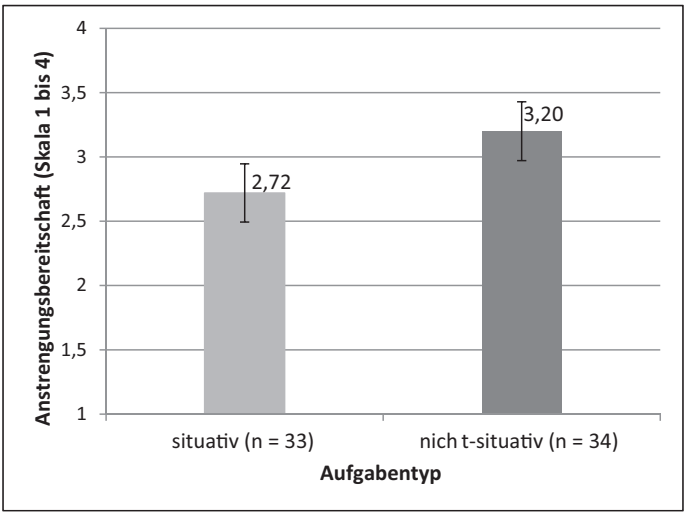


Abb. 20: Vergleichende Darstellung der selbstberichteten Anstrengungsbereitschaft (Kunter, 2002) nach dem Test für situative und nicht-situative Aufgaben, Mittelwerte und 95 %-Konfidenzintervall sind angegeben

Im Rahmen der Darstellung der Ergebnisse wurde auf eine Bewertung und Interpretation der Ergebnisse vorerst weitgehend verzichtet. Eine Zusammenfassung und kritische Diskussion der Ergebnisse wird im folgenden Kapitel 10 vorgenommen.

10 Diskussion

Nachdem die Ergebnisse der Hauptuntersuchung in Kapitel 9 weitgehend ohne Interpretation dargestellt wurden, erfolgen im ersten Abschnitt dieses Kapitels eine Zusammenfassung sowie die weiterführende Interpretation der Ergebnisse. Es folgt eine Herausarbeitung theoretischer und methodischer Limitationen der vorliegenden Arbeit und der daraus resultierenden Forschungsbedarfe. Der Abschnitt 10.2 schließt mit einer kritischen Diskussion der Funktion betriebswirtschaftlicher Wissenstests an Universitäten. Im letzten Teil der Arbeit werden ein Fazit sowie ein Ausblick auf zukünftige Forschung im Bereich der Messung betriebswirtschaftlichen Wissens von Studierenden formuliert (Abschnitt 10.3).

10.1 Zusammenfassung und Interpretation der empirischen Ergebnisse

Zusammenfassend bestätigten die Ergebnisse in weiten Teilen eine erfolgreiche Testentwicklung. Die situativen Items zeigten nach geringfügiger Itemselektion keine globalen und lokalen Verletzungen des Rasch-Modells. Die Verteilung der Schwierigkeiten über das Schwierigkeits-/Fähigkeitskontinuum sowie das Schwierigkeitsniveau der Aufgaben war angemessen. Lediglich eine Ergänzung sehr leichter Items wäre wünschenswert. Die Reliabilität der 20 selektierten Items war mit .60 nur bedingt zufriedenstellend. Für eine um sechs Items verlängerte Version des Tests, die an einer Substichprobe von 35 Personen eingesetzt wurde, ergab sich ein Cronbachs Alpha von .71, was als zufriedenstellend bewertet wurde. Neben der Verlängerung des Tests durch sechs Items wurden auch die Testbedingungen verbessert, indem das Testheft nur die situativen Items und eine verkürzte Version des Begleitfragebogens enthielt, was zu mehr Zeit für die Bearbeitung der situativen Items führte. Zudem wurde die Möglichkeit des voneinander Abschreibens dadurch reduziert, dass zwei Testhefte im Wechsel ausgegeben wurden. Das Testheft A enthielt 23 ausgewählte Items aus dem BAKT (Bothe, 2003) sowie den verkürzten Begleitfragebogen, und das Testheft B enthielt die 30 situativen Items mit verkürztem Begleitfragebogen.

Wie vorhergesagt, ließen sich die situativen Items am besten über ein eindimensionales Modell beschreiben. Dieses Ergebnis wies darauf hin, dass betriebswirtschaftliches Wissen auf Bachelorniveau ein breites, verschiedene betriebswirtschaftliche Inhaltsgebiete umfassendes Konstrukt darstellt und durch einen einzigen Testwert repräsentiert werden kann. Strukturelle Analysen zum

Verhältnis von situativen Items zu nicht-situativen Items wiesen für eine getrennte Modellierung der beiden Testtypen (deklarativ und situativ) die beste Modellpassung auf. Die Bestätigung des zweidimensionalen Modells kann dahingehend interpretiert werden, dass es gelungen ist, einen Test zu entwickeln, der im Vergleich zum BAKT (Bothe, 2003) ein distinktes Wissenskonstrukt misst. Der latente Zusammenhang zwischen diesen beiden Dimensionen war mit .83 sehr ausgeprägt, sodass davon auszugehen ist, dass beide Tests trotzdem ein inhaltlich sehr ähnliches Konstrukt messen. Entsprechend knapp fiel der Vergleich zwischen den alternativen Modellen aus, sodass eine Trennung der beiden Testtypen zwar als Arbeitsgrundlage für weitere Berechnungen in dieser Arbeit herangezogen wurde, es jedoch durchaus möglich ist, dass bisher nicht getestete dimensionale Modelle eine noch bessere Passung aufweisen. Um ein Testinstrument mit guter Reliabilität zu erreichen, ist es durchaus denkbar, unter Selektion modellverletzender Items alle Items gemeinsam zu skalieren.

Neben den konfirmatorischen faktoranalytischen Untersuchungen stärkten Subgruppenvergleiche die Befunde zur Konstruktvalidität des Tests. Die situativen Items wurden gezielt so entwickelt, dass sie sowohl für Studierende der Betriebswirtschaftslehre als auch für Studierende der Wirtschaftspädagogik ein geeignetes Maß betriebswirtschaftlichen Wissens darstellen sollten. Wie von einem solchen Test zu erwarten war, schnitten Studierende der Betriebswirtschaftslehre und Studierende der Wirtschaftspädagogik in dem Test etwa gleich gut ab. Studierende der Volkswirtschaft und anderer Studiengänge wiesen signifikant schlechtere Testleistungen auf. Zwischen Studierenden im Bachelorstudium und Studierenden im Masterstudium lag kein signifikanter Unterschied bezüglich der Testleistung vor. Wurden die situativen Items im Testheft nach den Items des BAKTs (Bothe, 2003) platziert, so war über alle Testteilnehmer hinweg ein Trend zu geringeren Leistungen erkennbar. Dieser Trend war jedoch nicht signifikant und wurde im Testdesign der Haupterhebung über eine randomisierte Darbietung beider Testblöcke ausgeglichen. Die Betrachtung der schwierigkeitsbestimmenden Aufgabenmerkmale zeigte den erwartungskonformen Trend dahingehend, dass Items mit komplexen Aufgabenstämmen schwieriger waren als Items mit weniger komplexen Aufgabenstämmen. Ebenso waren Items, die mathematische Modellierungsleistungen erforderten, schwieriger als Items, die ohne mathematische Modellierungsleistung gelöst werden konnten. Das bei der Testentwicklung anvisierte kognitive Anforderungsniveau im Sinne der vereinfachten Taxonomie von Anderson und Krathwohl (2001) (vgl. Abschnitt 2.3.2) spiegelte sich nicht in der erwarteten hierarchischen Schwierigkeitsrangfolge von „Reproduzieren“ über „Anwenden“ bis „Reflektieren“ wider. Vielmehr schienen Aufgaben, die entwickelt wurden, um das kognitive Anforderungsniveau „Anwenden“ abzudecken,

tendenziell leichter zu sein als Aufgaben, die reproduzierendes oder reflektierendes Wissen bei der Aufgabenlösung erforderten. Die Ergebnisse zur Analyse der schwierigkeitsbestimmenden Aufgabenmerkmale sollten aufgrund der geringen Itemzahlen, auf denen sie basieren, nicht generalisiert werden. Trotzdem sind die nicht-erwartungskonformen Ergebnisse vor dem Hintergrund gemischter Ergebnisse bezüglich der Modellierung des kognitiven Anforderungsniveaus in vorangegangenen Studien zu kaufmännischen Kompetenzen (Lehmann & Seeber, 2007; Seeber, 2008; Witt, 2006) nicht überraschend und sie unterstreichen die Notwendigkeit, die Forschung in diesem Bereich weiterzuführen und auszubauen.

Eine Forschungsfrage der vorliegenden Arbeit war, ob situative Aufgaben für Studierende schwerer oder leichter sind als nicht-situative Aufgaben und ob ein solcher Schwierigkeitseffekt von der Vorerfahrung mit situativen beruflichen Anforderungen abhängt. Auf der Grundlage des Vergleichs der Lösungsquoten pro Item zwischen den beiden Testtypen ließ sich feststellen, dass situative Aufgaben tendenziell eine geringere Lösungsquote aufwiesen als nicht-situative Aufgaben. Das heißt, sie waren im Mittel etwas schwerer. Dieser Unterschied war jedoch nicht signifikant, was möglicherweise auf geringe Itemzahlen zurückzuführen ist. Was die Testleistung betrifft, zeigte sich ein deutlicher Effekt zu Gunsten der nicht-situativen Aufgaben. Die Items des BAKT (Bothe, 2003) wurden signifikant häufiger richtig gelöst als die Items des situativen Tests.

An dieser Stelle muss einschränkend vermerkt werden, dass dieser Unterschied nicht unbedingt durch den Aufgabentyp zustande gekommen sein muss, sondern der situative Test möglicherweise unabhängig von seinem Aufgabenformat schwieriger ist, weil zum Beispiel schwerere betriebswirtschaftliche Inhalte abgefragt wurden. Zwar wurde bei der Testentwicklung auf eine vergleichbare Abdeckung der Iteminhalte geachtet, diese sollte jedoch von erfahrenen Lehrkräften an der Universität erneut überprüft und beurteilt werden.

Ein zentraler Befund der Arbeit war, dass Personen mit kaufmännischer Berufsausbildung bei situativen Aufgaben einen größeren Leistungsvorsprung vor Personen ohne kaufmännische Ausbildung aufwiesen als bei nicht-situativen Aufgaben. Es ist gelungen, mittels situativer Aufgaben ein Wissenskonstrukt zu erfassen, dessen Erfassung durch das Absolvieren einer kaufmännischen Ausbildung besonders begünstigt wurde. Der Vorteil von Personen mit kaufmännischer Ausbildung kann zum einen durch deren Berufserfahrung bedingt sein, zum anderen aber auch durch die gezielte Vorbereitung auf situative Testaufgaben im Rahmen der Abschlussprüfungen der Industrie- und Handelskammern. Wahrscheinlich ist, dass beide Aspekte bei der Bearbeitung von situativen Aufgaben eine Rolle spielen. Eine wichtige Erkenntnis aus den Ergebnissen

ist, dass verschiedene Aufgabentypen Personen mit unterschiedlichen Merkmalen leichter oder schwerer fallen (vgl. dazu auch Winther, 2011). Die vorliegende Arbeit hat gezeigt, dass für eine differenzierte Abbildung betriebswirtschaftlichen Wissens auf Bachelorniveau unterschiedliche Aufgabentypen eingesetzt werden sollten. Darüber hinaus untermauern die beschriebenen Befunde die Konstruktvalidität des Tests.

Inwiefern es tatsächlich gelungen ist, prozedurale Wissensstrukturen zu erfassen, kann auf Grundlage dieser Berechnungen jedoch nicht abschließen geklärt werden und sollte zukünftig durch Fachexperten sowie Experten der Fachdidaktik beurteilt werden. Anzumerken ist, dass die situativen Aufgaben des Tests kaum direkt prozedurales Wissen im Sinne der realen Durchführung von Routinen und Prozeduren erfassen, sondern den Schwerpunkt auf der Erfassung des Wissens über die Anwendung von Prozeduren und deren situationsabhängige Angemessenheit haben. Im Sinne von Wittmann, Süß und Oberauer (1996) wäre dieses Wissen als deklarativ gespeichertes Handlungswissen zu bezeichnen. Für die Erfassung prozeduralen Handlungswissens (Wittmann et al., 1996) wären andere Testarrangements, wie zum Beispiel computerbasierte Simulationen, notwendig (vgl. ALUMSIM Winther, 2010). In der Regel zeigen sich zwischen computerbasierten Simulationen, die prozedurales Wissen erfassen sollen, und papierbasierten deklarativen Aufgaben Dimensionen, die wahrscheinlich aufgrund der methodischen Umsetzung deutlich ausgeprägter sind als diejenigen zwischen dem BAKT (2003) und den situativen Aufgaben. Auf den Zusammenhang zwischen Testarrangement und erfasster Wissensart wird in den Abschnitten 10.2 und 10.3 erneut eingegangen.

Die aus der Literatur bekannten Kriterien zur Validierung wirtschaftswissenschaftlicher Wissenstests (kaufmännische Ausbildung, Note der Hochschulzugangsberechtigung, Geschlecht, Anzahl- und Note der besuchten Grundlagenmodule) (vgl. Kapitel 3) erwiesen sich in einer multiplen OLS-Regression als signifikant in vorhersagekonformer Richtung. Für alle Hypothesen zu diesen klassischen Validierungskriterien wurde die H0 verworfen (H2–1 bis H2–5). Von den als Prädiktor ausgewählten individuellen studienbezogenen Faktoren, die potenziell einer Intervention von Seiten der Lehrenden an Hochschulen zugänglich wären (fachliches Interesse, Lernstrategien, akademisches Selbstkonzept und Leistungsmotivation), zeigte sich nur das Interesse an betriebswirtschaftlichen Fachinhalten als bedeutsamer Prädiktor für die Testleistung. Das akademische Selbstkonzept wies zwar einen signifikanten Zusammenhang mit der Testleistung auf, war aber sowohl mit den Noten der Hochschulzugangsberechtigung als auch mit den Noten der Modulabschlussprüfungen so hoch korreliert, dass es nicht in die Gleichung zur Vorhersage der Testleistung aufgenommen wurde. Insgesamt kann die Kriteriumsvalidierung anhand „klassi-

scher" Kriterien als gelungen eingestuft werden. Mit dem fachspezifischen Studieninteresse wurde zusätzlich eine Variable mit der Testleistung in Verbindung gebracht, die potenziell durch Interventionen von Seiten der Hochschule zum Positiven veränderbar ist. Der Versuch, durch veränderbare personenbezogene Konstrukte wie Leistungsmotivation und Lernstrategien weitere Erklärungsfaktoren für die Testleistung zu identifizieren, ist nicht gelungen. Für die Hypothesen H2–6 und H2–7 wurde die Nullhypothese beibehalten; für die Hypothese H2–8 wurde die Nullhypothese mit Einschränkungen verworfen.

Dass der ausgewählte Prädiktor „Verwendung von Lernstrategien" keinen Zusammenhang mit der Testleistung aufwies, ist gegebenenfalls durch dessen Operationalisierung zu begründen. Möglicherweise waren die Testitems nicht ausreichend an den Kontext von Lernsituationen im betriebswirtschaftlichen Studium angepasst. Zudem steht in Frage, ob über retrospektive Selbstberichte eine valide Erfassung von Lernstrategien möglich ist. Studien haben gezeigt, dass es Personen häufig nicht gelingt, ihr eigenes Lernverhalten rückblickend korrekt zu beurteilen (Artelt, 2000). Zukünftig sollten weitere Aspekte des Lernens im Zusammenhang mit Testleistungen untersucht werden. Vielversprechend scheint zum Beispiel die Erfassung von studienbezogenen Metakognitionen. Metakognitionen beschreiben das Wissen und die Kontrolle über eigene kognitive Funktionen einer Person und weisen bedeutsame Zusammenhänge mit Lernleistungen auf (Hasselhorn & Labuhn, 2008).

Der fehlende Zusammenhang zwischen allgemeiner Leistungsmotivation und Testleistung ist in Anbetracht der relativ belastbaren Forschungsbefunde, die diesen Zusammenhang nahelegen (Brunstein & Heckhausen, 2010), nicht nachvollziehbar. Möglicherweise wäre eine fachspezifische betriebswirtschaftsspezifische Operationalisierung des Konstrukts, wie beim Interessenstest FSI (Schiefele et al., 1993), sinnvoll gewesen. Die Varianzaufklärung des Regressionsmodells von $R^2 = .19$ wies darauf hin, dass es neben den bekannten Einflussfaktoren auf die Testleistung noch weitere Faktoren gibt, die identifiziert werden sollten, um wissenschaftliche Erkenntnisse zur Genese betriebswirtschaftlicher Testleistung auszubauen. Zu den Ergebnissen der Regressionsanalyse ist kritisch hinzuzufügen, dass ein signifikanter Zusammenhang zwischen einem Prädiktor (z. B. Fachinteresse) und dem Kriterium (Testleistung) nicht zwingend einen Kausalzusammenhang von Fachinteresse auf Testleistung belegt. Theoretisch ist es auch möglich, dass Studierende, die in dem Test gut abgeschnitten haben, a posteriori angeben, mehr Interesse an betriebswirtschaftlichen Inhalten zu haben als Studierende, die in dem Test weniger gut abschnitten. Um den Raum für diese umgekehrt kausalen Interpretationen der Beziehung zwischen Prädiktor und Kriterium möglichst gering zu halten, wurden in der vorliegenden Arbeit nur Prädiktoren aufgenommen, deren kausale

Wirkung auf die Testleistung durch vorangegangene theoretische und empirische Arbeiten nahegelegt wurde (vgl. Abschnitt 9.1). Eine elaboriertere Methode der Bestätigung von Kausalität zwischen Prädiktor und Kriterium wäre ein Erhebungsdesign mit mehreren Messzeitpunkten, das bei zukünftigen Forschungsvorhaben umgesetzt werden sollte.

Als ein Indikator der Testfairness identifizierte die Analyse von itemspezifischem Differential Item Functioning (DIF) einige Items speziell bezogen auf das Geschlecht als problematisch. Deutlich weniger kritisch waren die Ergebnisse bezüglich der standort- und studiengangspezifischen Fairness. Der Test zeigte kaum Unterschiede hinsichtlich der Messqualität zwischen den Standorten Göttingen, Hamburg und Mannheim. Das heißt, der Test kann, wie in der Testentwicklung beabsichtigt, bei Studierenden der Wirtschaftspädagogik und Studierenden der Betriebswirtschaftslehre gleichermaßen eingesetzt werden. Die Verletzung der itembezogenen Messinvarianz zwischen den Geschlechtern war deutlich ausgeprägter als zwischen den Standorten oder Studiengängen. Nach Ausschluss der kritischen Items konnte das globale DIF auf ein nicht-signifikantes Niveau gesenkt werden. Die Tendenz, dass weibliche Testteilnehmer schlechtere Testleistungen erzielten als männliche, blieb aber auch nach Ausschluss der kritischen Items bestehen (vgl. Anhang F, Tabelle F-3). Auch weitere Beziehungen zu Außenkriterien blieben unverändert, was darauf hinweist, dass die Ergebnisse der Validierungsstudie nicht maßgeblich durch die als kritisch identifizierten Items verzerrt wurden.

Bis dato ist wenig darüber bekannt, warum bestimmte Aufgaben für Frauen schwerer sind als für Männer und vice versa. Da auch in der vorliegenden Arbeit keine systematischen Ursachen für die DIF-Effekte erkennbar waren, wurde DIF als ungewollter Störeinfluss behandelt. Differential Item Functioning kann jedoch auch ein Hinweis auf Mehrdimensionalität sein. Anstelle des Ausschlusses oder der Überarbeitung der Items, bieten neue Forschungsstrategien die Möglichkeit, auf der Grundlage von DIF zu ermitteln, welche Merkmalskombinationen in der Personengruppe dazu führen, dass ein Item für diese Gruppe, trotz ansonsten gleicher Fähigkeit, besonders leicht oder besonders schwer ist (Strobl, Kopf & Zeileis, 2011). Weiterführende Berechnungen zur Erklärung der DIF-Effekte wurden in der vorliegenden Arbeit nicht vorgenommen, hier besteht jedoch ein Forschungsbedarf, der in Abschnitt 10.2 näher erläutert wird.

Die Einschätzung der situativen Aufgaben durch 34 Studierende auf fünf ausgewählten Facetten der Testakzeptanz fiel im Vergleich zu den deklarativen Aufgaben schlechter aus. Lediglich die Dimension „Augenscheinvalidität", die sich auf die wahrgenommene berufliche Relevanz der Aufgaben bezog, fiel für situative Aufgaben erwartungskonform etwas höher aus als für deklarative

Aufgaben. Der Unterschied war jedoch so gering, dass kaum davon auszugehen ist, dass der Test durch seine intendierte Nähe zu beruflichen Tätigkeiten besonders motivierend ist. Diese Vermutung wurde durch den Befund bestätigt, dass Studierende, die die nicht situativen Items bearbeitet hatten, bezogen auf den Test eine signifikant höhere Anstrengungsbereitschaft aufwiesen als Studierende, die den situativen Test bearbeitet hatten. Insgesamt wurde die berufliche Augenscheinvalidität für beide Tests als sehr gering eingestuft. Das heißt, es ist nicht gelungen, einen situativen Test zu entwickeln, der aus Sicht der Studierenden spätere berufliche Anforderungen abdeckt. Dies war jedoch nicht das vornehmliche Ziel der Testentwicklung. Vielmehr sollte der Test in erster Linie curriculare und damit universitäre Anforderungen abbilden, was sowohl aus Sicht der Studierenden gelungen ist als auch im Rahmen der kriterialen Validierung bestätigt wurde. Trotzdem ist die geringe berufliche Augenscheinvalidität im Hinblick auf die prognostische Validität des Tests in Bezug auf spätere berufliche Tätigkeiten kritisch zu hinterfragen. Unter Berücksichtigung der berufspraktischen Erfahrungen der Studierenden änderte sich die Bewertung der situativen Testaufgaben im Vergleich zu den deklarativen Testaufgaben in erwartungskonforme Richtung (vgl. Anhang F, Abbildung F-3). Der Effekt war jedoch so gering und nicht signifikant, dass keine verallgemeinerbaren Aussagen über den Zusammenhang zwischen Berufserfahrung und Bewertung der Aufgaben getroffen werden können. Der sich abzeichnende Trend, dass die Augenscheinvalidität eines Tests vor dem Hintergrund eigener beruflicher Vorerfahrungen bewertet wird, sollte jedoch in zukünftigen Forschungsarbeiten berücksichtigt werden. Um unabhängig von der Einschätzung der Studierenden zu prüfen, ob der Test den späteren beruflichen Erfolg vorhersagen kann, bedürfte es jedoch einer Untersuchung mit mindestens einer Follow-up-Erhebung, wie sie in Abschnitt 10.2 diskutiert wird.

Ein positives Ergebnis der Studie zur Testakzeptanz war die gute Bewertung der curricularen Validität beider Tests. Es scheint durch das in Kapitel 6 beschriebene Vorgehen bei der Testentwicklung gelungen zu sein, die Inhalte des Studiums in den Items zu verankern. Die Ausrichtung von Curriculum und Assessments ist eine wichtige Forderung der bildungswissenschaftlichen Erfassung von Lernergebnissen (Achtenhagen, 2012; Wilson, 2005) und wird in Abschnitt 10.2 mit Bezug auf die vorliegende Arbeit kritisch diskutiert.

Die Ergebnisinterpretation der Erhebung zur Einschätzung der beiden Tests durch Studierende unterliegt der Einschränkung, dass die verlängerte Version des situativen Tests mit 30 Items sieben Items länger war als der deklarative Test. Es bleibt zu prüfen, ob die Einschätzung von zwei exakt gleich langen Versionen zu vergleichbaren Ergebnissen führen würde. Zu vermuten ist, dass die verhältnismäßig langen Aufgabenbeschreibungen zu erhöhtem Belastungs-

empfinden und weniger Kontrollerleben bei den Studierenden führten. In Abschnitt 10.2 wird dieser Gedanke aufgegriffen und hinsichtlich multimedialer Umsetzung situativer Items diskutiert.

Durch die quantitativ-empirische Untersuchung situativer Aufgaben ist es gelungen, ein differenziertes Eigenschaftsprofil situativer Aufgaben im Anwendungskontext Hochschule zu erstellen. Insbesondere der Interaktionseffekt zwischen Aufgabentyp und absolvierter kaufmännischer Ausbildung (vgl. Abbildung 17) untermauerte, dass durch situative Aufgaben Wissen erfasst wird, dass mit berufsspezifischen situativen Vorerfahrungen zusammenhängt. Assessments mit dem Ziel, ein möglichst umfassendes Bild betriebswirtschaftlichen Wissens von Studierenden zu zeichnen, sollten dementsprechend sowohl deklarative als auch situative Aufgaben miteinbeziehen. Welche Besonderheiten in zukünftigen Forschungsvorhaben berücksichtigt werden sollten, wird im folgenden Abschnitt beschrieben.

10.2 Limitation der empirischen Ergebnisse und Forschungsbedarfe

Wie jede Forschungsarbeit, weist auch die vorliegende Grenzen auf, die bei der Interpretation der Befunde zu berücksichtigen sind und die zugleich den Anstoß für zukünftige Forschungsarbeiten geben. Obwohl theoretische und methodische Schwachstellen in der Anlage einer Arbeit häufig miteinander verbunden sind, werden diese im Folgenden getrennt voneinander aufgeführt, kritisch betrachtet und weitere Forschungsdesiderata daraus abgeleitet. Im ersten Teil dieses Unterkapitels wird die theoretische Fundierung der vorliegenden Arbeit diskutiert. Im zweiten Teil werden methodische Grenzen der Arbeit dargestellt und Alternativen für zukünftige methodische Umsetzungen aufgezeigt sowie die Funktion von Assessments im Hochschulbereich kritisch betrachtet.

Kritische Betrachtung der theoretischen Einbettung

Das Forschungsfeld der quantitativen Messung betriebswirtschaftlichen Wissens an deutschen Hochschulen ist bisher wenig erschlossen (vgl. Kapitel 3). Damit sind die Forschungsergebnisse dieser Arbeit für den deutschen Sprachraum neuartig und weisen das Potenzial auf, als Grundlage für weitere Forschung in diesem Bereich zu dienen. Zu beachten ist dabei, dass hinsichtlich der theoretischen Basis dieser Arbeit aufgrund der bisher sehr geringen Forschungsaktivitäten in diesem Feld nicht auf umfassende Befunde zurückgegriffen werden konnte. Auch unter Einbezug der Erträge berufs- und wirtschaftspädagogischer Forschung ist die theoretische Fundierung der Testentwicklung

in der betriebswirtschaftlichen Domäne als noch nicht abgeschlossen zu bewerten.

Nach einem Abwägen unterschiedlicher Zugänge (vgl. Abschnitt 4.1) wurde die Domänenmodellierung in der vorliegenden Arbeit in erster Linie aus einer curricularen fachinhaltsbezogenen Perspektive vorgenommen. Die Auswahl der Universitäten, deren Studienordnungen und Modulbeschreibungen analysiert wurden, erfolgte ressourcenbedingt eklektisch und war auf drei Standorte beschränkt. Die Beschränkung auf drei Studienstandorte hat eine eingeschränkte Gültigkeit der Domänenstruktur zur Folge und lässt die Frage offen, ob der Test auch für den Einsatz an weiteren Universitäten in Deutschland geeignet wäre. Die Entwicklung eines standortübergreifenden Domänenmodells ist, zumindest unter der Bedingung eines deutschlandweiten Forschungsinteresses, ein dringendes Forschungsdesiderat, das möglicherweise im Rahmen des in Abschnitt 3.3 beschriebenen laufenden Forschungsprojektes WiWiKom Realisierung findet.

Ein zweiter Zugang, der im Rahmen der Testentwicklung gewählt wurde, war die Integration der Anforderungen beruflicher Tätigkeiten für Absolventen der Betriebswirtschaftslehre. Die Items wurden so gestaltet, dass sie Anforderungen widerspiegelten, die in, an junge Absolventen der Betriebswirtschaftslehre gerichteten, Stellenanzeigen gefordert wurden. Ein systematischer Vergleich zwischen den curricularen Anforderungen und Anforderungen an potenziellen Arbeitsplätzen von Absolventen der Betriebswirtschaftslehre wurde in der vorliegenden Arbeit nicht vorgenommen. Dieses mögliche Spannungsfeld wurde umgangen, indem in erster Linie die curricularen Anforderungen rahmengebend für die Testentwicklung waren und die Ergänzung der Anforderungen aus Stellenanzeigen erst in einem zweiten Schritt erfolgte. Das heißt, der Test deckte primär curriculare Inhalte ab und entsprach nur sekundär den in Stellenanzeigen formulierten Anforderungen, was durch die Einschätzung der Studierenden bestätigt wurde. Zum Beispiel wurden die Berufsfelder Personalwesen und Wirtschaftsinformatik, die unumstritten auf dem Arbeitsmarkt eine Einmündungsmöglichkeit für Absolventen der Betriebswirtschaftslehre bieten, aufgrund nicht eindeutiger curricularer Verortung nicht in den Test aufgenommen. Die Inhaltsbereiche des Tests decken somit nur ein Teilgebiet betriebswirtschaftlicher Handlungsbereiche ab. Dementsprechend sind die in der vorliegenden Arbeit erzielten Ergebnisse nur für ausgewählte Tätigkeitsfelder gültig.

Für zukünftige Forschungsvorhaben wäre es wichtig und für ein umfassendes Verständnis der Domäne notwendig zu untersuchen, inwiefern curriculare Anforderungen und Anforderungen an den Arbeitsplätzen von Absolventen der Betriebswirtschaftslehre deckungsgleich sind und in welchen Bereichen sie

voneinander abweichen. Neben den Einmündungsmöglichkeiten von Absolventen der Betriebswirtschaftslehre in sehr unterschiedliche Fachbereiche, stellen auch Tätigkeiten auf unterschiedlichen Hierarchieebenen eine große Herausforderung bei dieser Analyse beruflicher Anforderungen dar. Je nachdem auf welcher Hierarchiestufe eine Person in einem Unternehmen betriebswirtschaftlich tätig ist, unterscheiden sich die Anforderungen an professionelle Kompetenzen. Ein mögliches Resultat dieser Forschungsbemühung könnte die Feststellung sein, dass zwei getrennte Tests entwickelt werden müssen: ein Test, der die Anforderungen tatsächlicher beruflicher Tätigkeiten abdeckt, und ein Test, der vom Curriculum ausgeht. Einen ersten Hinweis, dass es sich zumindest aus Sicht der Studierenden um zwei unterschiedliche Konstrukte handelt, liefert die Analyse der Facetten der Testakzeptanz: Während sowohl der deklarative als auch der situative Test hohe Werte bezüglich ihrer curricularen Validität erzielen, wird die berufliche Augenscheinvalidität als gering eingestuft. Ob es jedoch sinnvoll wäre, Tests in starker Anlehnung an einzelne Handlungen in beruflichen Tätigkeitsfeldern zu entwickeln, ist vor dem Hintergrund der derzeitigen universitären Bildung, ohne verpflichtende Praxiserfahrung im unternehmerischen Bereich, fraglich. Ein solches Vorgehen würde einen Bruch zwischen dem Curriculum, der Instruktion und dem Assessment bedeutet (Wilson, 2005), was für die Funktion von Assessments im Bildungswesen kritisch ist, wie im letzten Teil dieses Abschnitts herausgearbeitet wird. Schlussendlich kann die curricular orientierte Umsetzung der situativen Items mit Anteilen aus Stellenanzeigen als guter Mittelweg betrachtet werden.

Ein weiterer Aspekt der Testentwicklung, der durch die theoretische Basis beeinflusst wurde, war die Generierung von Aufgaben auf unterschiedlichen kognitiven Anforderungsniveaus. In der vorliegenden Arbeit wurde versucht, drei potenziell schwierigkeitsbestimmende Aufgabencharakteristika (kognitive Prozessdimension, Komplexität und mathematisches Modellieren) systematisch über die Aufgaben hinweg zu variieren. Die Klassifikation nach kognitiven Prozessen stellte sich in der Entwicklungsphase als wenig trennscharf heraus, und die vermutete Schwierigkeitshierarchie wurde empirisch nicht bestätigt. Diese Problematik ist bereits aus anderen Forschungsarbeiten bekannt (Seeber, 2008; Witt, 2006). Sie unterstreicht den Forschungsbedarf dahingehend, dass zu klären ist, wie das kognitive Anforderungsniveau einer Aufgabe über die genannten Merkmale hinaus bestimmt wird und wie dieses die Schwierigkeit der Aufgabe beeinflusst. Zu diesem Zweck sollten die Items (1) von mehreren fachinhaltlichen und fachdidaktischen Experten erneut beurteilt und kategorisiert werden. (2) sollte unter Verwendung der Methode des lauten Denkens (Weidle, 1994) analysiert werden, welche Lösungsschritte und Strategien die Testteilnehmer beim Lösen der Aufgaben tatsächlich verwenden und inwiefern der Lösungsprozess durch a priori bestimmte Aufgabeneigenschaften beein-

flusst wird. Dabei sollte das Augenmerk darauf gelegt werden, wie der Bekanntheitsgrad mit einer Aufgabe oder einem Aufgabentyp den Lösungsprozess und den Lösungserfolg der Testteilnehmer bei einer Aufgabe beeinflusst.

In der vorliegenden Arbeit wurde sowohl bezüglich der Strukturierung der Domäne als auch bei der Modellierung der kognitiven Anforderungen ein modellgeleiteter Ansatz gewählt, der jedoch nicht dem Ideal einer vollständig modellgeleiteten Testentwicklung entspricht. Kritisch ist dabei zu vermerken, dass ein Bruch zwischen theoretischer Grundlage und Umsetzung des Testmaterials den Rückschluss von den empirischen Ergebnissen auf die Theorie schwächt. Die vorliegende Arbeit kann als ausbaufähige Grundlage theoriegeleiteter Testentwicklung im Bereich betriebswirtschaftlichen Hochschulwissens betrachtet werden. Insbesondere die zahlreichen empirischen Befunde, die die Validität des Tests untermauern, legen nahe, dass es trotz der oben genannten Verbesserungspotenziale gelungen ist, einen validen Test zu entwickeln. Zusätzlich zu den aufgezeigten Entwicklungsbedarfen in der theoretischen Fundierung und der Generalisierbarkeit des Tests werden im Folgenden methodische Limitationen der vorliegenden Arbeit betrachtet.

Kritische Betrachtung der methodischen Umsetzung

Ein methodischer Aspekt des Tests, der näherer Betrachtung bedarf, ist dessen relativ geringe Reliabilität. Obwohl der Test in seiner verlängerten Form (26 Items) eine Reliabilität von $\alpha = .71$ erreicht und somit die Anforderungen pädagogisch-psychologischer Diagnostik erfüllt (vgl. Kapitel 7), handelt es sich dabei nur um ein akzeptables Niveau. Dieses Niveau sollte verbessert werden und muss sich darüber hinaus an einer größeren Stichrobe und unter der Parameterschätzung durch das Rasch-Modell erneut etablieren. Obwohl Reliabilitäten zwischen .50 und .70 speziell bei situierten Messverfahren häufig publiziert werden (z. B. Abele et al., 2012; Schmitt & Chan, 2006; Winther, 2010), sollten trotzdem die Implikationen einer geringen Reliabilität betrachtet werden.

Eine geringe Reliabilität spricht dafür, dass die Messung von unbekannten Fehlerquellen überlagert wird. Diese unbekannten Fehlerquellen können sowohl innerhalb des Instruments liegen als auch durch äußere Faktoren der Testsituation bedingt werden. Eine mögliche Fehlerquelle, die vom Messinstrument ausgeht, ist nicht erkannte Mehrdimensionalität. Werden neben dem eigentlichen Zielkonstrukt noch weitere Konstrukte erfasst, die zwischen den Testteilnehmern variieren, so schlägt sich das in einer geringen Reliabilität nieder. Zwar wurde eine eindimensionale Struktur der Items im Rahmen eines Modellvergleichs als wahrscheinlich identifiziert, die geringen Unterschiede in der Passung zu alternativen Modellen (vgl. Tabelle 15) legen jedoch die Vermutung nahe, dass in den Daten eine unerkannte mehrdimensionale Struktur verbor-

gen sein könnte. Ohne begründete Vermutungen darüber, welcher Art diese mehrdimensionale Struktur sein könnte („within"- oder „between"-Item) und welche Items auf welchen Faktor laden, ist es jedoch keine sinnvolle Forschungsstrategie, willkürlich multidimensionale Modelle miteinander zu vergleichen. Aufgrund der unzureichend entwickelten theoretischen Grundlagen zur Modellierung der Mehrdimensionalität wurde in der vorliegenden Arbeit die Strategie verfolgt, den Test durch Itemselektion in Richtung Eindimensionalität zu optimieren. Für weitere Forschungsvorhaben ist es daher empfehlenswert, bereits bei der Testentwicklung theoretisch begründete mehrdimensionale Modelle in Betracht zu ziehen. Für situative Items weisen „within-item"-Modelle großes Potenzial auf heterogene Leistungsanforderungen innerhalb eines Items systematisch zu erfassen (vgl. Abschnitt 5.3.2).

Neben den in der vorliegenden Arbeit vollzogenen faktoriellen Untersuchungen ist die Analyse von DIF ein bedeutungsvoller Zugang zur Aufdeckung von Mehrdimensionalität. Während in der vorliegenden Arbeit Items mit DIF in erster Linie als überarbeitungs- oder ausschlusswürdig klassifiziert wurden, kann DIF alternativ der Aufdeckung von ungewollten systematischen Einflussfaktoren dienen. Für die gängigen DIF-Analysen (z. B. Wald-Test) muss jedoch im Vorfeld bestimmt werden, welche Subgruppen der Stichprobe auf DIF untersucht werden sollen. Damit ist nicht ausgeschlossen, dass wichtige Subgruppen mit DIF nicht untersucht und somit nicht erkannt werden. Eine Methodengruppe, in der diese Vorabspezifikation nicht notwendig ist, sind latente Klassenanalysen im Rahmen des Rasch-Modells (Rost, 1990). In diesem sogenannten Mixed-Rasch-Modell werden Gruppen identifiziert, in denen das Rasch-Modell gilt, die Itemparameter sich aber zwischen den Gruppen unterscheiden (Rost, 1990). Die gemeinsamen Eigenschaften der Personen innerhalb einer Klasse, die ursächlich für DIF sind, müssen in diesem Verfahren jedoch post-hoc ermittelt werden. Das erschwert die Interpretation der Klassen in Bezug auf ihre spezifischen Merkmale.

Einen vielversprechenden Ansatz zur Aufdeckung und Interpretation von DIF entwickelten Strobl, Kopf und Zeileis (2013). In ihrem Ansatz, der auf modelbasierter rekursiver Partitionierung beruht, muss (1) das Teilungskriterium nicht vorgegeben werden, sondern wird im Verfahren ermittelt, und (2) werden Kombinationen von Personeneigenschaften aufgedeckt, die DIF verursachen. Dieses Vorgehen verspricht ein differenziertes Bild darüber, welche Eigenschaftsausprägungen und Kombinationen von Eigenschaftsausprägungen zu signifikanten Abweichungen in der Itemparameterschätzung führen. Eine eingehende Analyse von DIF unter Einbezug der beschriebenen statistischen Methode und die Analyse und Diskussion der Ergebnisse mit fachdidaktischen und pädagogischen Experten sind ein weiteres Forschungsdesiderat der Zukunft.

Schlussendlich bietet die Zusammenführung der statistisch ermittelten DIF-Kennwerte mit den dafür ursächlichen Eigenschaftskombinationen ein gutes Bild darüber, wie und warum Aufgaben für unterschiedliche Subgruppen funktionieren. Dies ist sowohl von testtheoretischem als auch von didaktischem Erkenntnisinteresse.

Ebenso wie unentdeckte Mehrdimensionalität können weitere Eigenschaften des Testinstruments sowie der Testsituation ursächlich für Fehlerquellen in der Messung eines Merkmals sein.

Im Speziellen sind Multiple-Choice-Tests anfällig dafür, von Ratetendenzen verfälscht zu werden. Je weniger Antwortoptionen dargeboten werden, umso wahrscheinlicher ist es, für Testteilnehmende durch Raten oder durch Ausschluss eindeutig falscher Antwortoptionen einen Punkt in einer Aufgabe zu erzielen, ohne über das lösungsrelevante Wissen zu verfügen. Solchen Ratetendenzen wurde in der vorliegenden Arbeit in zwei Schritten entgegengewirkt: (1) Die Pilotierung wurde genutzt, um diejenigen Distraktoren zu identifizieren, die offensichtlich so falsch waren, dass kaum ein Studierender sie ausgewählt hat (diese Antwortoptionen wurden für die Haupterhebung „attraktiver" gestaltet, um ein systematisches Ausschließen von Antwortoptionen zu erschweren). (2) Eine der vier Antwortalternativen wurde so formuliert, dass sie teilrichtige Elemente enthielt. Dieses Design der Antwortoptionen erhöht die Schwierigkeit der Items, da mit Teilwissen möglicherweise die teilrichtige Lösung als vermeintlich richtige Lösung gewählt wird. Unter Reliabilitätsgesichtspunkten wurde neben dem dichotom bepunkteten Rasch-Modell Berechnungen durchgeführt, in denen teilrichtige Lösungen mit einem Punkt und richtige Lösungen mit zwei Punkten bewertet wurden. Das dafür verwendete Partial-Credit-Modell von Masters (Masters, 1982) wies allerdings auf Itemebene eine so schlechte Passung auf, dass das Modell verworfen werden musste und auf eine ausstehende Überarbeitung des Teilpunktelösungsschlüssels (siehe Anhang E, Tabelle E-2) verwiesen werden muss.

Neben den Testeigenschaften sollten auch die Testbedingungen, unter denen die Daten erhoben wurden, einer kritischen Betrachtung unterzogen werden. Das Vorgehen bei der Datenerhebung in der vorliegenden Arbeit wird dem low-stakes-testing zugeordnet. Damit werden Testsituationen beschrieben, in denen die Testteilnahme und das Testergebnis ohne bedeutsame Konsequenzen für die teilnehmende Person bleiben. Im Vergleich zum high-stakes-testing, in dem bedeutsame Konsequenzen mit der Testteilnahme verbunden sind, wird insbesondere die Testmotivation in „low-stakes"-Testsituationen in der einschlägigen Literatur problematisiert (Wise & DeMars, 2005). Es wird kritisiert, dass Testteilnehmer ohne äußere Anreize nicht die Leistung erbringen, die sie in einer „ernsten" Testsituation erbringen würden. Die Befundlage zur

Wirkung der Setzung von Anreizen in Leistungstests ist jedoch nicht eindeutig und weist auf Komplexe Interaktionen zwischen Person und Incentive hin (z. B. Bielinska-Kwaspisz, 2013). Die Datenerhebung der vorliegenden Arbeit erfolgte für alle Teilnehmer freiwillig. Lediglich am Standort Mannheim erhielten die Studierenden als Ausgleich für den Zeitaufwand eine Versuchspersonenstunde[14], die unabhängig von der Leistung im Test im Anschluss an die Teilnahme gutgeschrieben wurde. Da kein Leistungsunterschied zwischen den Standorten vorlag, ist diese Variation in den Testbedingungen jedoch vernachlässigbar. Neben den vermuteten motivationsmindernden Faktoren kann eine freiwillige Testteilnahme auch Selektionseffekte in Richtung einer erhöhten Testmotivation begünstigen. Die Durchführung einer freiwilligen Testung wurde den Studierenden mindestens eine Woche vor der Testung angekündigt, sodass hier möglicherweise eine Selbstselektion besonders motivierter Studierender stattgefunden hat. Damit ist gemeint, dass womöglich nur Studierende an dem Test teilgenommen haben, die über ein Mindestmaß an intrinsischer Motivation verfügt haben und damit möglicherweise eine Verzerrung der mittleren Testwerte zum Positiven begünstigt wurde.

Insgesamt war die Erhebungssituation in Vorlesungsräumen mit bis zu 180 Studierenden nicht vollständig kontrollierbar, was möglicherweise zu unsystematischen Störeinflüssen auf die Messung geführt hat. Der Itempositionseffekt (vgl. Abschnitt 9.3.2.2) wies darauf hin, dass die Konzentration und Motivation im Verlauf der Testung nachgelassen haben. Unter wenig standardisierten Bedingungen, in denen auf die freiwillige Teilnahme der Testteilnehmer gebaut wird, kann eine hohe Testakzeptanz dazu führen, die Motivation und möglicherweise auch die Leistungsfähigkeit über einen längeren Zeitraum positiv zu beeinflussen. Die vergleichsweise schlechten Ergebnisse für die Akzeptanz (vgl. Abschnitt 9.3.4.2) der situativen Items sind vor diesem Hintergrund besonders kritisch zu bewerten. Studierende erlebten die situativen Items tendenziell als belastender, hatten während der Testung weniger Kontrollerleben und schätzten sowohl die curriculare Validität als auch die Messqualität des Tests schlechter ein als die des deklarativen BAKTs (Bothe, 2003). Insbesondere die letzten beiden Einschätzungen von Seiten der Studierenden sind nicht deckungsgleich mit den empirischen Befunden. Die 23 ausgewählten Items des BAKT (Bothe, 2003) wiesen genau das gleiche Cronbachs Alpha von .71 wie die 26 situativen Items auf. Bezüglich der curricularen Validität wiesen die empirischen Ergebnisse eher darauf hin, dass die Anforderungen der deklarativen Wissensitems nicht mit den Anforderungen der BWL-Pflichtmodule zusammenhängen. Trotz dieser Diskrepanzen zwischen der Einschätzung der Items durch die Stu-

14 Die Prüfungsordnung gibt vor, dass eine bestimmte Stundenzahl als Teilnehmer in Studien der Fakultät absolviert werden muss (Universität Mannheim, 2009).

dierenden und den Ergebnissen der Validierung sollte die Einschätzung der Studierenden ernst genommen werden, insbesondere weil auch die selbstberichtete Anstrengungsbereitschaft für situative Aufgaben geringer ausfiel als für die deklarativen Aufgaben. Unter Bezug auf die Forschungsergebnisse von Kanning (2008) (vgl. Abschnitt 5.2) birgt eine multimediale Umsetzung der situativen Items möglicherweise eine Chance, die situativen Items für Studierende attraktiver zu gestalten. Eine Umsetzung der Situationen in Video oder Bild mit möglichst realistischen Elementen könnte den Realitätsbezug der Aufgaben erhöhen und somit ansprechender für Studierende sein. Eine computergestützte Einbettung der Items würde außerdem eine randomisierte Darbietung aller Items erleichtern und die Menge an zu lesendem Text reduzieren. Damit wäre gleichzeitig der Effekt der Lesefähigkeit auf die Messung des eigentlichen Konstrukts minimiert (vgl. die kritische Betrachtung situativer Items in Abschnitt 5.3.2). Bei einer digitalen Umsetzung von Testitems ist jedoch zu beachten, dass die Testteilnehmer hohe Anforderungen an die technischen Umsetzungen solcher Tests haben. Mit einer anwenderfreundlichen und möglichst realitätsnahen Umsetzung ist ein finanzieller und zeitlicher Testentwicklungsaufwand verbunden, der erst ab sehr großen Fallzahlen durch Vorteile der automatischen Datenspeicherung und Datenweiterverarbeitung ausgeglichen wird. Trotzdessen wird die multimediale Umsetzung von situativen Testitems als zukunftsweisend bewertet und sollte unbedingt weiter erforscht werden. In diesem Zusammenhang wäre es interessant, zu prüfen, welche prognostische Validität die beiden in der vorliegenden Arbeit eingesetzten Tests in Bezug auf spätere berufliche Tätigkeiten aufweisen und inwiefern diese zum Beispiel durch eine multimediale Umsetzung der (situativen) Testitems erhöht werden kann. Für die Prüfung der prognostischen Validität wäre es notwendig, im zeitlichen Versatz zum Einsatz der Wissenstests Kriterien des Berufserfolgs zu erheben. Ein solches Follow-up-Erhebungsdesign ist vielversprechend in Hinblick auf die erwarteten Forschungsergebnisse, jedoch mit großem Aufwand verbunden. Je nachdem, in welchem zeitlichen Abstand die Nachbefragung oder die Nachbefragungen stattfinden, ist mit einer hohen Panelmortalität zu rechnen, die wiederum weitere methodische Probleme nach sich zieht (Windzio & Grotheer, 2002). Trotzdem sollte die Realisierung von Studien mit mehreren Messzeitpunkten inklusive Follow-up-Untersuchungen zur prognostischen Validität in zukünftigen Forschungsvorhaben nicht aus dem Blick verloren werden.

Neben der kritischen Betrachtung der theoretischen und methodischen Umsetzung der vorliegenden Arbeit sollen abschließend einige allgemeine Potenziale für verbesserte Forschungsdesigns im Bereich der Messung von Wissen an Hochschulen aufgezeigt werden. Dazu zählt unter anderem die kritische Betrachtung von Daten aus Selbstberichten als Validierungskriterien. Zwar erwie-

sen sich selbstberichtete Noten von Gymnasiasten als weitgehend realitätsgetreu (Sparfeldt, Buch, Rost & Lehmann, 2008), die Einschätzung der eigenen Lernstrategien ist hingegen weitaus weniger verlässlich (Lind & Sandmann, 2003). Eine Möglichkeit, Selbstberichte zu umgehen, ist zum Beispiel die handlungsnahe Erfassung des Lernstrategiegebrauchs über die Auswertung von Protokollen des lauten Denkens (Lind & Sandmann, 2003). Die Anstrengungsbereitschaft während eines Tests könnte in einer computergestützten Testung zum Beispiel über die Zeit, die in die Lösung einer Aufgabe investiert wurde (time on task), gemessen werden. Darüber hinaus bieten moderne Methoden die Möglichkeit, den Verlauf der visuellen Aufmerksamkeit am Computer über Augenbewegungen zu dokumentieren und darüber Rückschlüsse auf Fokus und Verlauf der Aufmerksamkeit zu ziehen (Duchowski, 2007). Die tatsächliche Lern- und Leistungsmotivation wäre möglicherweise durch die tatsächliche Stundenzahl am Schreibtisch oder in der Bibliothek besser beschrieben als durch die Beantwortung der Frage auf einer Skala von sehr selten (1) bis sehr oft (7). Der Versuch, Wissensprädiktoren zu erfassen, die nicht auf Selbstberichtdaten basieren, wirft noch viele Fragen der Umsetzbarkeit auf (z. B. begrenzte Fallzahlen durch hohen Auswertungsaufwand, technische Herausforderungen, datenschutzrechtliche Bedenken), könnte aber die Forschung zu Determinanten des Wissens von Studierenden maßgeblich bereichern.

Für die Vorhersage von Testleistungen sollte darüber hinaus die Entwicklung und Absicherung komplexer Modelle weiter betrieben werden. Modelle des universitären Lernens, wie zum Beispiel das 3P (presage-process-product) Modell von Biggs (1993) bieten eine gute Grundlage, Lehren und Lernen an Hochschulen besser zu verstehen und Bedingungen von Lernleistungen zu identifizieren. Als Bedingungen des Lernens an Hochschulen werden sowohl Studierendencharakteristika (Vorwissen, kognitive Voraussetzungen, Lernstile, Motivation und Erwartungen) als auch Kontextvariablen wie Lehrmethode und Lehrqualität berücksichtigt. Zur Beschreibung der Prozessvariable wird die Tiefe der kognitiven Verarbeitung herangezogen (tief vs. oberflächlich) (Biggs, 1993). Als Lernprodukt können sowohl Noten als auch Testleistungen herangezogen werden. Des Weiteren wäre es, wie bereits im vorangegangenen Abschnitt beschrieben, interessant, als langfristiges Lernergebnis Kriterien des Berufserfolges in die Untersuchungen einzubeziehen. In diesem Zusammenhang sollten durch längsschnittliche Forschungsdesigns mit Erhebungen zu mehreren Messzeitpunkten Wachstumsmodelle getestet werden, um Wissensentwicklung an Hochschulen abzubilden. Auch Rahmen dieser Modelle studentischer Wissens- und Kompetenzentwicklung sollten verschiedene Wissensarten und deren Vernetzung über die Zeit betrachtet werden (für eine erste Konzeptualisierung von Kompetenzentwicklungsmodellen in der kaufmännischen Bildung siehe Winther (2011)). Anhand von Längsschnittdaten ließen sich auch

Fragen der Kausalität besser beantworten als mit nur einem Messzeitpunkt. Die statistische Überprüfung komplexer Modelle bedarf sowohl einer großen Itemzahl als auch möglichst großer Fallzahlen, die in zukünftigen Forschungsvorhaben realisiert werden sollten.

Kritische Betrachtung der Funktion von Assessments an Universitäten

In der Einleitung dieser Arbeit wurde aufgezeigt, welchen potenziellen Nutzen standardisierte Assessments im Bildungssystem haben. Dabei wurden Nutzenaspekte auf individueller und institutioneller Ebene sowie auf Ebene des Bildungssystems und im Bereich der Forschung herausgestellt. Zu einer kritischen Auseinandersetzung mit der vorliegenden Arbeit gehört neben der Betrachtung theoretischer und methodischer Entwicklungsbedarfe auch die Klärung der Frage nach der Funktion und dem praktischen Nutzen von standardisierten Assessments an Hochschulen.

Der Diskussion um die Funktion von Assessments im Bildungswesen liegen zwei Forschungstraditionen zugrunde. Vornehmlich in den Vereinigten Staaten von Amerika hat es sich über Jahrzehnte etabliert, standardisierte Assessments und deren Ergebnisse unabhängig von den jeweiligen Curricula oder der üblichen Instruktionspraxis zu entwickeln, durchzuführen und zu interpretierten (Achtenhagen, 2012). Im europäischen Forschungsraum und speziell in Deutschland lag der Fokus der Bildungsforschung lange Zeit auf curricularen und didaktischen Überlegungen unter weitgehender Vernachlässigung des dazugehörigen Assessmentaspektes (Achtenhagen, 2012). Aktuell wird sowohl in der deutschen als auch in der englischsprachigen Literatur kritisch diskutiert, wie Assessments im Bildungswesen gestaltet sein müssen, um ein integrativer Bestandteil erfolgreicher Umsetzung von Lehr-Lernprozessen zu sein. Die „Curriculum-Instruction-Assessment Triade" von Pellegrino (2006) bietet dabei ein überzeugendes Modell, das beide oben aufgezeigten Forschungstraditionen vereint. In dem Modell umschreibt das Curriculum, welches Wissen und welche Fertigkeiten innerhalb eines Fachbereiches gelehrt werden und gelernt werden sollen (Pellegrino, 2006, S. 2). Instruction steht für den Instruktionsaspekt und somit für die Lehrmethoden und Lernaktivitäten, die die Lernenden dabei unterstützen, die im Curriculum spezifizierten Lernziele zu erreichen (Pellegrino, 2006, S. 2). Assessment ist in diesem Zusammenhang das Mittel, um Ergebnisse dieses Bildungsprozesses im Hinblick auf die Leistungen der Studierenden und deren Aneignung wichtiger Kompetenzen messbar zu machen (Pellegrino, 2006, S. 2). Im Einklang mit Pellegrino (2006) argumentierte Achtenhagen (2012), dass alle drei Aspekte der Triade aufeinander abgestimmt sein müssen, damit Assessments innerhalb eines Bildungssystems ihren vollen Nutzen entfalten können und vor allem ein faires Beurteilungsinstrument darstel-

len. Denn nur, wenn Assessments auf die curricularen und instruktionalen Besonderheiten in einem Bereich abgestimmt sind, können diese für Rückschlüsse über die Wirksamkeit von Lehren und Lernen genutzt werden und Verbesserungspotenziale im Curriculum identifizieren. Gelingt es, Curriculum, Instruktion und Assessment aufeinander auszurichten, so dienen Assessments nicht nur der Kontrolle von Lernergebnissen, sondern können ebenfalls genutzt werden, um rückwirkend instruktionale Praktiken zu verbessern oder auf notwendige Veränderungen im Curriculum hinzuweisen.

Genau diese Passung zwischen Curriculum, Instruktion und Assessment ist im Bereich der deutschen Hochschulbildung noch schwach ausgeprägt. Pellegrino (2006, S. 2) fasst die Konsequenzen einer solchen schlechten Passung treffend mit dem Ausdruck „overall incoherence in the educational enterprise" zusammen. Das Deutsche Hochschulsystem befindet sich seit der Umsetzung des Bologna-Prozesses (Bologna-Declaration, 1999) in einem stetigen Veränderungsprozess (Zinger, 2012). Rahmencurricula liegen für betriebswirtschaftliche Studiengänge nicht vor[15]. Dementsprechend ist eine kohärente Ausrichtung von Curriculum, Instruktion und Assessment in diesem Bereich zumindest auf formaler Ebene kaum erkennbar und aufgrund mangelnder Forschungsarbeiten in diesem Bereich auch nicht systematisch überprüfbar. Die Forderung, dass der Bachelorabschluss berufsqualifizierend sein soll, hat bisher keine übergreifende, sondern höchstens punktuelle Verankerung in den betriebswirtschaftlichen Studiengängen, den Lehr-Lernformen und den hochschultypischen Assessments gefunden (Wolter & Banscherus, 2012). Zum jetzigen Zeitpunkt liegt mit dem Bologna-Prozess eine klare übergeordnete Zielsetzung für Bachelorstudiengänge vor. Bisher wurde aber nicht verbindlich fixiert, wie diese in einzelnen Studiengängen erreicht werden soll. An dieser Situation wird die Problematik der Passung zwischen bildungspolitischen Zielvorgaben und curricularer sowie instruktionaler Umsetzung deutlich. Zum Beispiel bleibt auf institutioneller Ebene unklar, wie welche Wissensarten im Bachelorstudium erlernt werden sollen und welche kognitiven Anforderungen am Ende eines Studiums bewältigbar sein sollen. Ebenso unsystematisch variieren die Lehr-Lernformen innerhalb und zwischen den Universitäten. So sind Praktika an einigen Universitäten im Bachelorstudium der Betriebswirtschaftslehre als verpflichtend verankert und anderen nicht. Vor dem Hintergrund dieser derzeit brüchigen Beziehung zwischen den Komponenten der Triade kommt Assessments im Hochschulbereich nicht die Aufgabe zu, Curriculum und Instruktion zu über-

15 Derzeit wird nach Beschluss der Mitgliederversammlung der Sektion Berufs- und Wirtschaftspädagogik in Oldenburg am 25. März 2003 an einem Basiscurriculum für das universitäre Studienfach Berufs- und Wirtschaftspädagogik gearbeitet. (Beschluss abrufbar unter: http://www.bwp-dgfe.de/sektion/positionen/curriculum/)

prüfen oder studentische Leistungen über Studienstandorte hinweg zu vergleichen. Vielmehr liegt ihr Potenzial darin, Impulse für die Entwicklung von Curriculum und Instruktion sowie deren stringente gemeinsame Ausrichtung zu geben. Die Ergebnisse der vorliegenden Arbeit zum Leistungsunterschied von Studierenden bei deklarativen und anwendungsorientierten Tests werfen zum Beispiel die Frage auf, ob das Universitätsstudium der Betriebswirtschaftlehre die Studierenden ausreichend zur Lösung handlungsorientierter Aufgaben befähigt. Folgt man dem Anspruch, dass der Bachelor berufsqualifizierend sein soll, weisen die Ergebnisse darauf hin, dass das Anwenden von Wissen auf berufliche Handlungssituationen in der Lehrpraxis der Universitäten weiter ausgebaut werden könnte.

10.3 Fazit und Ausblick

In Europa und in Deutschland ist das Interesse an der Erforschung des betriebswirtschaftlichen Wissens und Könnens von Studierenden in den letzten Jahren kontinuierlich gestiegen. Die vorliegende Arbeit leistet durch die Entwicklung eines validen situativen Testinstrumentes einen Beitrag auf dem Weg hin zu einem besseren Verständnis der Erfassung und Entstehung betriebswirtschaftlichen Wissens an deutschen Hochschulen. Ausdrücklich wird die Forderung untermauert, Wissen nicht als Entität zu betrachten, sondern unterschiedliche Wissensfacetten durch geeignete Messinstrumente differenziert abzubilden und zu analysieren. Dafür wurden situative Aufgaben entwickelt und unter methodischen Gesichtspunkten kritisch betrachtet. Aus der kritischen Betrachtung des Forschungsstandes und den Limitationen der vorliegenden Arbeit haben sich drei zentrale Forschungsdesiderata ergeben, die für den Fortschritt der empirischen Bildungsforschung im Bereich der Hochschulbildung als wichtig erachtet werden. Dazu gehören:

(1) eine fachübergreifende Modellierung der betriebswirtschaftlichen Domäne sowie eine stringente Ausrichtung von Curriculum, Instruktion und Assessment.

(2) die Entwicklung und Prüfung komplexer Modelle zur Vorhersage der Studien- und Testleistung und deren Entwicklung über die Zeit sowie die Prüfung mehrdimensionaler und mehrparametrischer Messmodelle.

(3) die professionelle Umsetzung und Erforschung computergestützter Assessments im Hochschulbereich, die die Möglichkeit bieten, insbesondere beruflich relevante handlungsnahe Wissenskonstrukte zu erfassen.

Sollte es in Zukunft gelingen, durch den Einsatz wissenschaftlich abgesicherter Tests mehr Licht in die „Black Box" des Wissenserwerbs und der Wissensent-

wicklung an Hochschulen zu bringen, so würde das Anliegen dieser Arbeit fortgesetzt.

Literaturverzeichnis

Abele, S., Greiff, S., Gschwendtner, T., Wüstenberg, S., Nickolaus, R., Nitzschke, A. & Funke, J. (2012). Dynamische Problemlösekompetenz. *Zeitschrift für Erziehungswissenschaft, 15* (2), 363–391.

Abele, S. & Nickolaus, R. (2013). Probleme reliabler und valider Kompetenzmessung bei eingeschränkten Testzeiten – Theoretische Lösungsansätze und erste Ergebnisse empirischer Untersuchungen [Abstract]. In V. Bank (Hrsg.), *Beruf und Erziehung Abstractband zur Chemnitzer Herbsttagung der Sektion Berufs- und Wirtschaftspädagogik der Deutschen Gesellschaft für Erziehungswissenschaften. Berichte aus der Berufs- und Wirtschaftspädagogik Papers and Proceedings in Vocationomics.* (S. 30). Chemnitz: Professur für Wirtschaftspädagogik.

Achtenhagen, F. & Winther, E. (2011). Fachdidaktische Perspektiven der Kompetenzmessung–am Beispiel des kaufmännisch-verwaltenden Bereichs. In O. Zlatkin-Troitschanskaia (Hrsg.), *Stationen Empirischer Bildungsforschung. Traditionslinien und Perspektiven* (S. 352–367). Hohengehren: Schneider.

Achtenhagen, F. (2012). The curriculum-instruction-assessment triad. *Empirical Research in Vocational Education and Training, 4* (1), 5–25.

Adams, R. J., Wilson, M. & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21* (1), 1–23.

AERA, APA & NCME. (2002). *Standards for educational and psychological testing (5. Aufl.)*. Washington, DC: American Psychological Association.

Agarwal R. & Day, A. E. (1998). The impact of the internet on Economic Education. *The Journal of Economic Education, 29* (2), 99–110.

AHELO Consortium. (2013). *Assessment of higher education learning outcomes AHELO. Feasibility study report volume 2 – Data analysis and national experiences*. OECD. Zugriff am 12.09.2013 http://www.oecd.org/edu/skills-beyond-school/AHELOFSReportVolume2.pdf

Ahmed, H., Rhydderch, M. & Matthews, P. (2012). Can knowledge tests and situational judgement tests predict selection centre performance? *Medical Education, 46* (8), 777–784.

Alagumalai, S. & Curtis, D. (2005). Classical Test Theory. In J. P. Keeves & S. Alagumalai (Hrsg.), *Applied Rasch measurement. A book of exemplars: papers in honour of John P. Keeves* (1–14). Dordrecht: Springer.

Amelang, M. & Schmidt-Atzert, L. (2006). *Psychologische Diagnostik und Intervention* (4. Aufl.). Berlin: Springer.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38* (1), 123–140.

Anderson, G., Benjamin, D. & Fuss, M. (1994). The determinants of success in university introductory economics courses. *The Journal of Economic Education, 25* (2), 99–119.

Anderson, J. R. (1976). *Language, memory, and thought*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior, 22,* 261–295.

Anderson, J. R. (1996). A simple theory of complex cognition. *American Psychologist, 51* (4), 355–365.

Anderson, J. R. (2001). *Kognitive Psychologie* (3. Aufl.). Heidelberg: Spektrum.

Anderson, J. R. & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, N.J.: Lawrence Erlbaum Associates.

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: Longman.

Andrich, D. (2011). Rating scales and Rasch measurement. *Expert review of pharmacoeconomics & outcomes research, 11* (5), 571–585.

Artelt, C. (2000). Wie prädiktiv sind retrospektive Selbstberichte über den Gebrauch von Lernstrategien für strategisches Lernen. *Zeitschrift für Pädagogische Psychologie, 14* (2), 72–84.

Auspurg, K., Hinz, T., Liebig, S. & Sauer, C. (2009). Auf das Design kommt es an. Experimentelle Befunde zu komplexen Settings in Faktoriellen Surveys. *soFid Methoden und Instrumente der Sozialwissenschaften*, 23–40. Zugriff am 12.09.2013 http://pub.uni-bielefeld.de/luur/download?func=downloadFile&recordOId=1901715&fileOId=2242211

Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2003). *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung* (10. Aufl.). Berlin: Springer.

Bagamery, B. D., Lasik, J. J. & Nixon, D. R. (2005). Determinants of success on the ETS business major field exam for students in an undergraduate multisite regional university business program. *Journal of Education for Business, 81* (1), 55–63.

Bank, V. & Retzmann, T. (2011, Oktober). *Bedarfsermittlung im Wirtschaftskundlichen Unterricht, oder: Der WBT und seine Eignung zur Weiterbildungsbedarfsanalyse bei Lehrkräften.* Vortrag auf der Herbsttagung der Sektion Berufs- und Wirtschaftspädagogik der DGfE in Konstanz.

Bartlett, S., Peel, M. J. & Pendlebury, M. (1993). From fresher to finalist: a three year analysis of student performance in the first college course. *Accounting Education, 2* (2), 111–122.

Baumert, J., Bos, W. & Lehmann, R. (Hrsg.). (2000). *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe.* Opladen: Leske + Budrich.

Baumert, J. & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft, 9* (4), 469–520.

Beaton, A. E. & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics, 17* (2), 191–204.

Beck, K., Krumm, V. & Dubs, R. (1998). *Wirtschaftskundlicher Bildungs-Test (WBT)*. Göttingen: Hogrefe.

Bender, N. (2011). Die Abbildung vernetzten Wissens zur privaten Ver- und Überschuldung mit Concept Maps. In U. Faßhauer, J. Aff, B. Fürstenau & E. Wuttke (Hrsg.), *Lehr-Lernforschung und Professionalisierung. Perspektiven der Berufsbildungsforschung* (S. 99–110). Opladen: Barbara Budrich.

Bielinska-Kwapisz, A. & Brown, F. W. (2013). Differential gender performance on the major field test–business. *Journal of Education for Business, 88* (3), 159–166.

Biggs, J. B. (1993). From theory to practice: A cognitive systems approach. *Higher education research and development, 12* (1), 73–85.

Black, H. T. & Duhon, D. L. (2003). Evaluating and improving student achievement in business programs: The Effective Use of Standardized Assessment Tests. *Journal of Education for Business, 79* (2), 90–98.

Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung & Bundesagentur für Arbeit (2004). Studien- und Berufswahl – Informationen und Entscheidungshilfen. Nürnberg: BW Bildung und Wissen.

Blömeke, S. (2013). Ja, aber … – Lehren aus TEDS-M und KoKoHs für eine Teilnahme an AHELO. In Hochschul Informations System GmbH (Hrsg.), *AHELO goes Germany? Dokumentation des GfHf- & HIS-HF-Workshops* (S. 13–20). Zugriff am 12.09.2013 http://www.his.de/pdf/pub_fh/fh-201302.pdf

Bloom, B. S. (1972). *Taxonomie von Lernzielen im kognitiven Bereich*. Weinheim: Beltz.

Bologna Declaration. (1999). *The Bologna Declaration of 19 June 1999. Joint declaration of the European Ministers of Education* (European Union, Hrsg.), Brussels. Zugriff am 12.09.2013 http://www.bologna-berlin2003.de/pdf/bologna_declaration.pdf

Bologna Working Group. (2005). *A framework for qualifications of the European higher education area* (Ministry of Science Technology, Hrsg.), Kopenhagen. Zugriff am 19.09.2013 http://www.bologna-bergen2005.no/Docs/00-Main_doc/050218_QF_EHEA.pdf

Bond, T. G. & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, N.J.: Lawrence Erlbaum Associates.

Bortz, J. (1999). *Statistik. Für Sozialwissenschaftler* (5. Aufl.). Berlin: Springer.

Bothe, T. (2003). *Entwicklung und Erprobung eines deklarativen Wissenstests*. Unveröffentlichte Diplomarbeit, Universität Mannheim.

Bothe, T., Wilhelm, O. & Beck, K. (2005). Assessment of declarative business administration knowledge: Measurement development and validation. Unveröffentlichtes Manuskript.

Brunstein, J. C. & Heckhausen, H. (2010). Leistungsmotivation. In J. Heckhausen & H. Heckhausen (Hrsg.), *Motivation und Handeln* (S. 145–192). Berlin: Springer.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3. Aufl.). München: Pearson Studium.

Bühner, M. & Ziegler, M. (2009). *Statistik für Psychologen und Sozialwissenschaftler*. München: Pearson Studium.

Bühner, M., Ziegler, M., Krumm, S. & Schmidt-Atzert, L. (2006). Ist der I-S-T 2000 R Rasch-skalierbar? *Diagnostica, 52* (3), 119–130.

Bülow-Schramm, M. & Braun, E. (2013). Einleitung. In Hochschul Informations System GmbH (Hrsg.), *AHELO goes Germany? Dokumentation des GfHf- & HIS-HF-Workshops* (S. 1–8). Zugriff am 12.09.2013 http://www.his.de/pdf/pub_fh/fh-201302.pdf

Bycio, P. & Allen, J. S. (2007). Factors associated with performance on the educational testing service (ETS) major field achievement test in business (MFAT-B). *The Journal of Education for Business, 82* (4), 196–201.

Campbell, D. T. & Fiske, D. W. (1959b). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56* (2), 81.

Cattell, R. B. (1974). Radial parcel factoring-vs-item factoring in defining personality structure in questionnaires: Theory and experimental checks. *Australian Journal of Psychology*, *26* (2), 103–119.

Chen, Y., & Hoshower, L. B. (2003). Student evaluation of teaching effectiveness: An assessment of student perception and motivation. *Assessment & evaluation in higher education*, *28*(1), 71–88.

Christian, M. S., Edwards, B. D. & Bradeley, J. C. (2010). Situational judgement Tests. Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology, 63* (1)*,* 83–117.

Cleary, T. A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement, 5* (2), 115–124.

Cohen, A. S. & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42* (2), 133–148.

Cohn, E. & Johnson, E. (2006). Class attendance and performance in principles of economics. *Education Economics, 14* (2), 211–233.

Cox, S. S., Charles, L., Chen, L. J.-S. & Totten, L. J. (2011). Is it more than just GPA? An examination of work experience and test preparation effects on MFT-B scores. *Journal of Academic Administration in Higher Education, 7* (1), 53–70.

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

Dickhäuser, O., Schöne, C., Spinath, B. & Stiensmeier-Pelster, J. (2002). Die Skalen zum akademischen Selbstkonzept. *Zeitschrift für Differentielle und Diagnostische Psychologie, 23* (4), 393–405.

Donlon, T. F. & Angoff, W. H. (1971). The scholastic aptitude test. *The College Board admissions testing program.* New York: College Entrance Examination Board.

Dörfler, R. (2005). Diplom, Bachelor, Master aus Sicht der Wirtschaft. *Beiträge zur Hochschulforschung, 27* (3), 108–112.

Dörner, D., Schaub, H. & Strohschneider, S. (1999). Komplexes Problemlösen Königsweg der Theoretischen Psychologie. *Psychologische Rundschau, 50* (4), 198–205.

Dostal, W. & Reinberg, A. (1999). *Ungebrochener Trend in die Wissensgesellschaft. Entwicklung der Tätigkeiten und Qualifikationen.* IAB Kurzbericht: Aktuelle Analysen aus dem Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit: 10. Zugriff am 26.06.2012 http://doku.iab.de/kurzber/1999/kb1099.pdf.

Draney, K. & Wilson, M. (2008). A LLTM approach to the examination of teachers' ratings of classroom assessment tasks. *Psychology Science, 50* (3), 417.

Duchowski, A. T. (2007). *Eye tracking methodology: Theory and practice.* Berlin: Springer.

Duff, A. (2004). Understanding academic performance and progression of first-year accounting and business economics undergraduates: the role of approaches to learning and prior academic achievement. *Accounting Education, 13* (4), 409–430.

Edelmann, W. (1996). *Lernpsychologie*. Belz: Weinheim.

Educational Testing Service. (2011). *Find out how to prove — and improve — the effectiveness of your Business program with the ETS® Major Field Tests.* Zugriff am 09.10.2013 http://www.ets.org/Media/Tests/MFT/pdf/mft_testdesc_business_4 cmf.pdf.

Educational Testing Service. (2012). *Major Field Test Comparative Data Guide (CDG)*. Princeton: Educational Testing Service (ETS). Zugriff am 09.10.2013 http://www.ets.org/s/mft/pdf/business4gmf.pdf.

Eggert, S. & Bögeholz, S. (2010). Students' use of decision-making strategies with regard to socioscientific issues: An application of the Rasch partial credit model. *Science Education, 94* (2), 230–258.

Eskew, R. K. & Faley, R. H. (1988). Some determinants of student performance in the first-college-level financial accounting course. *The Accounting Review, 63* (1), 137–147.

Fahrmeir, L., Kneib, T., & Lang, S. (2009). *Regression*. *Modelle, Methoden und Anwendungen* (2. Aufl.). Berlin: Springer.

Fiege, C., Reuther, F. & Nachtigall, C. (2011). Faire Vergleiche?–Berücksichtigung von Kontextbedingungen des Lernens beim Vergleich von Testergebnissen aus deutschen Vergleichsarbeiten. *Zeitschrift für Bildungsforschung, 1* (2), 133–149.

Förster, M., Happ, R. & Zlatkin-Troitschanskaia, O. (2012). Valide Erfassung des volkswirtschaftlichen Fachwissens von Studierendenden der Wirtschaftswissenschaften und Wirtschaftspädagogik. Eine Untersuchung der diagnostischen Eignung des Wirtschaftskundlichen Bildungstests (WBT). *bwp@ – Berufs- und Wirtschaftspädagogik – online, 22,* 1–21. Zugriff am 12.09.2013 http:// www.bwpat.de/ausgabe22/foerster_etal_bwpat22.pdf

Förster, M. & Zlatkin-Troitschanskaia, O. (2010). Wirtschaftliche Fachkompetenz bei Studierenden mit und ohne Lehramtsperspektive in den Diplom- und Bachelorstudiengängen. Messverfahren und erste Befunde. *Lehrerbildung auf dem Prüfstand (Sonderheft)* (3), 106–125.

Garey, C. & Ellis, L. (1995). The effects of attendance on student learning in principles of economics. *The American Economic Review, 85* (2), 343–346.

George, D. & Mallery, P. (2000). *SPSS for Windows: A simple guide and reference*. Boston: Allyn & Bacon.

Gottschalk-Mazouz, N. (2007). Was ist Wissen? Überlegungen zu einem Komplexbegriff an der Schnittstelle von Philosophie und Sozialwissenschaften. In S. Ammon, C. Heineke & K. Selbmann (Hrsg.), *Wissen in Bewegung. Vielfalt und Hegemonie in der Wissensgesellschaft* (S. 21–40). Weilerswis: Velbrück Wissenschaft.

Götz, T., Frenzel, A. C. & Pekrun, R. (2009). Psychologische Bildungsforschung. In R. Tippelt & B. Schmidt (Hrsg.), *Handbuch Bildungsforschung* (S. 71–91). Berlin: Springer.

Größler, A., Wilhelm, O., Wittmann, W. W. & Milling, P. M. (2002). *Measuring Business Knowledge for Persormel Selection in Small and Medium Sized Companies*. Institut für Mittelstandsforschung der Universität Mannheim (Hrsg.) (Nr. 44).

Group of National Experts on the AHELO Feasibility Study. (2012). *Economics assessment development report. AHELO Feasibility Study* (OECD, Hrsg.). Zugriff am 12.09.2013 http://search.oecd.org/officialdocuments/displaydocument pdf/?cote=edu/imhe/ahelo/gne%282011%2919/ANN4/FINAL&doclanguage=en.

Gruber, H. (1999). *Erfahrung als Grundlage kompetenten Handelns* (1. Aufl.). Bern: Huber.

Gruber, H., Mandl, H. & Renkl, A. (2000). Was lernen wir in Schule und Hochschule: Träges Wissen? In H. Mandl & J. Gerstenmaier (Hrsg.), *Die Kluft zwischen Wissen und Handeln. Empirische und theoretische Lösungsansätze* (S. 139–152). Göttingen: Hogrefe.

Gschwendtner, T. (2008). Raschbasierte Modellierung berufsfachlicher Kompetenz in der Grundbildung von KraftfahrzeugmechatronikerInnen. In K. Breuer, T. Deißinger & D. Münk (Hrsg.), *Probleme und Perspektiven der Berufs- und Wirtschaftspädagogik aus nationaler und internationaler Sicht* (S. 21–30). Opladen: Barbara Budrich.

Gschwendtner, T., Geißel, B. & Nickolaus, R. (2010). Modellierung beruflicher Fachkompetenz in der gewerblich-technischen Grundbildung. In E. Klieme, D. Leutner & M. Kenk (Hrsg.) Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes. *Zeitschrift für Pädagogik. 56,* 258–269 [Themenheft]. Weinheim: Beltz.

Gul, F. A. & Fong, S.C.C. (1993). Predicting success for introductory accounting students: some further Hong Kong evidence. *Accounting Education, 2* (1), 33–42.

Gutenberg-Universität Mainz. (2013, 05. Februar). *Kompetenzen im Hochschulsektor*. Zugriff am 09.10.2013 http://www.kompetenzen-im-hochschulsektor.de/index.php.

Halpern, N. (2007). The impact of attendance and student characteristics on academic achievement: findings from an undergraduate business management module. *Journal of Further and Higher Education, 31* (4), 335–349.

Happ R. (2013, März). *Entwicklung des volkswirtschaftlichen Fachwissens von Studierenden im Rahmen des ILLEV-Projektes.* Vortrag auf der Frühjahrstagung der Sektion Berufs- und Wirtschaftspädagogik der Deutschen Gesellschaft für Erziehungswissenschaft in Kassel.

Hartig, J. (2007). Skalierung und Definition von Kompetenzniveaus. In B. Beck (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung; DESI-Studie (Deutsch-Englisch-Schülerleistungen-International)* (S. 83–99). Weinheim: Beltz.

Hartig, J. & Frey, A. (2012). Konstruktvalidierung und Skalenbeschreibung in der Kompetenzdiagnostik durch die Vorhersage von Aufgabenschwierigkeiten. *Psychologische Rundschau, 63* (1), 43–49.

Hartig, J., Frey, A. & Jude, N. (2012). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 143–171). Berlin: Springer.

Hartig, J. & Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within-item and between-item multidimensionality. *Journal of Psychology, 216* (2), 89–101.

Hartig, J. & Jude, N. (2007). Empirische Erfassung von Kompetenzen und psychometrische Kompetenzmodelle. In J. Hartig & E. Klieme (Hrsg.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik* (Bildungsforschung, Bd. 20) (S. 17–36). Berlin: BMBF.

Hasselhorn, M., & Labuhn, A. S. (2008). Metakognition und selbstreguliertes Lernen. In W. Schneider & M. Hasselhorn (Hrsg.), *Handbuch der pädagogischen Psychologie* (S. 28–37). Göttingen: Hogrefe.

Henning, C. & Henning, W. (2009). *Studienführer Wirtschaftswissenschaften. Betriebswirtschaft, Volkswirtschaft, internationale Betriebswirtschaft, internationale Volkswirtschaft, Wirtschaftswissenschaften/Ökonomie, Wirtschaftsinformatik, Wirtschaftspädagogik* (7. Aufl.). Eibelstadt: Lexika-Verlag.

Hofmeister, W. (2005). Erläuterung der Klassifikationsmatrix zum ULME-Kompetenzstufenmodell. *bwp@ – Berufs- und Wirtschaftspädagogik- online, 8,* 1–21. Zugriff am 18.06.2012 http://www.bwpat.de/ausgabe8/hofmeister_bwpat8.shtml

Horváth, P., Gleich, R. & Voggenreiter, D. (2001). *Controlling umsetzen.* Stuttgart: Schäffer-Poeschel.

HRK, KMK & BMBF. (2005). Qualifikationsrahmen für Deutsche Hochschulabschlüsse. *Zugriff am 12.09.2013 http://www. Kmk.org/fileadmin/veroeffentlichungen_beschluesse/2005/2005_04_21-Qualifikationsrahmen-HS-Abschluesse.pdf [Stand Juli 2010].*

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6* (1), 1–55.

Hunter, D. R. (2003). Measuring general aviation pilot judgment using a situational judgment technique. *The international journal of aviation psychology, 13* (4), 373–386.

Hußtegge, R. (2011). *Selbstreguliertes Wollen als Bedingung für Studienerfolg an der Universität. Band 2*. Dissertation, Carl von Ossietzky Universität Oldenburg. Zugriff am 12.09.2013 http://oops.uni-oldenburg.de/1197/2/hussel11_2.pdf

Ingenkamp, K. (1997). *Lehrbuch der pädagogischen Diagnostik* (4. Aufl.). Weinheim: Beltz.

Jacobs, C., Heubrock, D., Petermann, F., Kubinger, K. D. & Wurst, E. (2003). Adaptives Intelligenz Diagnostikum 2 (AID 2). *Diagnostica, 49* (4), 184–188.

Jähnig, C. C. (2013). Assessing business knowledge of students in German higher education. In U. Faßhauer, B. Fürstenau & E. Wuttke (Hrsg.), *Jahrbuch der berufs- und wirtschaftspädagogischen Forschung* (S. 47–60). Opladen: Barbara Budrich.

Jahn, K. (2012). *Fachliche Anforderungen an Absolventen der Betriebswirtschaftslehre. Eine Stellenanzeigenanalyse.* Unveröffentlichte Diplomarbeit, Georg-August-Universität Göttingen.

Jong, T. de & Ferguson-Hessler, M. G. (1996). Types and qualities of knowledge. *Educational Psychologist, 31* (2), 105–113.

Kailer, N., Stockinger, A., Daxner, F., Wimmer-Wurm, B., Böhm, D. & Zweimüller, R. (2013). *Entrepreneurship Education in technischen Studienrichtungen.* GRIN Verlag.

Kanning, U. P. (2008). Videogestützte Situational Judgement Tests – Ergebnisse zweier Studien. In W. Sarges & D. Scheffer (Hrsg.), *Innovative Ansätze für die Eignungsdiagnostik* (S. 87–96). Göttingen: Hogrefe.

Kersting, M. (2008). Zur Akzeptanz von Intelligenz-und Leistungstests. *Report Psychologie, 33* (9), 420–433.

Klieme, E. (2004). Was sind Kompetenzen und wie lassen sie sich messen? *Pädagogik, 56* (6)*,* 10–13.

Klieme, E. (2010). *PISA 2009. Bilanz nach einem Jahrzehnt*. Münster: Waxmann.

Klieme, E. & Hartig, J. (2007). Kompetenzkonzepte in der Sozialwissenschaft und im erziehungswissenschaftlichen Diskurs. In M. Prenzel, I. Gogolin & H.-H. Krüger (Hrsg.), *Kompetenzdiagnostik. Zeitschrift für Erziehungswissenschaft Sonderheft 8* (S. 6–11). Wiesbaden: VS Verlag für Sozialwissenschaften.

Klieme, E. & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingereichten Schwerpunktprogramms der DFG. *Zeitschrift für Pädagogik, 52* (6), 876–903.

Klieme, E., Maag-Merki, K. & Hartig, J. (2007). Kompetenzbegriff und Bedeutung von Kompetenzen im Bildungswesen. In J. Hartig & E. Klieme (Hrsg.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik* (S. 1–15). Bonn, Berlin: Bundesministerium für Bildung und Forschung.

Köller, O., Trautwein, U., Lüdtke, O. & Baumert, J. (2006). Zum Zusammenspiel von schulischer Leistung, Selbstkonzept und Interesse in der gymnasialen Oberstufe. *Zeitschrift für Pädagogische Psychologie, 20* (1), 27–39.

König, J. & Blömeke, S. (2009). Pädagogisches Wissen von angehenden Lehrkräften. *Zeitschrift für Erziehungswissenschaft, 12* (3), 499–527.

Krapp, A., Schiefele, U. & Schreyer, I. (1993). Metaanalyse des Zusammenhangs von Interesse und schulischer Leistung. *Zeitschrift für Entwicklungspsychologie und pädagogische Psychologie, 10* (2), 120–148.

Kubinger, K. D. (2000). Und für die Psychologische Diagnostik hat es doch revolutionäre Bedeutung. *Psychologische Rundschau, 51* (1), 33–34.

Kunter, M. (2002). *PISA 2000: Dokumentation der Erhebungsinstrumente.* Berlin: Max-Planck-Institut für Bildungsforschung. Zugriff am 12.09.2013 http://edoc.mpg.de/get.epl?fid=3501&did=14414&ver=0

Latham, G. P., Saari, L. M., Pursell, E. D. & Campion, M. A. (1980). The situational interview. *Journal of Applied Psychology, 56* (4)*,* 422–427.

Lehmann, R. & Seeber, S. (2007). *ULME III. Untersuchungen von Leistungen, Motivation und Einstellungen der Schülerinnen und Schüler in den Abschlussklassen der Berufsschulen.* Hamburg: HIBB.

Leonard, N. & Insch, G. S. (2010). Tacit knowledge in academia: A proposed model and measurement scale. *The Journal of Psychology, 139* (6), 495–512.

Liepmann, D., Beauducel, A., Brocke, B. & Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000 R (IST 2000 R). Manual* (2. Aufl.). Göttingen: Hogrefe.

Lievens, F., Peeters, H. & Schollaert, E. (2008). Situational judgment tests: a review of recent research. *Personnel Review, 37* (4), 426–441.

Lievens, F. & Sackett, P. R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology, 91* (5), 1181–1188.

Lind, G. & Sandmann, A. (2003). Lernstrategien und domänenwissen. *Journal of Psychology, 211* (4), 171–192.

Ling, G. (2012). *Why the major field test in business does not report subscores: Reliability and construct validity evidence*. Princeton, N. J: Educational Testing Service (ETS). Zugriff am 12.09.2013 http://www.ets.org/Media/Research/pdf/RR-12–11.pdf

Linstone, H. A. & Turoff, M. (1975). *The Delphi method*. Reading, Mass.: Addison-Wesley.

Lovett, M. C., Daily, L. Z. & Reder, L. M. (2000). A source activation theory of working memory: cross-task prediction of performance in ACT-R. *Cognitive Systems Research, 1* (2), 99–118.

Lüttinger, P. & König, W. (1988). Die Entwicklung einer international vergleichbaren Klassifikation für Bildungssysteme. *ZUMA Nachrichten, 22* (12), 1–14.

Maier, U. (2012). *Lehr-Lernprozesse in der Schule: Studium: Allgemeindidaktische Kategorien für die Analyse und Gestaltung von Unterricht.* Stuttgart: UTB.

Mair, P., Hatzinger, R. & Maier, J. M. (2012). eRm: Extended Rasch modeling. R package version 0.15–1. Zugriff am 23.09.2013 http://CRAN.R-project.org/package=eRm

Mandl, H. & Gerstenmaier, J. (Hrsg.). (2000). *Die Kluft zwischen Wissen und Handeln. Empirische und theoretische Lösungsansätze*. Göttingen: Hogrefe.

Mandl, H. & Krause, U.-M. (2001). *Lernkompetenz für die Wissensgesellschaft. (Forschungsbericht Nr. 145)*. München: Ludwig-Maximilians-Universität München, Lehrstuhl für Empirische Pädagogik und Pädagogische Psychologie. Zugriff am 19.09.2013 http://epub.ub.uni-muenchen.de/253/1/FB_145.pdf

Marcus, B., Funke, U. & Schuler, H. (1997). Integrity Tests als spezielle Gruppe eignungsdiagnostischer Verfahren: Literaturüberblick und metaanalytische Befunde zur Konstruktvalidität. *Zeitschrift für Arbeits-und Organisationspsychologie*, *41* (1), 2–17.

Marsh, H., Hau, K.-T. & Weng, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural equation modeling, 11* (3), 320–341.

Marsh, H. W. & O'Neill, R. (1984). Self description questionnaire III: the construct validity of multidimensional self-concept ratings by late adolescents. *Journal of Educational Measurement, 21* (2), 153–174.

Martin, M. O., Gregory, K. D., Stemler, S. E. & Foy, P. (2000). *TIMSS 1999 technical report*. Boston: International Study Center.

Marzano, R. J. & Kendall, J. S. (2008). *Designing & assessing educational objectives. Applying the new taxonomy*. Thousand Oaks: Corwin Press.

Mason, P. M., Coleman, B. J., Steagall, J. W., Gallo, A. A. & Fabritius, M. M. (2011). The use of the ETS major field test for assurance of business content learning: Assurance of waste? *Journal of Education for Business, 86* (2), 71–77.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika, 47* (2), 149–174.

McCloy, R. A., Campbell, J. P. & Cudeck, R. (1994). A confirmatory test of a model of performance determinants. *Journal of Applied Psychology, 79* (4), 493–505.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L. & Grubb, L. W. (2007). Situational judgement tests, response instructions, and validity. A meta-analysis. *Personnel Psychology, 60* (1)*,* 63–91.

Mirchandani, D., Lynch, R. & Hamilton, D. (2001). Using the ETS major field test in business: Implications for assessment. *The Journal of Education for Business, 77* (1), 51–56.

Mislevy & Risconscente (2005). *Evidence-Centered Assessment Design: Layers, structures, and terminology. PADI Principled Assessment Designs for Inquiry. Technical Report 9.* Menlo Park, CA: SRI International.

Möller, J. H. (2010). *Kundenorientierung im Hotelfach. Die Entwicklung und Validierung eines Situational Judgment Tests*. Münster: MV.

Möller, J. & Köller, O. (2004). Die Genese akademischer Selbstkonzepte. *Psychologische Rundschau, 55* (1), 19–27.

Moosbrugger, H. & Kelava, A. (2012a). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In H. Moobrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 8–26). Berlin: Springer.

Moosbrugger, H. & Kelava, A. (2012b). Klassische Testtheorie (KTT). In H. Moobrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 104–117). Berlin: Springer.

Motowidlo, S. J. & Beier, M. E. (2010). Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgment test. *Journal of Applied Psychology, 95* (2), 321.

Motowidlo, S. J., Crook, A. E., Kell, H. J. & Naemi, B. (2009). Measuring procedural knowledge more simply with a single-response situational judgment test. *Journal of Business and Psychology, 24* (3), 281–288.

Motowidlo, S. J., Dunnette, M. & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75* (6), 640–647.

Motowidlo, S., Hooper, A. C. & Jackson, H. L. (2006a). A theoretical basis for situational judgment tests. In J. A. Weekley & R. E. Ployhart (Hrsg.), *Situational judgment tests. Theory, measurement, and application* (S. 57–82). Mahwah, N.J.: Lawrence Erlbaum Associates, Publishers.

Motowidlo, S. J., Hooper, A. C. & Jackson, H. L. (2006b). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology, 91* (4), 749.

Müller, G. F. (2003). *Fragebogen zur Diagnose unternehmerischer Potenziale: (F-DUP)*. Zugriff am 12.09.2013 http://www.gruendungszuschuss.de/fileadmin/media/downloads/sonstige_pdfs/F-DUP_2004_TV.pdf.

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

Murphy, K. R., & Davidshofer, C. O. (1991). *Psychological testing*: principles and applications Upper Saddle River, N. J.: Prentice-Hall.

Muthén, L. & Muthén B.O. (2010). *Mplus Users's Guide. Sixth Edition*. Los Angeles: Muthén & Muthén.

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

Nunnally, J. (1967). *Psychometric Theory*. New York: McGraw-Hill.

OECD (Hrsg.). (2011). *Tuning-AHELO conceptual framework of expected and desired learning outcomes in economics.* OECD Education Working Papers: 59. Zugriff am 12.09.2013 http://dx.doi.org/10.1787/5 kghtchwb3nn-en.

Osterlind, S. J. (1989). *Test item bias* (4. Aufl.). Newbury Park: Sage Publications.

Osterlind, S. J. & Everson, H. T. (2009). *Differential item functioning* (2, Aufl.). Thousand Oaks, CA: Sage Publications.

Parshall, C. G., Spray, J., Kalohn, J. & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer.

Patterson, F., Ashworth, V., Mehra, S. & Falcon, H. (2012). Could situational judgement tests be used for selection into dental foundation training? *British Dental Journal, 213* (1), 23–26.

Päzold, G. & Wahle, M. (2003). Das duale System der Berufsbildung zwischen Erosionstendenzen und Modernisierungschancen. In A. Bredow, R. Dobischat & J. Rottmann (Hrsg.), *Berufs- und Wirtschaftspädagogik von A-Z. Grundlagen, Kernfragen und Perspektiven (Band 4)* (S. 471–489). Hohengehren: Schneider-Verlag.

Pellegrino, J. W. (2006). *Rethinking and redesigning curriculum, instruction and assessment: What contemporary research and theory suggests.* Washington, DC: National Center on Education and the Economy for the New Commission on the Skills of the American Workforce.

Pellegrino, J. W., Chudowsky, N. & Glaser, R. (2001). Knowing what students know: The science and design of educational assessment. Washington, DC: The National Academies Press.

Penfield, R. D. & Lam, T. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19* (3), 5–15.

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of consumer research, 21* (2), 381–391.

Pomplun, M., Frey, S., & Becker, D. F. (2002). The score equivalence of paper-and-pencil and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement*, *62* (2), 337–354.

Popper, K. R. (2009). *Vermutungen und Widerlegungen: das Wachstum der wissenschaftlichen Erkenntnis.* Tübingen: Mohr Siebeck.

Pöttker, J. (2009). *Personaldiagnostik in Ausbildungsberufen des Handwerks. Entwicklung eines Situational-Judgment-Tests und Überprüfung verschiedener Scoringmethoden* (1., Aufl.). Münster: Monsenstein und Vannerdat.

Preiss, P. & Klauser, F. (1992). Lehrerbildung und Problemlöseforschung mit einem LAN-Unternehmensplanspiel (Jeansfabrik). In K. Dette, D. Haupt & C. Polze (Hrsg.), *Multimedia und Computeranwendungen in der Lehre* (S. 495–502). Berlin: Springer.

Preuß, A. & Wehrmaker, M. (2008). Verfälschungssicher Online-Fähigkeitstests. In W. Sarges & D. Scheffer (Hrsg.), *Innovative Ansätze für die Eignungsdiagnostik* (S. 229–238). Göttingen: Hogrefe.

Raju, P., Lonial, S. & Mangold, G. M. (1995). Differential effects of subjective knowledge, objective knowledge, and usage experience on decision making: An exploratory investigation. *Journal of Consumer Psychology, 4* (2), 153–180.

Ramm, M. & Multrus, F. (2006). *Das Studium der Betriebswirtschaftslehre. Eine Fachmonographie aus studentischer Sicht*. Bonn, Berlin: Bundesministerium für Bildung und Forschung. Zugriff am 12.09.2013 http://www.bmbf.de/pub/studium_der_betriebswirtschaftlehre.pdf

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* Oxford, England: Nielsen & Lydiche.

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Zugriff am 12.09.2013 http://www.R-project.org/.

Renkl, A. (2009). Lehren und Lernen. In R. Tippelt & B. Schmidt (Hrsg.), *Handbuch Bildungsforschung* (S. 737–751). Berlin: Springer.

Renkl, A., Gruber, H., Mandl, H. & Hinkhofer, L. (1994). Hilft Wissen bei der Identifikation und Kontrolle eines komplexen ökonomischen Systems? *Unterrichtswissenschaft, 22* (3)*,* 195–202.

Riese, J. & Reinhold, P. (2012). Die professionelle Kompetenz angehender Physiklehrkräfte in verschiedenen Ausbildungsformen. *Zeitschrift für Erziehungswissenschaft, 15* (1), 111–143.

Rigdon, E. E. (1996). CFI versus RMSEA: A comparison of two fit indexes for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 3* (4), 369–379.

Rigney, T. (2002). A study of the relationship between entry qualifications and achievement of third level business studies students. *Irish Journal of Management, 23* (2), 117–139.

Rosendahl, J. & Straka, G. A. (2011). Kompetenzmodellierungen zur wirtschaftlichen Fachkompetenz angehender Bankkaufleute. *Zeitschrift für Berufs- und Wirtschaftspädagogik, 107* (2), 190–217.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14* (3), 271–282.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (2. Aufl.). Bern: Huber.

Rost, J. (2008). Zur Messung von Kompetenzen einer Bildung für nachhaltige Entwicklung. In I. Bormann & de Haan Gerhard (Hrsg.), *Kompetenzen der Bildung für nachhaltige Entwicklung. Operationalisierung, Messung, Rahmenbedingungen, Befunde* (S. 61–73). Wiesbaden: VS Verlag für Sozialwissenschaften.

Roth, H. (1971). *Entwicklung und Erziehung: Grundlagen einer Entwicklungspädagogik.* Braunschweig: Schroedel.

Rumsey, D. (2008). *Weiterführende Statistik für Dummies.* (1. Aufl.). Weinheim: Wiley-VCH.

Ryle, G. (1949). *The concept of mind*. Chicago: University of Chicago Press.

Sailer, M. (2009). *Anforderungsprofile und akademischer Arbeitsmarkt. Die Stellenanzeigenanalyse als Methode der empirischen Bildungs- und Qualifikationsforschung* (Erziehung & Bildung – Eichstätter Studien, Bd. 3). Münster: Waxmann.

Schäfer, B. (1999). Entwicklung von Handlungskompetenz zur Gestaltung beruflicher Handlungsfelder — Eine didaktische Reflexion des Lernfeld-Konzeptes. In P. Sloane, R. Bader & G. Straka (Hrsg.), *Lehren und Lernen in der beruflichen Ausbildung. Ergebnisse der Herbsttagung 1998* (S. 163–174). Wiesbaden: Springer Fachmedien.

Saunders, P. (1991). The third edition of the test of unterstanding in college economics. *Journal of Economic Education, 81* (2)*,* 255–272.

Schaeper, H. & Wolter, S. (2008). Hochschule und Arbeitsmarkt im Bologna-Prozess. *Zeitschrift für Erziehungswissenschaft, 11* (4), 607–625.

Schermelleh, K. & Schweizer, K. (2012). Multitrait-Multimethod-Analysen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 346–361). Berlin: Springer.

Schermelleh-Engel, K. & Werner, C. S. (2012). Methoden der Reliabilitätsbestimmung. H. Moosbrugger & A. Kelava (Hrsg.),*Testtheorie und Fragebogenkonstruktion* (S. 119–141). Berlin: Springer.

Schiefele, U., Krapp A. Wild K. P. & Winteler, A. (1993). Der „Fragebogen zum Studieninteresse" (FSI). *Diagnostica, 39* (4)*,* 335–351.

Schiefele, U. & Moschner, B. (1997). *Selbstkonzept, Lernmotivation, Lernstrategien, epistemologische Überzeugungen, Instruktionsqualität und Studienleistung: längsschnittliche Verläufe und kausale Zusammenhänge. Antrag an die Deutsche Forschungsgemeinschaft*. Bielefeld: Universität, Abteilung für Psychologie.

Schiefele, U., Streblow, L., Ermgassen, U. & Moschner, B. (2003). Lernmotivation und Lernstrategien als Bedingungen der Studienleistung. *Zeitschrift für Pädagogische Psychologie, 17* (3/4), 185–198.

Schmidt-Atzert, L. (2008). Tests und Tools. *Zeitschrift für Personalpsychologie, 7* (3), 141–143.

Schmidt, F. L. & Hunter, J. E. (1993). Tacit knowledge, practical intelligence, general mental ability, and job knowledge*, 2* (1), 8–9.

Schmitt, N. & Chan, D. (2006). Situational Judgment Tests: Method or Construct? In J. A. Weekley & R. E. Ployhart (Hrsg.), *Situational judgment tests. Theory, measurement, and application* (S. 135–155). Mahwah, N.J.: Lawrence Erlbaum Associates.

Schneider, M. (2005). *Konzeptuelles und prozedurales Wissen als latente Variablen: Ihre Interaktion beim Lernen mit Dezimalbrüchen*. Dissertation, Technische Universität Berlin.

Schöps, K., Senkbeil, M. & Schütte, K. (2009). Umweltbezogene Einstellungen von Jugendlichen in Deutschland—Ergebnisse aus PISA 2006. In *Vertiefende Analysen zu PISA 2006* (S. 53–77). Berlin: Springer.

Schubert, S., Ortwein, H., Dumitsch, A., Schwantes, U., Wilhelm, O. & Kiessling, C. (2008). A situational judgement test of professional behaviour: development and validation. *Medical Teacher, 30* (5), 528–533.

Schumann, S. (2013, Juli). *Oekoma – Ökonomische Kompetenzen von Berufsmaturanden in der Schweiz.* Gastvortrag an der Universität Göttingen.

Schumann, S. & Eberle, F. (2011). Bedeutung und Verwendung schwierigkeitsbestimmender Aufgabenmerkmale für die Erfassung ökonomischer und beruflicher Kompetenzen. In U. Faßhauer, B. Fürstenau & E. Wuttke (Hrsg.), *Grundlagenforschung zum dualen System und Kompetenzentwicklung in der Lehrerbildung* (S. 77–89). Opladen: 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, Hrsg.). Zürich: Institut für Gymnasial- und Berufspädagogik.

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 (Hrsg.), *Lehr-Lernforschung und Professionalisierung. Perspektiven der Berufsbildungsforschung,* Schriftenreihe der Sektion Berufs- und Wirtschaftspädagogik der Deutschen Gesellschaft für Erziehungswissenschaft (DGfE), (S. 51–64). Leverkusen: Barbara Budrich.

Seeber, S. (2008). Ansätze zur Modellierung beruflicher Fachkompetenz in kaufmännischen Ausbildungsberufen. *Zeitschrift für Berufs- und Wirtschaftspädagogik, 104* (1), 74–97.

Seeber, S., Nickolaus, R., Winther, E., Achtenhagen, F., Breuer, K., Frank, I. et al. (2010). Kompetenzdiagnostik in der Berufsbildung. Begründung und Ausgestaltung eines Forschungsprogramms. *Beiheft zur Berufsbildung in Wissenschaft und Praxis, 1,* 1–15.

Segerer, R., Marx, A. & Marx, P. (2012). Unlösbare Items im KFT 4–12+ R. *Diagnostica, 58* (1), 45–50.

Settlage, D. M. & Settlage, L. A. (2011). A statistical framework for assessment using the ETS Major Field Test in Business. *Journal of Education for Business, 86* (5), 274–278.

Soper, J. (1979). *The test of economic literacy: Discussion guide and rationale*. New York: Joint Council on Economic Education.

Sparfeldt, J. R., Buch, S. R., Rost, D. H. & Lehmann, G. (2008). Akkuratesse selbstberichteter Zensuren. *Psychologie in Erziehung und Unterricht, 55* (1), 68–75.

Statistisches Bundesamt. (2013). Tabellen Studierende Betriebswirtschaftslehre. Zugriff am 18.09.2013 https://www.destatis.de/DE/ZahlenFakten/Indikatoren/LangeReihen/Bildung/lrbil02.html

Stemler, S. E. & Sternberg, R. J. (2006). Using situational judgment tests to measure practical intelligence. In J. A. Weekley & R. E. Ployhart (Hrsg.), *Situational judgment tests. Theory, measurement, and application* (S. 107–134). Mahwah, N.J.: Lawrence Erlbaum Associates.

Sternberg, R. J. (Hrsg.). (1982). *Handbook of human intelligence*. Cambridge: University Press.

Sternberg, R. J., Wagner, R. K., Williams, W. M. & Horvath, J. A. (1995). Testing common sense. *American Psychologist, 50* (11), 912–922.

Stevens, M. J. & Campion, M. A. (1994). The knowledge, skill, and ability requirements for teamwork. Implications for human resource management. *Journal of Management, 20* (2), 503–530.

Stratmann, J., Preussler, A. & Kerres, M. (2009). Lernerfolg und Kompetenz: Didaktische Potenziale der Portfolio-Methode im Hochschulstudium. *Zeitschrift für Hochschulentwicklung, 4,* (1)*,* 90–103.

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

Strobl, C. (2012). *Das Rasch-Modell. Eine verständliche Einführung für Studium und Praxis*. München: Rainer Hampp Verlag.

Strobl, C., Kopf, J. & Zeileis, A. (2013). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*. (Accepted)

Süß, H.-M. (1996). *Intelligenz, Wissen und Problemlösen*. Göttingen: Hogrefe.

Süß, H.-M. (1999). Intelligenz und komplexes Problemlösen. *Psychologische Rundschau, 50* (4), 220–228.

Teichler, U. (Hrsg.). (2005). *Hochschullandschaft im Wandel.* Weinheim: Beltz.

Teichler, U. (2007). Studium und Berufschancen: Was macht den Unterschied aus. *Beiträge zur Hochschulforschung, 29* (4), 10–31.

Tramm, T. & Seeber, S. (2006). Überlegungen und Analysen zur Spezifizität kaufmännischer Kompetenz. In G. Minnameier & E. Wuttke (Hrsg.), *Berufs- und wirtschaftspädagogische Grundlagenforschung. Lehr-Lern-Prozesse und Kompetenzdiagnostik Festschrift für Klaus Beck* (S. 273–288)*.* Frankfurt am Main: Peter Lang.

Tremblay, K., Lalancette, D. & Roseveare, D. (2012). *Assessment of Higher Education Learning Outcomes. Feasibility Study Report Volume 1 – Design and Implementation* (OECD, Hrsg.). Zugriff am 12.09.2013 http://www.oecd.org/edu/skills-beyond-school/AHELOFSReportVolume1.pdf

Universität Göttingen, Wirtschaftswissenschaftliche Fakultät. (Oktober 2010). *Modulverzeichnis für die Bachelor-Studiengänge der Wirtschaftswissenschaftlichen Fakultät* (Präsident der Georg-August-Universität Göttingen, Hrsg.) (Nr. 27), Göttingen. Zugriff am 12.09.2013 http://www.uni-goettingen.de/de/184937.html.

Universität Göttingen, Wirtschaftswissenschaftliche Fakultät. (Oktober 2013). *Abschluss: „Bachelor of Arts B.A."* Zugriff am 09.10.2013 http://www.uni-goettingen.de/de/81147.html

Universität Hamburg. (Juni 2010). *Neufassung der Fachspezifischen Bestimmungen für den Bachelorstudiengang Betriebswirtschaftslehre (B. Sc.) im Fachbereich Betriebswirtschaftslehre der Fakultät Wirtschafts- und Sozialwissenschaften der Universität Hamburg* (Der Präsident der Universität Hamburg Referat 31 – Qualität und Recht, Hrsg.) (Nr. 31), Hamburg. Zugriff am 12.09.2013 http://www.uni-hamburg.de/campuscenter/studienorganisation/ordnungen-satzungen/pruefungs-studienordnungen/wirtschafts-und-sozialwissenschaften/FSB_FakWiSo_BSc_BWL_Neufassung_20090415.pdf

Universität Mannheim. (Juni 2009). *Bachelor of Science (B.Sc.) „Betriebswirtschaftslehre". Modulkatalog* (Universität Mannheim, Hrsg.), Mannheim. Zugriff am 12.09.2013 http://www.bwl.uni-mannheim.de/fileadmin/files/dekanat/files/pa/modulkatalog/Modulkatalog_BScBWL_v20.pdf

Urban, D. (2004). *Neue Methoden der Längsschnittanalyse.* Münster: LIT.

Wagenknecht, H. (1980). *Lexikon der Psychologie*. Freiburg: Herder.

Weekley, J. A. & Ployhart, R. E. (Hrsg.). (2006). *Situational judgment tests. Theory, measurement, and application*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Weekley, J. A., Ployhart, R. E. & Holtz, B. C. (2006). On the development of situational judgment tests: Issues in item development, scaling, and scoring. In J. A. Weekley & R. E. Ployhart (Hrsg.), *Situational judgment tests. Theory, measurement, and application* (S. 157–182). Mahwah, N.J.: Lawrence Erlbaum Associates.

Weidle, R. & Wagner, A. C. (1994). Die Methode des Lauten Denkens. In G. L. Huber & H. Mandl (Hrsg.), *Verbale Daten: Eine Einführung in die Grundlagen und Methoden der Erhebung und Auswertung* (S. 81–103). Weinheim: Beltz.

Weinert, F. E. (2002). Vergleichende Leistungsmessung in Schulen-eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 17–23). Weinheim: Beltz.

Weishaupt, H. (2010). *Bildung in Deutschland. Ein indikatorengestützter Bericht mit einer Analyse zur Zukunft des Bildungswesens im demografischen Wandel*. Bielefeld: Bertelsmann.

Westers, P. & Kelderman, H. (1992). Examining differential item functioning due to item difficulty and alternative attractiveness. *Psychometrika, 57* (1), 107–118.

Wiepcke, C. (2009). Employability in the Bologna process: an area of tension between society, businesses and students. *The international journal of learning, 16* (4), 435–445.

Wild, K.-P., Krapp, A., Schiefele, U., Lewalter, D. & Schreyer, I. (1995). *Dokumentation und Analyse der Fragebogenverfahren und Tests. (Berichte aus dem DFG-Projekt «Bedingungen und Auswirkungen berufsspezifischer Lernmotivation)* (Nr. 2). München: Universität der Bundeswehr, Institut für Erziehungswissenschaft und Psychologie.

Wild, K.-P. & Schiefele, U. (1994). Lernstrategien im Studium. Ergebnisse zur Faktorenstruktur und Reliabilität eines neuen Fragebogens. *Zeitschrift für Differentielle und Diagnostische Psychologie* (15), 185–200.

Williams, M. L., Waldauer, C. & Duggal, V. G. (1992). Gender differences in economic knowledge: An extension of the analysis. *Journal of Economic Education, 23* (3)*,* 219–231.

Wilson, M. (2005). Constructing measures: an item response modeling approach. New York: Psychology Press.

Windzio, M., & Grotheer, M. (2002). Bleiben die Erfolgreichen übrig? Die Kombination von Sequenzmusteranalysen und log-linearen Pfadmodellen bei der Analyse des Zusammenhangs von Berufserfolg und Panelmortalität. *Zeitschrift für Soziologie*, *31* (6), 514–528.

Winther, E. (2010). *Kompetenzmessung in der beruflichen Bildung*. Bielefeld: Bertelsmann.

Winther, E. (2011). Das ist doch nicht fair! – Mehrdimensionalität und Testfairness in kaufmännischen Assessments. *Zeitschrift für Berufs- und Wirtschaftspädagogik, 107* (2), 218–238.

Winther, E. & Achtenhagen, F. (2008). Kompetenzstrukturmodell für die kaufmännische Bildung. Adaptierbare Forschungslinien und theoretische Ausgestaltung. *Zeitschrift für Berufs- und Wirtschaftspädagogik, 104* (4), 511–538.

Winther, E. & Achtenhagen, F. (2009). Skalen und Stufen kaufmännischer Kompetenz. *Zeitschrift für Berufs- und Wirtschaftspädagogik, 105* (4), 521–556.

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

Witt R. (2006). Kompetenzstufenmodell zur Messung ökonomischer Kompetenz. In G. Minnameier & E. Wuttke (Hrsg.), *Berufs- und Wirtschaftspädagogische Grundlagenforschung. Lehr-Lern-Prozesse und Kompetenzdiagnostik* (S. 407–419). Frankfurt am Main: Peter Lang.

Wittmann, W., Süß, H.-M. & Oberauer, K. (1996). *Determinanten komplexen Problemlösens* (Bericht 9). Zugriff am 12.09.2013 http://www.psychologie.uni-mannheim.de/psycho2_alt/publi/ps/ber09.pdf

Wolter, A., & Banscherus, U. (2012). Praxisbezug und Beschäftigungsfähigkeit im Bologna-Prozess–"A never ending story "?. In W. Schubarth, K. Speck, A. Seidel, C. Gottmann, C. Kamm & M. Krohn (Hrsg.), *Studium nach Bologna: Praxisbezüge stärken?! Praktika als Brücke zwischen Hochschule und Arbeitsmarkt* (S. 21–36). Springer Fachmedien Wiesbaden.

Wu, M. L., Adams, R. J. & Wilson, M. (1998). *ACER ConQuest: Generalised Item Reponse Modelling Software.* Camberwell, Victoria: ACER press.

Zeegers, P. (2004). Student learning in higher education: a path analysis of academic achievement in science. *Higher Education Research & Development, 23* (1), 35–56.

Zinger, B. (2012). *Das Hochschulstudium nach Bologna: zwischen Strukturreform und didaktischer Neuausrichtung* (Vol. 11). Doktorarbeit an der Universität Kassel.

Zlatkin-Troitschanskaia, O. (2013, März). *Innovativer Lehr-Lernortverbund (IL-LEV) in der akademischen Hochschulausbildung.* Vortrag auf der Frühjahrstagung der Sektion Berufs- und Wirtschaftspädagogik in Kassel.

Zlatkin-Troitschanskaia, O., & Breuer, K. (2010). Wirtschaftspädagogisches Studium an der Uni Mainz. Ein polyvalenter Bachelor und Master of Science in Wirtschaftspädagogik. *Erziehungswissenschaft*, *21*(40), 125–133.

Zlatkin-Troitschanskaia, O., Förster, M. & Happ, R. (2012). Bologna-Reform – Ergebnisse aus einer vergleichenden empirischen Studie zwischen den auslaufenden Diplom- und den neuen Bachelor-/Masterstudiengängen. *Zeitschrift für Berufs- und Wirtschaftspädagogik, 108* (3), 420–437.

Zlatkin-Troitschanskaia, O. & Kuhn, C. (2010). Messung akademisch vermittelter Fertigkeiten und Kenntnisse von Studierenden bzw. Hochschulabsolventen: Analyse zum Forschungsstand. *Arbeitspapiere Wirtschaftspädagogik (Nr. 56)*. Universität Mainz: Lehrstuhl für Wirtschaftspädagogik. Zugriff am 19.09.2013 http://www.wipaed.uni-mainz.de/ls/382.php

Zöllösi-Janze, M. (2004). Wissensgesellschaft in Deutschland: Überlegungen zur Neubestimmung der deutschen Zeitgeschichte über Verwissenschaftlichungsprozesse. *Geschichte und Gesellschaft, 30* (2), 277–313.

Anhang A Synoptische Darstellung der Modulbeschreibungen

Tab. A-1: Ziele und Inhalte der Pflichtmodule „Finanz- und Rechnungswesen" in den Studiengängen Betriebswirtschaftslehre und Wirtschaftspädagogik

Inhaltsbereich: Finanz- und Rechnungswesen		
Universität Göttingen	**Universität Hamburg**	**Universität Mannheim**
Einführung die Finanzwirtschaft (Universität Göttingen, 2010, S. 2218)	**Finanzierung** (Universität Hamburg, 2010, S. 23)	**Finanzwirtschaft I** (Universität Mannheim, 2009, S. 11)
• Grundbegriffe der betrieblichen Finanzwirtschaft • Grundlegende finanzwirtschaftliche Fragen und erste Lösungsansätze • Neben einem Verständnis des finanzwirtschaftlichen Denkens und der ökonomischen Grundlagen des Faches, die für das weitere Studium benötigt werden, soll auch ein praktisches Wissen, insbesondere hinsichtlich der Methoden der Investitionsrechnung, erworben werden.	• Informationseffizienz von Finanzmärkten und Nettobarwert von Finanzierungen • Finanzierungsinstrumente und ihre Begebung • Ausschüttungen aus empirischer und normativer Sicht • Verschuldung, Risiko und Kapitalkosten • Kapitalstrukturtheorien • Finanzierungsinstrumente und Optionen • Zeitstruktur der Zinssätze • Optionen und die Bewertung von Forderungs- und Beteiligungstiteln eines Unternehmens	• Moderne Finanzierungstheorie und -praxis und wesentliche Institutionen und analytischen Konzepte, die zum Verständnis moderner Kapitalmärkte erforderlich sind • Barwertmethoden • Verfahren zur Kapitalbudgetierung • Bewertung von Aktien und Anleihen • Theorie der Portfolioallokation • das Capital Asset Pricing Modell (CAPM)
Interne Unternehmensrechnung (Universität Göttingen, 2010, S. 2165)	**Kosten- und Leistungsrechnung** (Universität Hamburg, 2010, S. 16)	**Internes Rechnungswesen I** (Universität Mannheim, 2009, S. 25)
• Ziele der internen Unternehmensrechnung • Informationsinteressen der planenden, leitenden und kontrollierenden Personen in Unternehmen • Methoden und Verfahren der internen Unternehmensrechnung	• Einführung in die interne Unternehmensrechnung • Grundlagen der Kostentheorie • Instrumentarium der Kosten- und Leistungsrechnung • Kostenrechnung • Kostenstellenrechnung	• kostentheoretische Grundlagen • Kosten-, Erlös- und Ergebnisplanung • Kosten-, Erlös- und Ergebniskontrolle • Kostenmanagement • Kosten- und Erlösrechnungssysteme

(Fortsetzung Tab. A-1)

Inhaltsbereich: Finanz- und Rechnungswesen		
Universität Göttingen	**Universität Hamburg**	**Universität Mannheim**
• Spezielle Entscheidungsprobleme	• Selbstkostenrechnung • Kurzfristige Erfolgsrechnung • Systeme der Kosten- und Leistungsrechnung • Normalkostenrechnung • Plankostenrechnung • Teilkosten- und Deckungsbeitragsrechnung • Kontroll- und Entscheidungsrechnungen auf Basis von Kosten und Erlöse	

Tab. A-2: Ziele und Inhalte des Pflichtmoduls „Unternehmensführung" in den Studiengängen Betriebswirtschaftslehre und Wirtschaftspädagogik

Inhaltsbereich: Unternehmensführung		
Universität Göttingen	**Universität Hamburg**	**Universität Mannheim**
Unternehmensführung und Organisation (Universität Göttingen, 2010, S. 2166)	**Unternehmensführung 1: Grundlagen des Managements** (Universität Hamburg, 2010, S. 14)	**Management I: Strategic & International Management** (Universität Mannheim, 2009, S. 6)
• organisationstheoretische Grundlagen der Unternehmensführung • Prozesse strategischer Planung und Entscheidung von Unternehmen	• Entstehung der Managementfunktion der Organisation • Grundkenntnisse organisatorischer Gestaltung und Steuerung • Führungsansätze und Führungsrichtungen • Grundlagen verhaltensorientierten Managements • Entwicklung der Management Disziplin • Die Rollen von Manager/innen • Die Managementfunktionen • Planung, Organisation, Führung, Controlling	• Managementansätze, Führungsziele, -system, Führungsprozess und -kompetenz • Controlling mit den Phasen Planung, Steuerung und Überwachung • Konzepte des strategischen und internationalen Managements • Zusammenhänge, Probleme und Lösungen des Managements • Aufgaben, Einsatzfelder und Instrumente eines Controllers • Grundlagen der strategischen Unternehmensführung

(Fortsetzung Tab. A-2)

Inhaltsbereich: Unternehmensführung		
Universität Göttingen	**Universität Hamburg**	**Universität Mannheim**
		• Konzepte für das strategische und internationale Management

Tab. A-3: Ziele und Inhalte des Pflichtmoduls „Marketing" in den Studiengängen Betriebswirtschaftslehre und Wirtschaftspädagogik

Inhaltbereich: Marketing		
Universität Göttingen	**Universität Hamburg**	**Universität Mannheim**
Beschaffung und Absatz (Universität Göttingen, 2010, S. 2166)	**Einführung ins Marketing** (Universität Hamburg, 2010, S. 25)	**Marketing I** (Universität Mannheim, 2009, S. 8)
• Grundkenntnisse des Beschaffungs- und Absatzkanals • Analysieren von Waren- und Informationsströmen • Fragen sowie Methoden, mit denen Waren- und Informationsströme analysiert werden können • Absatzpolitische Instrumente • Grundlagen des Konsumentenverhaltens • Grundlagen Marktforschung • Rahmenbedingungen und Entscheidungen bei der Ausgestaltung der Absatzpolitik • Interdependenzen zu • den Entscheidungen im Beschaffungsbereich • Methoden, mit denen • die Entscheidungsfindung unterstützt werden kann	• Was ist Absatz/Marketing? • Verständnis für den Kunden entwickeln • Märkte analysieren • Ziele und Strategien festlegen • Marketing-Mix-Maßnahmen gestalten • Markenoptionen gestalten • Produkte und Services gestalten • Kommunikation managen • Preise bilden • Distributionsentscheidungen treffen • Marketing-Mix optimieren • Ziele, Strategien und Maßnahmen kontrollieren	• Grundlagen des Marketing • Instrumente des Marketing-Mix und Kundenbeziehungs-management • Bezug zu relevanten Theorien und Marktforschungsmethoden • Strategisches Marketing • Institutionelle Besonderheiten des Marketings • Dienstleistungs-, Business-to-Business Marketing • Internationales Marketing

Tab. A-4: Ziele und Inhalte des Pflichtmoduls „Produktion" in den Studiengängen Betriebswirtschaftslehre und Wirtschaftspädagogik

Inhaltsbereich: Produktion		
Universität Göttingen	**Universität Hamburg**	**Universität Mannheim**
Produktion und Logistik (Universität Göttingen, 2010, S. 2167)	**Produktion** (Universität Hamburg, 2010, S. 24)	**Produktion I** (Universität Mannheim, 2009, S. 26 f)
• Produktionsprozesse sowie die enge Verzahnung von Produktion und Logistik • Betriebliche Abläufe mit Hilfe • geeigneter Planungsmodelle effizient zu gestalten • Produktionsfunktionen • lineare Optimierung • Produktions- und Kostentheorie • Produktionsprogrammplanung mit linearer Optimierung • Bereitstellungsplanung/ Beschaffungslogistik • Durchführungsplanung • Produktionslogistik • Distributionslogistik • Simulation und Visualisierung von Produktions- u. Logistikprozessen	• Einführung in die Produktions- und Kostentheorie Produktionstypen • Grundlagen der strategischen, taktischen und operativen Produktionswirtschaft • Ausgewählte Entscheidungsmodelle in der Produktion • Aufbau und Inhalt von Standardsoftware (z. B. Produktionsplanungs- und -steuerungssysteme, Advanced Planning Systeme)	• Grundzüge der Produktions- und Kostentheorie • Konzepte der Fertigungsstrategien • Ansätze des Produktionsmanagements • Produktionsplanung und -steuerung

Anhang B Pilotierungsergebnisse

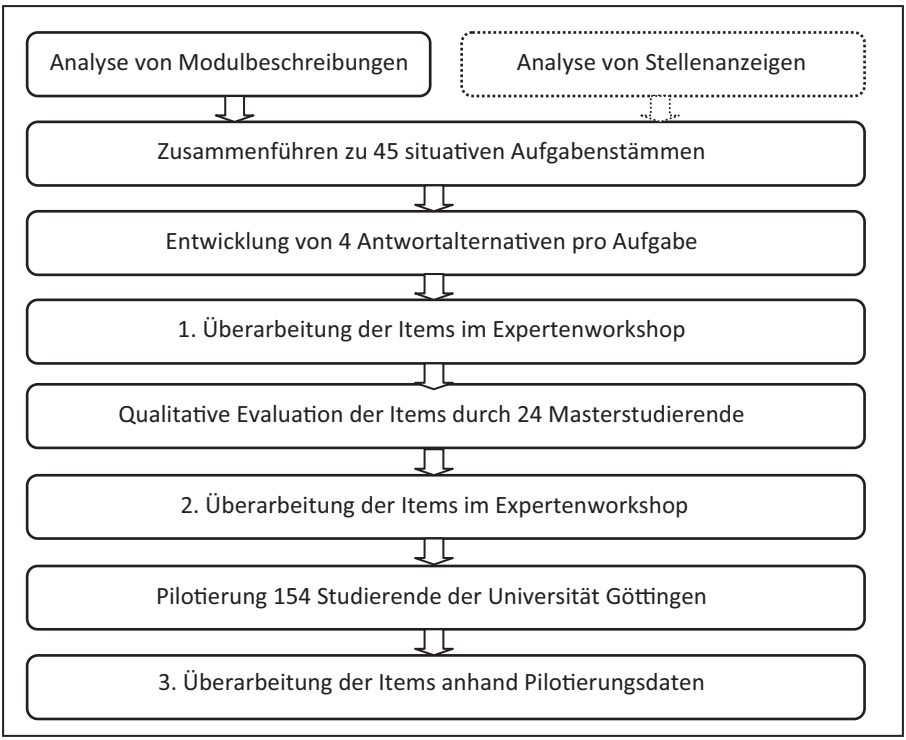


Abb. B-1: Grafische Darstellung des Prozesses der Itementwicklung

Tab. B-1: Zusammenfassende Darstellung der Pilotierungsergebnisse (problematische Kennwerte sind fett gedruckt; Items, die für die Haupterhebung selektiert wurden sind grau unterlegt)

Modulbezug	Item Nr.	Inhaltliche Anforderung	Schwierigkeit	Trennschärfe	WMNS Q[1)]	Gender DIF weiblich	Neue Item Nr.[2)]
Unternehmensführung	1.1	Laterale Kommunikation umsetzten	1.354	**0.02**	1.01	**0.530 (0.221)**	X
	1.2	Konsequenzen von Dezentralisierungsprozessen kommunizieren	–0.506	0.42	0.97	0.281 (0.174)	1
	1.3	Planungstechniken in der strategischen Projektplanung anwenden	0.326	**0.19**	1.01	0.206 (0.170)	X[3)]
	1.4	Hauptversammlung organisieren	–1.130	**0.02**	1.04	–0.313 (0.186)	X
	1.5	BCG Matrix in Produktpräsentation nutzen	0.867	0.21	0.99	0.153 (0.183)	2
	1.6	SWOT-Analyse durchführen	–1.130	**0.09**	1.02	**–0.452 (0.185)**	X
	1.7	Leitbild verfassen	–0.969	**0.19**	1.01	–0.080 (0.182)	3
	1.8	Geeignete Lohnart bestimmen	0.055	**0.17**	1.02	–0.084 (0.167)	4
	1.9	Preis bei Merger erhöhen	–1.533	**0.17**	1.00	–0.144 (0.204)	5
	1.10	Outsourcing Optionen beurteilen	–0.907	0.20	1.00	–0.159 (0.177)	X
	1.11	Geeignete Leitungskonfiguration bestimmen	0.162	0.22	1.00	–0.047 (0.168)	6
Finanz- und Rechnungswesen	2.1	Deckungsbeitrag errechnen	–1.452	0.23	0.99	**0.814 (0.258)**	7
	2.2	Methode zur Darstellung von Ergebnissen unter verschiedenen Bedingungsfaktoren aufzeigen	–0.078	0.27	1.00	0.098 (0.167)	8
	2.3	Erstellen eines Berichtsformulars im Vertrieb	–0.847	0.22	1.00	0.062 (0.178)	9
	2.4	Soll-Ist Vergleich in der Produktion	0.440	0.20	1.00	–0.132 (0.172)	10
	2.5	Periodenerfolgsrechnung im internationalen Unternehmen auswählen	–0.024	**0.04**	1.03	–0.335 (0.168)	X

(Fortsetzung Tab. B-1)

Modulbezug	Item Nr.	Inhaltliche Anforderung	Schwierigkeit	Trennschärfe	WMNS Q[1)]	Gender DIF weiblich	Neue Item Nr.[2)]
	2.6	Outsourcingoptionen mit Kriterien bewerten	–0.907	**0.17**	1.00	**–0.415 (0.179)**	11
	2.7	Unternehmenskennzahlen bewerten	0.586	0.29	0.99	–0.157 (0.177)	12
	2.8	F & E Quote bewerten	0.646	**0.17**	1.00	–0.129 (0.178)	13
	2.9	Variable Stückkosten berechnen	2.954	**–0.09**	1.01	0.048 (0.357)	X
Produktion	3.1	ABC Analyse durchführen	–0.024	0.36	0.98	0.143 (0.167)	14
	3.2	Kriterien für Wahl eines Produktionsverfahrens	–1.231	0.33	0.98	–0.236 (0.191)	15
	3.3	Anpassungsstrategien finden	1.399	**0.05**	1.01	0.204 (0.213)	X
	3.4	Gesamtkostenverlauf prognostizieren	0.082	**0.17**	1.02	–0.318 (0.169)	X
	3.5	Engpässe in Produktion beheben	–0.343	**0.19**	1.00	0.251 (0.170)	X
	3.6	Passende Methode zur Bestimmung von Produktionsmengen	–0.343	**0.14**	0.99	–0.126 (0.167)	X2
	3.7	Collaborative Planning umsetzen	0.272	0.30	0.99	0.207 (0.170)	16
	3.8	Arbeitsunfälle in der Produktion vermeiden	0.707	**0.13**	1.01	**–0.410 (0.186)**	X
	3.9	Auslieferungsformen mit Kriterien bewerten	–0.237	**0.06**	1.02	0.086 (0.167)	X
	3.10	Verbrauchmaterialien bestellen	–1.197	0.41	0.96	0.144 (0.197)	17
	3.11	Produktionssysteme modellhaft abbilden	0.526	**0.16**	1.00	–0.113 (0.174)	X
Absatz & Marketing	4.1	Methode für Umfrage zur Akzeptanz von Onlinekatalogen	0.707	**0.19**	1.00	–0.245 (0.183)	X2
	4.2	Kognitive Dissonanz beim Kunden reduzieren	2.203	**0.10**	1.00	–0.336 (0.283)	X

(Fortsetzung Tab. B-1)

Modulbezug	Item Nr.	Inhaltliche Anforderung	Schwierigkeit	Trennschärfe	WMNS Q[1)]	Gender DIF weiblich	Neue Item Nr.[2)]
	4.3	Kriterien für Marktsegmentierung benennen	0.676	**0.08**	1.01	–0.073 (0.180)	X
	4.4	Passenden Testmarkt wählen	1.190	**0.18**	1.00	–0.010 (0.199)	X
	4.5	Marketing Slogan entwerfen	–1.575	**0.02**	1.02	0.045 (0.212)	X
	4.6	Produktpolitik umsetzen	–1.900	0.21	1.00	0.455 (0.268)	18
	4.7	Studiendesign für Analyse des Kaufverhaltens wählen	0.439	0.27	0.99	–0.013 (0.172)	19
	4.8	Family Branding umsetzen	–0.506	0.31	0.98	0.281 (0.174)	20
	4.9	Marktforschungsauftrag vergeben	–0.818	0.26	1.00	–0.074 (0.175)	21
	4.10	Methode zur Bestimmung der Stärke des Einflusses von Werbebudget auf Absatz benennen	0.867	**0.17**	1.00	–0.051 (0.183)	X
	4.11	Erwartungswert errechnen	1.397	0.21	0.99	0.017 (0.212)	22
	4.12	Preisbereitschaft bestimmen	1.353	**0.08**	1.01	0.022 (0.206)	X
	4.13	Einzelpreis bestimmen	–0.674	0.36	0.97	**0.387 (0.181)**	23
	4.14	Bündelpreis bestimmen	–0.877*	0.35	0.98	0.017*	24

[1)] Alle Abweichungen von 1 n. s.
[2)] vgl. Abbildung 11
[3)] Items wurden aufgrund von qualitativen Rückmeldungen zu mangelnder curricularer Übereinstimmung nicht in die Haupterhebung aufgenommen
* Parameter wurde fixiert

Tab. B-2: Multiple OLS-Regression zur Vorhersage der geschätzten Personenparameter im situativen betriebswirtschaftlichen Wissenstest von N = 126 Studierenden der Universität Göttingen

	Leistung im situativen BWL-Test				
Variable	***ß***	**stand. *ß***	***SE***	***t***	***p***
Konstante	.416				
Absolvierte kaufmännische Berufsausbildung	.272*	.182*	0.063	2.006	.047
Anzahl der besuchen BWL-Pflichtmodule	.168**	.233**	0.063	2.678	.008
Durchschnittliche Note der Hochschulzugangsberechtigung	–.376*	–.207*	0.164	–2.292	.024
Männliches Geschlecht	.332**	.235**	0.122	–2.727	.007
korr. R^2	.108				
F	4.805**				
* p < .05. ** p < .01					

Anhang C Instruktion und Begleitfragebogen zum Test

Tab. C-1: Aufbau des Testhefts und Quellen der Skalen für die Haupterhebung

1. Abschnitt: Wissenstests (situative und nicht-situative Items wurden in randomisierter Reihenfolge dargeboten)		
Itembezeichnung	**Skala**	**Quelle**
24 situative Items	0 falsch 1 richtig	Eigene Entwicklung (vgl. Kapitel 6)
23 nicht-situative Items	0 falsch 1 richtig	Ausgewählte Items aus dem Business Administration Knowledge Test (BAKT) (Bothe, 2003)
2. Abschnitt: Fragen zur Person		
Itembezeichnung	**Skala**	**Quelle**
Geschlecht	0 männlich 1 weiblich	Eigene Entwicklung
Geburtsjahrgang	Freitext	
Deutsch Muttersprache	1 ja 2 nein 3 mehrsprachig	
Höchster Bildungsabschluss beider Elternteile	CASMIN	Lüttinger und König (1988)
Schulform der Hochschulzugangsberechtigung	1 Gymnasium 2 Wirtschaftsgymnasium 3 Gesamtschule 4 Abendgymnasium 5 Fachoberschule 6 nicht wirtschaftliches Fachgymnasium 7 Sonstige	Eigene Entwicklung
Durchschnittliche Note der Hochschulzugangsberechtigung	Freitext	
Kaufmännische Berufsausbildung	1 nein 2 ja, vor Studium 3 ja, nach Studium	

(Fortsetzung Tab. C-1)

Branche des kaufmännischen berufsbildenden Abschlusses	1 Bank 2 Industrie 3 Versicherungen 4 Spedition- und Logistikdienstleistung 5 Groß- und Außenhandel 6 Bürokommunikation 7 anderer Abschluss	
Note des berufsbildenden Abschlusses	Freitext	
3. Fragen zum Studium		
Studienabschnitt	1 Bachelor 2 Mater 3 Diplom 4 Promotion	Eigene Entwicklung
Fachsemester	Freitext	
Universität vorheriger Studienabschlüsse	1 Universität Göttingen[16] 2 eine andere Universität	
Aktueller Studiengang	1 Betriebswirtschaftslehre 2 Wirtschaftspädagogik 3 Wirtschaftsinformatik 4 Volkswirtschaft 5 Anderer Studiengang	
Besuch testrelevanter Module8 • Internes Rechnungswesen • Unternehmensführung und Organisation • Einführung in die Finanzwirtschaft • Produktion und Logistik • Beschaffung und Absatz	1 ja 2 nein 3 ein Modul mit ähnlichen Inhalten	
Note in testrelevanten (siehe oben) Modulen8	Von 1–4 mit einer Dezimalestelle	

16 Ortspezifische Angaben wurden für die Standorte Mannheim und Hamburg entsprechend angepasst.

(Fortsetzung Tab. C-1)

Wochen in kaufmännischen Praktika	Freitext	
Wochen in kaufmännischer Nebentätigkeit	Freitext	
4. Fragen zu individuellen studienbezogenen Faktoren		
Itembezeichnung	**Skala**	**Quelle**
9 Items zu BWL-spezifischem Studieninteresse	1 trifft gar nicht zu 2 trifft sehr begrenzt zu 3 trifft weitgehend zu 4 trifft völlig zu	Fragebogen zum Studieninteresse (FSI) (Schiefele et al., 1993)
4 Items zur studienbezogenen Leistungsmotivation 3 Items zur Wettbewerbsmotivation	von 1 stimmt gar nicht bis 6 stimmt genau	Wild, Krapp, Schiefele, Lewalter und Schreyer (1995), abgeändert nach Hußtegge (2011)
9 Items zum akademischen Selbstkonzept 8 Items zum mathematischen Selbstkonzept	von 1 stimmt gar nicht bis 6 stimmt genau	Verkürzte Skalen aus dem SDQ (March & O' Neill, 1984), abgeändert nach Hußtegge (2011)
32 Items zur Verwendung von Lernstrategien	von 1 sehr selten bis 6 sehr oft	Wild und Schiefele (1994), abgeändert nach Hußtegge (2011)
Bereits an der Befragung teilgenommen?	0 nein 1 ja	Eigene Entwicklung

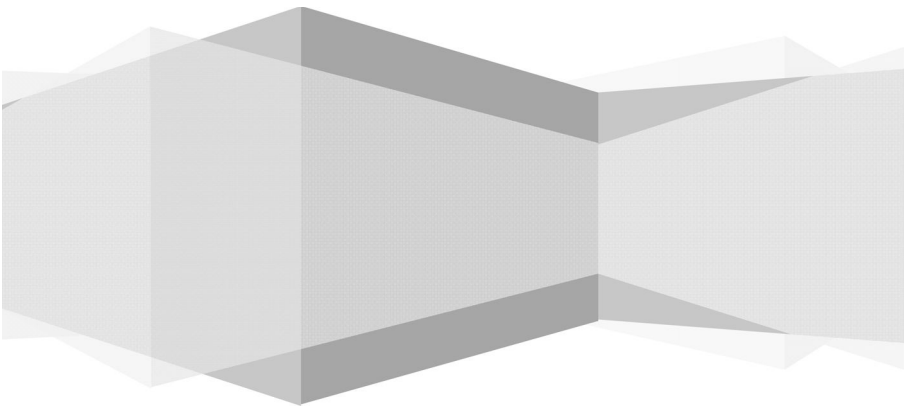
GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Testheft A (oder B)

Studie zu betriebswirtschaftlichem Wissen

Dipl.-Psych. Christine Caroline Jähnig

Georg-August-Universität Göttingen
Professur für Wirtschaftspädagogik und Personalentwicklung
Platz der Göttinger Sieben 5
Tel: 0551/39–44 08
Mail: christine-caroline.jaehnig@wiwi.uni-goettingen.de



GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Liebe Studierende,

vielen Dank, dass Sie sich die Zeit nehmen, an meiner Umfrage teilzunehmen. Zielgruppe der Untersuchung sind Studierende mit grundlegenden Kenntnissen der Betriebswirtschaft. Da Sie ein Bachelor-Studium mit betriebswirtschaftlichen Inhalten absolvieren, sind Ihre Antworten für diese wissenschaftliche Untersuchung äußerst wichtig.

Ziel der Untersuchung ist herauszufinden, wie gut Sie die im Fragebogen gestellten Aufgaben beantworten können und wie die Beantwortung der Aufgaben mit individuellen und institutionellen Faktoren zusammenhängt.

Zum Fragebogen

In dem folgenden Fragebogen werden Ihnen Fragen aus unterschiedlichen Bereichen der Betriebswirtschaft gestellt. Bitte lesen Sie sich die Fragstellung genau durch und wählen Sie aus den möglichen Antwortoptionen diejenige Antwort aus, die Sie am ehesten für richtig halten. Möglicherweise gibt es mehrere Antwortoptionen, die richtig sind oder zumindest richtige Anteile enthalten. **Sie dürfen jedoch pro Aufgabe nur ein Antwortkreuz setzen**. Wenn Sie sich unsicher sind, wählen Sie bitte die Antwortalternative, zu der Sie tendieren. Taschenrechner sind als Hilfsmittel erlaubt, aber nicht unbedingt notwendig.

Nach Abschluss der betriebswirtschaftlichen Aufgaben möchte ich Sie bitten, einige Fragen zu Ihrer Person und Ihren Erfahrungen im Studium zu beantworten.

Die Daten dieser Umfrage werden selbstverständlich anonym und nur zu wissenschaftlichen Zwecken ausgewertet. Wenn Sie jedoch mehr über die Studie erfahren möchten, Anmerkungen haben oder Hinweise geben möchten, dann können Sie unter *christine-caroline.jaehnig@wiwi.uni-goettingen.de* mit mir Kontakt aufnehmen.

Mit Ihrer gewissenhaften Teilnahme an dieser Befragung leisten Sie einen entscheidenden Beitrag für mein Promotionsvorhaben. Dafür möchte ich Ihnen im Voraus danken.

Herzlichen Dank für Ihre Bereitschaft zur Mitarbeit!

Caroline Jähnig

Angaben zu Ihrer Person

1.1: Welches Geschlecht haben Sie?

□ männlich □ weiblich

1.2: In welchem Jahr wurden Sie geboren? _____

1.3: Ist Deutsch Ihre Muttersprache?

□ Ja □ Nein □ Ich bin mehrsprachig aufgewachsen

1.4: Bitte geben Sie den höchsten Bildungsabschluss für jedes Elternteil von Ihnen an.

Hochschulabschluss	Mutter	Vater
Fachhochschulabschluss	□	□
Fachhochschulreife/Abitur und berufliche Ausbildung	□	□
Fachhochschulreife/Abitur ohne berufliche Ausbildung	□	□
Mittlere Reife und berufliche Ausbildung	□	□
Mittlere Reife ohne berufliche Ausbildung	□	□
Hauptschulabschluss und berufliche Ausbildung	□	□
Hauptschulabschluss ohne berufliche Ausbildung	□	□
Kein Abschluss	□	□

1.5: In welcher Schulform haben Sie Ihre Hochschulzugangsberechtigung erworben?

□ Gymnasium

□ Wirtschaftsgymnasium (berufliches Gymnasium mit Schwerpunkt Wirtschaft)

□ Gesamtschule

□ Abendgymnasium

□ Fachoberschule

□ anderes, nicht wirtschaftliches Fachgymnasium (berufliches Gymnasium)

□ Sonstige:

_____

1.6: Welche durchschnittliche Abschlussnote haben Sie erreicht? _____

1.7: Haben Sie eine kaufmännische Berufsausbildung absolviert?

□ Nein (weiter zu Frage 2.1) □ Ja, **vor** Erlangung der Hochschulzugangsberechtigung □ Ja, **nach** Erlangung der Hochschulzugangsberechtigung

1.8: Welchen berufsbildenden Abschluss haben Sie durch Ihre Ausbildung erworben?

□ Bankkaufmann/-frau

□ Industriekaufmann/-frau

□ Versicherungskaufmann/-frau

□ Kaufmann/-frau für Spedition und Logistikdienstleistung

□ Kaufmann/-frau im Groß- und Außenhandel

□ Kaufmann/-frau für Bürokommunikation

□ Einen anderen Abschluss, und zwar:

_____

1.9: Welche durchschnittliche Abschlussnote haben Sie in der Berufsausbildung erreicht? _____

Angaben zum Studium Göttingen

2.1a: In welchem Studienabschnitt befinden Sie sich?

□ Bachelor □ Master □ Diplom □ Promotion

2.1b: In welchem Fachsemester Ihres Studienabschnitts befinden Sie sich? _____

2.2: Falls Sie bereits im Master-/Promotionsstudium sind, an welcher Universität haben Sie Ihren Bachelorabschluss erworben?

□ Universität Göttingen

□ Einer anderen Universität, und zwar:

_____

2.3: In welchem Studiengang sind Sie derzeit eingeschrieben?

□ Betriebswirtschaft; ggf. Schwerpunkt angeben: _____

□ Wirtschaftspädagogik mit Zweitfach: _____

□ Wirtschaftsinformatik

□ Anderer Studiengang: _____

2.4: Haben Sie die unten aufgeführten Module besucht oder besuchen diese derzeit?

	Ja	Nein	Ein Modul mit ähnlichen Inhalten
Internes Rechnungswesen	□	□	□
Unternehmensführung und Organisation	□	□	□
Einführung in die Finanzwirtschaft	□	□	□
Produktion und Logistik	□	□	□
Beschaffung und Absatz	□	□	□

2.5: Falls ja, welche Note haben Sie in den unten angegebenen Modulen erzielt?

	1	1,3	1,7	2	2,3	2,7	3	3,3	3,7	4	5
Internes Rechnungswesen	□	□	□	□	□	□	□	□	□	□	□
Unternehmensführung und Organisation	□	□	□	□	□	□	□	□	□	□	□
Einführung in die Finanzwirtschaft	□	□	□	□	□	□	□	□	□	□	□
Produktion und Logistik	□	□	□	□	□	□	□	□	□	□	□
Beschaffung und Absatz	□	□	□	□	□	□	□	□	□	□	□
Notendurchschnitt Bachelor (falls zutreffend vorläufig)	□	□	□	□	□	□	□	□	□	□	□

2.6: Wie viele Wochen haben Sie während Ihres Studiums in Praktika mit kaufmännischen Bezügen verbracht und in welchen Bereichen waren Sie tätig? (Bitte zusätzlich Wochenarbeitsstunden angeben)

2.7: Wie viele Wochen sind Sie während Ihres Studiums einer Nebentätigkeit mit kaufmännischen Bezügen nachgegangen und welche Tätigkeiten haben Sie ausgeführt? (Bitte zusätzlich Wochenarbeitsstunden angeben)

2.8: Fragen zum Studieninteresse

	trifft gar nicht zu	trifft sehr begrenzt zu	trifft weitgehend zu	trifft völlig zu
Die Beschäftigung mit Stoffinhalten aus der Betriebswirtschaft wirkt sich positiv auf meine Stimmung aus.	□	□	□	□
Wenn ich genügend Zeit hätte, würde ich mich mit Betriebswirtschaft, auch unabhängig von Prüfungsanforderungen, intensiver beschäftigen.	□	□	□	□
Die Beschäftigung mit den Inhalten und Problemen der Betriebswirtschaft gehören zu meinen Lieblingstätigkeiten.	□	□	□	□
Die Beschäftigung mit Betriebswirtschaft hat für mich recht wenig mit Selbstverwirklichung zu tun	□	□	□	□
Über Inhalte von Betriebswirtschaft zu reden, macht mir Spaß.	□	□	□	□
Wenn ich in einer Bibliothek oder einem Buchladen bin, schmökere ich gerne in Zeitschriften oder Büchern, die Themen aus der Betriebswirtschaft ansprechen.	□	□	□	□
Es ist für mich von großer persönlicher Bedeutung, gerade Inhalte aus dem Fach der Betriebswirtschaftslehre studieren zu können.	□	□	□	□
Im Vergleich zu anderen mir sehr wichtigen Dingen (z. B. Hobbies, soziale Beziehungen) messe ich der Betriebswirtschaft eine geringe Bedeutung bei.	□	□	□	□
Wenn ich ehrlich sein soll, ist mir Betriebswirtschaft eher gleichgültig.	□	□	□	□

2.9: Nachfolgend finden Sie eine Reihe von Aussagen, die sich auf Ihr Lernen im Hauptfach beziehen. Schätzen Sie bei jeder Aussage ein, inwieweit diese auf Sie zutrifft.

Ich lerne im Studium …	stimmt gar nicht					stimmt genau
… weil ich mein Studium erfolgreich abschließen möchte.	□	□	□	□	□	□
… weil ich zu den Besten gehören möchte.	□	□	□	□	□	□
… weil ich gute Leistungen bringen möchte.	□	□	□	□	□	□
… weil mir Erfolg im Studium viel bedeutet.	□	□	□	□	□	□
… weil ich zeigen möchte, dass ich intelligenter bin als andere.	□	□	□	□	□	□
… weil ich herausragende Leistungen zeigen möchte.	□	□	□	□	□	□
… weil ich in den Prüfungen besser abschneiden möchte als andere.	□	□	□	□	□	□
… weil ich bei den Prüfungen möglichst gut abschneiden möchte.	□	□	□	□	□	□

2.10: Im Folgenden möchte ich gerne mehr darüber erfahren, wie Sie sich selber im Studium sehen.

	stimmt gar nicht					stimmt genau
In den meisten Lehrveranstaltungen kann ich mich auf meine Begabung verlassen.	□	□	□	□	□	□
Bei mathematischen oder rechnerischen Problemen werde ich von meinen Bekannten oft um Rat gefragt.	□	□	□	□	□	□
Ich weiß genau, was ich machen muss, um gute Noten zu bekommen.	□	□	□	□	□	□
In den meisten Lehrveranstaltungen erziele ich, aufgrund meiner intellektuellen Begabungen, gute Ergebnisse.	□	□	□	□	□	□
Auch wenn eine Prüfung sehr schwierig ist, weiß ich, was ich tun muss, um sie zu bestehen.	□	□	□	□	□	□
In den meisten Lehrveranstaltungen vertraue ich auf meine Intelligenz.	□	□	□	□	□	□
Bei Aufgaben, die mathematisches oder rechnerisches Denken verlangen, bin ich nicht besonders erfolgreich.	□	□	□	□	□	□
In den meisten Lehrveranstaltungen lerne ich – auch ohne mich anzustrengen – schnell etwas dazu.	□	□	□	□	□	□
In Lehrveranstaltungen, die mit Mathematik und Rechnen zu tun haben, bin ich immer gut gewesen.	□	□	□	□	□	□
Auch bei schwierigen Prüfungsvorbereitungen bin ich in der Lage, mich gezielt mit dem Lernstoff auseinanderzusetzen.	□	□	□	□	□	□
Für die meisten Lehrveranstaltungen sind meine Begabungen sehr hilfreich.	□	□	□	□	□	□

Alles, was mit Mathematik und Rechnen zu tun hat, ist mir schwer verständlich.	□	□	□	□	□	□
Ich bin ziemlich gut in Mathematik und Rechnen.	□	□	□	□	□	□
In den meisten Lehrveranstaltungen erziele ich, aufgrund meiner Fähigkeiten, gute Leistungsergebnisse.	□	□	□	□	□	□
In mathematisch-rechnerischen Fächern bin ich gewöhnlich besser als in anderen Fächern.	□	□	□	□	□	□

2.11: Im folgenden möchte ich gerne mehr über Ihr gegenwärtiges Lernverhalten in betriebswirtschaftlichen Modulen erfahren. Bitte geben Sie für jede der im folgenden genannten Aktivitäten die Häufigkeit an, mit der Sie diese üblicherweise ausführen, wenn Sie sich auf eine Prüfung vorbereiten (z. B. auf Klausuren, Hausarbeiten, Seminarvorbereitungen, Referate usw.).

	sehr selten					sehr oft
Ich versuche, Beziehungen zu den Inhalten verwandter Lehrveranstaltungen herzustellen.	□	□	□	□	□	□
Zu neuen Konzepten stelle ich mir praktische Anwendungen vor.	□	□	□	□	□	□
Ich versuche, neue Begriffe oder Theorien auf mir bereits bekannte Begriffe und Theorien zu beziehen.	□	□	□	□	□	□
Ich stelle mir Sachverhalte bildlich vor.	□	□	□	□	□	□
Ich versuche in Gedanken, das Gelernte mit dem zu verbinden, was ich schon darüber weiß.	□	□	□	□	□	□
Ich denke mir konkrete Beispiele zu bestimmten Lerninhalten aus.	□	□	□	□	□	□
Ich beziehe das, was ich lerne, auf meine eigenen Erfahrungen.	□	□	□	□	□	□
Ich überlege mir, ob der Lernstoff auch für mein Alltagsleben von Bedeutung ist.	□	□	□	□	□	□
Ich frage mich, ob der Text, den ich gerade durcharbeite, wirklich überzeugend ist.	□	□	□	□	□	□
Ich prüfe, ob die in einem Text dargestellten Theorien, Interpretationen oder Schlussfolgerungen ausreichend belegt und begründet sind.	□	□	□	□	□	□
Ich denke über Alternativen zu den Behauptungen oder Schlussfolgerungen in den Lerntexten nach.	□	□	□	□	□	□
Der Stoff, den ich gerade bearbeite, dient mir als Ausgangspunkt für die Entwicklung eigener Ideen.	□	□	□	□	□	□
Es ist für mich sehr reizvoll, widersprüchliche Aussagen aus verschiedenen Texten aufzuklären.	□	□	□	□	□	□
Ich gehe an die meisten Texte kritisch heran.	□	□	□	□	□	□
Ich vergleiche die Vor- und Nachteile verschiedener theoretischer Konzeptionen.	□	□	□	□	□	□
Das, was ich lerne, prüfe ich auch kritisch.	□	□	□	□	□	□

Ich präge mir den Lernstoff von Texten durch Wiederholen ein.	□	□	□	□	□	□
Ich lese meine Aufzeichnungen mehrmals hintereinander durch.	□	□	□	□	□	□
Ich lerne Schlüsselbegriffe auswendig, um mich in der Prüfung besser an wichtige Inhaltsbereiche erinnern zu können.	□	□	□	□	□	□
Ich lerne eine selbst erstellte Übersicht mit den wichtigsten Fachtermini auswendig.	□	□	□	□	□	□
Ich lese einen Text durch und versuche, ihn mir am Ende jedes Abschnitts auswendig vorzusagen.	□	□	□	□	□	□
Ich lerne Regeln, Fachbegriffe und Formeln auswendig.	□	□	□	□	□	□
Ich lerne den Lernstoff anhand von Skripten oder anderen Aufzeichnungen möglichst auswendig.	□	□	□	□	□	□
Ich wiederhole den Stoff anhand der vorgegebenen Kontrollfragen	□	□	□	□	□	□
Wenn ich mir ein bestimmtes Pensum zum Lernen vorgenommen habe, bemühe ich mich, es auch zu schaffen.	□	□	□	□	□	□
Ich strenge mich auch dann an, wenn der Stoff mir überhaupt nicht liegt.	□	□	□	□	□	□
Ich gebe nicht auf, auch wenn der Stoff sehr schwierig oder komplex ist.	□	□	□	□	□	□
Ich lerne auch spätabends und am Wochenende, wenn es sein muss.	□	□	□	□	□	□
Gewöhnlich dauert es lange, bis ich mich dazu entschließe, mit dem Lernen anzufangen.	□	□	□	□	□	□
Vor der Prüfung nehme ich mir ausreichend Zeit, um den ganzen Stoff noch einmal durchzugehen.	□	□	□	□	□	□
Ich nehme mir mehr Zeit zum Lernen als die meisten meiner Studienkollegen.	□	□	□	□	□	□
Ich arbeite so lange, bis ich mir sicher bin, die Prüfung gut bestehen zu können.	□	□	□	□	□	□

2.12: Ich habe an dieser oder einer ähnlichen Befragung bereits teilgenommen.

□ Ja □ Nein

Vielen Dank für Ihre Teilnahme! Hier ist Platz für Ihre Anmerkungen:

Anhang D Skalendokumentationen

Tab. D-1: Skalendokumentation zur Skala „Studieninteresse" aus der Kurzform des FSI (Schiefele et al., 1993), Skala von 1 bis 4

Itemtext	***M***	***SD***	***rit***
Die Beschäftigung mit Stoffinhalten aus der Betriebswirtschaft wirkt sich positiv auf meine Stimmung aus.	2.45	0.74	.70
Wenn ich genügend Zeit hätte, würde ich mich mit Betriebswirtschaft, auch unabhängig von Prüfungsanforderungen, intensiver beschäftigen.	2.51	0.88	.63
Die Beschäftigung mit den Inhalten und Problemen der Betriebswirtschaft gehören zu meinen Lieblingstätigkeiten	2.06	0.79	.69
Die Beschäftigung mit Betriebswirtschaft hat für mich recht wenig mit Selbstverwirklichung zu tun.[Item rekodiert]	2.58	0.86	.50
Über Inhalte von Betriebswirtschaft zu reden, macht mir Spaß.	5.65	0.86	.70
Wenn ich in einer Bibliothek oder einem Buchladen bin, schmökere ich gerne in Zeitschriften oder Büchern, die Themen aus der Betriebswirtschaft ansprechen.	2.04	0.83	.47
Es ist für mich von großer persönlicher Bedeutung, gerade Inhalte aus dem Fach der Betriebswirtschaftslehre studieren zu können.	2.52	0.85	.69
Im Vergleich zu anderen mir sehr wichtigen Dingen (z. B. Hobbies, soziale Beziehungen) messe ich der Betriebswirtschaft eine geringe Bedeutung bei. [Item rekodiert]	2.67	0.87	.52
Wenn ich ehrlich sein soll, ist mir Betriebswirtschaft eher gleichgültig. [Item rekodiert]	3.16	0.87	.66
N ***M*** ***SD*** **Minimum/Maximum** **Cronbachs Alpha**	317 2.48 0.33 2.04/3.16 .87		

Tab. D-2: Skalendokumentation zur Skala „Leistungsmotivation", Skala von 1 bis 6, nach Wild, Krapp, Schiefele, Lewalter und Schreyer (1995) entnommen aus Hußtegge (2011, S. 451)

Itemtext	***M***	***SD***	***rit***
Ich lerne im Studium,			
weil ich mein Studium erfolgreich abschließen möchte.	5.35	0.90	.46
weil ich gute Leistungen bringen möchte.	4.83	1.16	.75
weil mir Erfolg im Studium viel bedeutet.	5.01	1.07	.64
weil ich bei den Prüfungen möglichst gut abschneiden möchte.	4.39	1.24	.56
N ***M*** ***SD*** **Minimum/Maximum** **Cronbachs Alpha**	254 4.89 0.40 4.39/5.35 .79		

Tab. D-3: Skalendokumentation zur Skala „Wettbewerbsmotivation", Skala von 1 bis 6, nach Wild, Krapp, Schiefele, Lewalter und Schreyer (1995) entnommen aus Hußtegge (2011, S. 451)

Itemtext	***M***	***SD***	***rit***
Ich lerne im Studium,			
weil ich zu den Besten gehören möchte.	3.47	1.47	.67
weil ich zeigen möchte, dass ich intelligenter bin als andere.	2.45	1.36	.66
weil ich in den Prüfungen besser abschneiden möchte als andere.	2.84	1.78	.71
weil ich herausragende Leistungen zeigen möchte.	3.49	1.41	.66
N ***M*** ***SD*** **Minimum/Maximum** **Cronbachs Alpha**	259 3.04 0.51 2.45/3.49 .84		

Tab. D-4: Skalendokumentation zur Skala „akademisches Selbstkonzept", Skala von 1 bis 6, nach Marsh und O' Neill (1984) entnommen aus Hußtegge (2011, S. 438–442)

Itemtext	***M***	***SD***	***rit***
In den meisten Lehrveranstaltungen kann ich mich auf meine Begabung verlassen.	3.39	1.22	.59
Ich weiß genau, was ich machen muss, um gute Noten zu bekommen.	3.85	1.23	.48
In den meisten Lehrveranstaltungen erziele ich, aufgrund meiner intellektuellen Begabungen, gute Ergebnisse.	3.40	1.15	.73
Auch wenn eine Prüfung sehr schwierig ist, weiß ich, was ich tun muss, um sie zu bestehen.	4.00	1.78	.65
In den meisten Lehrveranstaltungen vertraue ich auf meine Intelligenz.	3.22	1.21	.63
In den meisten Lehrveranstaltungen lerne ich – auch ohne mich anzustrengen – schnell etwas dazu.	3.53	1.10	.53
Auch bei schwierigen Prüfungsvorbereitungen bin ich in der Lage, mich gezielt mit dem Lernstoff auseinanderzusetzen.	4.11	1.09	.56
Für die meisten Lehrveranstaltungen sind meine Begabungen sehr hilfreich.	3.71	1.16	.73
In den meisten Lehrveranstaltungen erziele ich, aufgrund meiner Fähigkeiten, gute Leistungsergebnisse.	3.79	1.11	.72
N ***M*** ***SD*** **Minimum/Maximum** **Cronbachs Alpha**	243 3.69 0.30 3.22/4.12 .87		

Tab. D-5: Skalendokumentation zur Skala „mathematisches Selbstkonzept", Skala von 1 bis 6 nach March und O' Neill(1984) entnommen aus Hußtegge (2011, S. 438–442)

Itemtext	***M***	***SD***	***rit***
Bei mathematischen oder rechnerischen Problemen werde ich von meinen Bekannten oft um Rat gefragt.	3.47	1.49	.70
Bei Aufgaben, die mathematisches oder rechnerisches Denken verlangen, bin ich nicht besonders erfolgreich. [Item rekodiert]	4.15	1.43	.64
In Lehrveranstaltungen, die mit Mathematik und Rechnen zu tun haben, bin ich immer gut gewesen.	4.34	1.44	.78
Alles, was mit Mathematik und Rechnen zu tun hat, ist mir schwer verständlich. [Item rekodiert]	3.78	1.39	.88
Ich bin ziemlich gut in Mathematik und Rechnen.	3.79	1.41	.90
In mathematisch-rechnerischen Fächern bin ich gewöhnlich besser als in anderen Fächern.	3.46	1.58	.73
N ***M*** ***SD*** **Minimum/Maximum** **Cronbachs Alpha**	250 3.83 0.36 3.46/4.34 .92		

Tab. D-6: Skalendokumentation zur Skala „Lernstrategien: Zusammenhangslernen", Skala von 1 bis 6, nach Wild und Schiefele (1994) entnommen aus Hußtegge (2011, S. 443–446)

Itemtext	***M***	***SD***	***rit***
Ich versuche, Beziehungen zu den Inhalten verwandter Lehrveranstaltungen herzustellen.	3.84	1.41	.61
Zu neuen Konzepten stelle ich mir praktische Anwendungen vor.	3.97	1.38	.66
Ich versuche, neue Begriffe oder Theorien auf mir bereits bekannte Begriffe und Theorien zu beziehen.	4.02	1.17	.56
Ich stelle mir Sachverhalte bildlich vor.	4.23	1.23	.57
Ich versuche in Gedanken, das Gelernte mit dem zu verbinden, was ich schon darüber weiß.	4.46	1.06	.66
Ich denke mir konkrete Beispiele zu bestimmten Lerninhalten aus.	4.17	1.37	.57
Ich beziehe das, was ich lerne, auf meine eigenen Erfahrungen.	4.26	1.37	.60
Ich überlege mir, ob der Lernstoff auch für mein Alltagsleben von Bedeutung ist.	4.20	1.46	.40
N ***M*** ***SD*** **Minimum/Maximum** **Cronbachs Alpha**	315 4.14 0.19 3.84/4.46 .84		

Tab. D-7: Skalendokumentation zur Skala „Lernstrategien: kritisches Hinterfragen", Skala von 1 bis 6, nach Wild und Schiefele (1994) entnommen aus Hußtegge (2011, S. 443–446)

Itemtext	***M***	***SD***	***rit***
Ich frage mich, ob der Text, den ich gerade durcharbeite, wirklich überzeugend ist.	3.91	1.34	.45
Ich prüfe, ob die in einem Text dargestellten Theorien, Interpretationen oder Schlussfolgerungen ausreichend belegt und begründet sind.	3.11	1.39	.69
Ich denke über Alternativen zu den Behauptungen oder Schlussfolgerungen in den Lerntexten nach.	3.33	1.31	.74
Der Stoff, den ich gerade bearbeite, dient mir als Ausgangspunkt für die Entwicklung eigener Ideen.	3.22	1.23	.61
Es ist für mich sehr reizvoll, widersprüchliche Aussagen aus verschiedenen Texten aufzuklären.	2.99	1.34	.63
Ich gehe an die meisten Texte kritisch heran.	3.55	1.23	.63
Ich vergleiche die Vor- und Nachteile verschiedener theoretischer Konzeptionen.	3.42	1.24	.64
Das, was ich lerne, prüfe ich auch kritisch.	3.37	1.25	.67
N ***M*** ***SD*** **Minimum/Maximum** **Cronbachs Alpha**	313 3.36 0.28 2.99/3.91 .87		

Tab. D-8: Skalendokumentation zur Skala „Lernstrategien: Wiederholungsstrategien", Skala von 1 bis 6, nach Wild und Schiefele (1994) entnommen aus Hußtegge (2011, S. 443 bis 446)

Itemtext	***M***	***SD***	***rit***
Ich präge mir den Lernstoff von Texten durch Wiederholen ein.	4.65	1.21	.57
Ich lese meine Aufzeichnungen mehrmals hintereinander durch.	4.55	1.35	.57
Ich lerne Schlüsselbegriffe auswendig, um mich in der Prüfung besser an wichtige Inhaltsbereiche erinnern zu können.	4.75	1.29	.65
Ich lerne eine selbst erstellte Übersicht mit den wichtigsten Fachtermini auswendig.	4.29	1.46	.48
Ich lese einen Text durch und versuche, ihn mir am Ende jedes Abschnitts auswendig vorzusagen.	2.51	1.52	.31
Ich lerne Regeln, Fachbegriffe und Formeln auswendig.	4.66	1.33	.65
Ich lerne den Lernstoff anhand von Skripten oder anderen Aufzeichnungen möglichst auswendig.	4.19	1.53	.64
Ich wiederhole den Stoff anhand der vorgegebenen Kontrollfragen	4.38	1.30	.38
N ***M*** ***SD*** **Minimum/Maximum** **Cronbachs Alpha**	315 4.25 0.73 2.51/4.75 .81		

Tab. D-9: Skalendokumentation zur Skala „Lernstrategien: Anstrengungsbereitschaft", Skala von 1 bis 6, nach Wild und Schiefele (1994), entnommen aus Hußtegge (2011, S. 443 bis 446)

Itemtext	***M***	***SD***	***rit***
Wenn ich mir ein bestimmtes Pensum zum Lernen vorgenommen habe, bemühe ich mich, es auch zu schaffen.	4.80	1.18	.61
Ich strenge mich auch dann an, wenn der Stoff mir überhaupt nicht liegt.	4.77	1.10	.55
Ich gebe nicht auf, auch wenn der Stoff sehr schwierig oder komplex ist.	4.69	1.08	.47
Ich lerne auch spätabends und am Wochenende, wenn es sein muss.	4.90	1.35	.23
Gewöhnlich dauert es lange, bis ich mich dazu entschließe, mit dem Lernen anzufangen. [Item rekodiert]	3.32	1.50	.40
Vor der Prüfung nehme ich mir ausreichend Zeit, um den ganzen Stoff noch einmal durchzugehen.	4.24	1.30	.51
Ich nehme mir mehr Zeit zum Lernen als die meisten meiner Studienkollegen.	3.23	1.40	.47
Ich arbeite so lange, bis ich mir sicher bin, die Prüfung gut bestehen zu können.	4.27	1.28	.56
N ***M*** ***SD*** **Minimum/Maximum** **Cronbachs Alpha**	300 4.28 .66 3.23/4.90 .78		

Anhang E Situative Items und Itemcharakteristika

Hinweis: Jede der folgenden 30 situativen Aufgaben enthält Hinweise auf die Perspektive aus der Sie die Frage beantworten sollen und welches Ziel mit der Beantwortung der Frage erreicht werden soll. Bitte versuchen Sie sich so gut wie möglich in die geschilderte Situation hinein zu versetzen und die Fragestellung vor dem Hintergrund der vorgegebenen Zielerreichung zu beantworten. Bitte kreuzen Sie nur eine Antwortalternative an.

S. 1: Im Rahmen einer Initiative für das Empowerment der Filialstruktur sollen Sie den Filialleitern Ihres Unternehmens Informationen über die geplante Dezentralisierung im Unternehmen weitergeben. Ihr Ziel ist es, die Filialleiter auf veränderte Anforderungen vorzubereiten. Was sprechen Sie an?

□ Ich bereite die Filialleiter auf größere Entscheidungsspielräume vor.

□ Ich bereite die Filialleiter auf Einbußen in der Planungsfreiheit vor.

□ Ich bereite die Filialleiter auf konsequente Erfolgskontrollen vor.

□ Ich bereite die Filialleiter auf eine größere Arbeitsbelastung vor.

S. 2: Sie sind im Produktmanagement tätig und derzeit auf der Jahrestagung Ihres großen internationalen Arbeitgebers. Kürzlich haben externe Unternehmensberater von der Boston Consulting Group (BCG) eine Einordnung aller Firmenprodukte in die Produkt-Portfolio-Matrix vorgenommen. Das von Ihnen betreute Produkt wurde als ein Produkt mit niedrigem relativem Marktanteil und geringem Marktwachstum eingestuft. In einer internen PowerPoint Präsentation sollen Sie Ihr Produkt kurz vorstellen. Ihr Ziel ist es, anhand einer Abbildung das Produkt möglichst positiv darzustellen. Wie präsentieren Sie Ihr Produkt?

□ Ich präsentiere als Aufhänger das Bild eines ärmlich aussehenden Hundes, der eine Kuh bewacht, um die positiven Funktionen meines Produktes hervorzuheben.

□ Ich beginne mit dem Bild einer Kuh, die ärmlich aussieht und unter einem Vordach Unterschlupf sucht, um zu verdeutlichen, dass das Produkt unter das Schutzdach des Unternehmens gestellt werden sollte.

□ Die Präsentation beginnt mit einer Kuh, die sich in ein Fragezeichen verwandelt, um darauf hinzuweisen, dass noch viel Potenzial in meinem Produkt steckt.

□ Die Präsentation beginnt mit einem ärmlich aussehenden Hund, der in eine Decke gewickelt ist. Die Decke soll auf den positiven Deckungsbeitrag des Produktes verweisen.

S. 3: Sie sind in einem großen Unternehmen tätig. Ihr Ziel ist es, ein Unternehmensleitbild zu erstellen. Wie gehen Sie vor?

□ Ich befrage den Vorstand nach den Unternehmenszielen und fasse diese zu einem Leitbild zusammen.

□ Das Leitbild hat hauptsächlich eine Kommunikationsfunktion an die Öffentlichkeit, deshalb sollte es von der Marketingabteilung erstellt werden.

□ Leitbilder haben eine wichtige Funktion für die Unternehmenskultur, daher organisiere ich einen Prozess, der die Mitarbeiter einbezieht.

□ Ich hole mir Anregungen bei veröffentlichten Leitbildern vergleichbarer Unternehmen und informiere mich über deren Erstellungsprozess.

S. 4: Sie sind im Vergütungsmanagement angestellt und für die Bewertung und Weiterentwicklung der bestehenden Vergütungssysteme im Unternehmen zuständig. Die Herstellung Ihrer teilweise komplexen Produkte erfordert von den Mitarbeitern unterschiedlichen Zeiteinsatz. Ziel ist es, den persönlichen Einsatz von Mitarbeitern in der Produktion zu erhöhen. Welche Lohnart wählen Sie?

□ Gewinnbeteiligung.

□ Zeitlohn.

□ Prämienlohn.

□ Akkordlohn.

S. 5: Sie sind im Besitz eines Unternehmens. Derzeit finden Verhandlungen über die Übernahme Ihres Unternehmens durch Konkurrenzunternehmen statt. Ihr Ziel ist es, den bestmöglichen Preis für die Übernahme zu generieren. Wie gehen Sie vor, um Ihr Ziel zu erreichen?

□ Ich informiere potenzielle weitere Interessenten.

□ Ich verlangsame den Verhandlungsprozess.

□ Ich veröffentliche geschönte Unternehmensinformationen.

□ Ich kollaboriere mit einem bevorzugten Interessenten.

S. 6: In Ihrem aufstrebenden mittelständischen Unternehmen wird diskutiert, das bisherige Einlininensystem der Leitungskonfiguration den neuen Bedingungen anzupassen. Ziel ist es, die Leitungsebene bei speziellen fachlichen Fragen zu entlasten, ohne die Weisungsbeziehungen zu verändern. Welche Leitungskonfiguration schlagen Sie vor?

□ Matrixkonfiguration.

□ Mehrlininenkonfiguration.

□ Stablininenkonfiguration.

□ Projektkonfiguration.

S. 7: Sie sind in einer strategischen Unternehmensberatung tätig und sollen ein Softwareunternehmen dahingehend beraten, ob eine business-to-business oder eine business-to-consumer Strategie verfolgt werden soll. Ihr Ziel ist es, zwei Businesspläne für die beiden Strategien zu entwickeln und diese dem Management vorzustellen. Welche Frage klären Sie im ersten Schritt der Erstellung der Business-Pläne?

□ Was sind die Kernkompetenzen des zu beratenen Unternehmens?

□ Wie ist die Größe des zu erschließenden Marktes?

□ Welche Strategie maximiert den Gewinn?

□ Besteht das Humankapital, um die jeweilige Strategie zu verfolgen?

S. 8: Nach dem Besuch eines Managementseminars zum Thema Work-Life-Balance ist es Ihr Ziel als Führungskraft, durch laterale Kommunikation die Implementierung eines Programms zur Verbesserung der Work-Life-Balance in Ihrem Unternehmen vorzubereiten. Wie gehen Sie vor?

- □ Ich stelle entsprechende Unterlagen zusammen und versuche die Geschäftsführung für mein Vorhaben zu gewinnen.
- □ Ich versuche als erstes die Kollegen auf meiner Führungsebene für das Programm zu gewinnen.
- □ Ich versuche die Mitarbeiter für das Programm zu begeistern.
- □ Ich versuche Zielgruppen aus unterschiedlichen Unternehmensbereichen und Führungsebenen für das Programm zu begeistern.

S. 9: Sie wollen in eine neue Produktionsanlage investieren. Ihr Ziel ist es, vorab festzustellen ob die Investition sich lohnt. Welche Methode legen Sie der Bewertung der Investition zugrunde?

- □ Kostenvergleichsmethode.
- □ Kapitalwertmethode.
- □ Rentabilitätsvergleichsrechnung.
- □ Gewinnvergleichsrechnung.

S. 10: Das Unternehmen, in dem Sie arbeiten, erzielte im letzten Jahr 1.200.000 € Umsatzerlöse. Davon erwirtschaftete die Abteilung, die Sie leiten, 700.000 €. Die direkt Ihrer Abteilung zuzuordnenden Kosten belaufen sich auf 340.000 €. Die Gesamtkosten Ihres Unternehmens belaufen sich auf 720.000 €. Ziel ist es, den Deckungsbeitrag der Abteilung zu den gesamten Fixkosten zu berechnen und zu bewerten. Wie gehen Sie vor?

- □ Die Abteilung trägt mit 360.000 € Deckungsbeitrag einen bedeutsamen Teil zur Deckung der Fixkosten bei.
- □ Die Abteilung trägt mit 500.000 € Deckungsbeitrag einen bedeutsamen Teil zur Deckung der Fixkosten bei.
- □ Die Abteilung ist mit 360.000 € Deckungsbeitrag nur knapp profitabel.
- □ Abteilung trägt mit 500.000 €nur knapp zur Deckung der Gesamtkosten bei.

S. 11: In Ihrer Business Unit zur Produktion von Radioweckern fallen monatlich Fixkosten von 75.000 € an. Die variablen Kosten für Ihr Produkt betragen 5 €. Sie verkaufen das Produkt am Markt für 20 €. Als Pricing Manger ist es Ihr Ziel, die Menge an zu produzierenden Radioweckern zu errechnen, bei der die Gewinnschwelle erreicht ist. Bei welcher Menge wird der Break-Even-Point erreicht?

- □ 5000 Stück
- □ 3750 Stück
- □ 3000 Stück
- □ 15000 Stück

S. 12: Sie sind im Bereich Controlling in einem produzierenden Unternehmen beschäftigt. Für jede Einheit wird in Ihrem Unternehmen mit einer Vorgabezeit von 2 Stunden gerechnet. In der letzten Abrechnungsperiode sind bei 442.000,00 € Ist-Kosten in 5.145 Stunden 2.450 Einheiten bearbeitet worden. Ziel ist es, einen Soll-Ist Vergleich durchzuführen. Zu welchem Ergebnis kommen Sie?

- □ Die Ist-Bearbeitungszeit entspricht genau der Soll-Zeit.
- □ Es liegen 0,6 % Ist- Soll-Abweichung vor.
- □ Der Ist-Wert liegt 6 Minuten unter dem Soll-Wert.
- □ Der Soll-Wert liegt 6 Minuten unter dem Ist-Wert.

S. 13: Ihr Unternehmen strebt an, sein Unternehmenswachstum durch ein Wachstum der Auslandsmärkte zu erreichen. Hierzu wird ein neues Vertriebsbüro in Spanien eingerichtet. Ihr Ziel ist es, ein Konzept für das Berichtsformular für das spanische Vertriebsbüro zu erstellen. Wie sieht Ihr Vorschlag aus?

- □ Das Berichtsformular wird so konzipiert, dass das Erreichen der Jahresziele überprüft werden kann.
- □ Das Berichtsformular wird so konzipiert, dass Umsätze und Kosten auf monatlicher Basis verglichen werden können.
- □ Das Berichtsformular wird so konzipiert, dass die Arbeitszeit der Vertriebsmitarbeiter erfasst wird.
- □ Das Berichtsformular wird so konzipiert, dass Kennzahlen über die Zahl der besuchten Händler und die Anzahl der Bestellungen erfasst werden können.

S. 14: Sie sind im Bereich Controlling in einem produzierenden Unternehmen beschäftigt. Für jede Einheit wird in Ihrem Unternehmen mit einer Vorgabezeit von 2 Stunden gerechnet. In der letzten Abrechnungsperiode sind bei 442.000,00 € Ist-Kosten in 5.145 Stunden 2.450 Einheiten bearbeitet worden. Ziel ist es, einen Soll-Ist Vergleich durchzuführen. Zu welchem Ergebnis kommen Sie?

- □ Die Ist-Bearbeitungszeit entspricht genau der Soll-Zeit.
- □ Es liegen 0,6 % Ist- Soll-Abweichung vor.
- □ Der Ist-Wert liegt 6 Minuten unter dem Soll-Wert.
- □ Der Soll-Wert liegt 6 Minuten unter dem Ist-Wert.

S. 15: In Ihrem Produktionsprozess gibt es verschiede Optionen des Outsourcings. Ziel ist es, die Entscheidungsfindung zu unterstützen und diese Outsourcingmöglichkeiten unter Einbeziehung folgender Kriterien zu bewerten: Qualität, Kernkompetenzen, Kapitalbindung, Flexibilität, Unabhängigkeit, Kundenorientierung und Prozessorientierung. Wie gehen Sie vor?

- □ Ich schreibe zu jeder Outsourcingoption und jedem Kriterium einen kurzen Bericht.
- □ Ich errechne für jede Outsourcingoption einen Kennwert und reihe die Optionen nach Größe des Kennwerts.
- □ Ich ordne die Optionen entsprechend ihrer Gewinnerwartung.
- □ Ich vergebe jeder Outsourcingoption Punktwerte für die Kriterien und errechne so eine Gesamtwertung.

S. 16: Sie sind für die Planung und Steuerung des Innovationsprozesses Ihres Unternehmens zuständig. Zur Bewertung des Prozesses liegen Ihnen verschiedene Innovationskennzahlen vor. Ziel ist es, die aktuelle F & E-Quote von 2 % mit der Zielausprägung von 2,5 % zu vergleichen. Wie bewerten Sie die Kennzahlen?

□ Sowohl Ist-Zustand als auch Zielquote sind konservativ niedrig.

□ Die Abweichung von 0,5 % sollte behoben werden.

□ Eine solch geringe Abweichung von Ist- und Soll- Zustand kann vernachlässigt werden.

□ Das Unternehmen steht bezüglich der F & E-Quote ausgesprochen gut da.

S. 17: Sie sind in einer Unternehmensberatung tätig. Ihr Ziel ist es, eine Geschäftseinheit Ihres Kunden (globales Industrieunternehmen) hinsichtlich der Wirtschaftlichkeit zu analysieren. Welche Faktoren vernachlässigen Sie in Ihrer Analyse?

□ Löhne und Gehälter.

□ Voraussichtliche Gewinne der Wettbewerber.

□ Kosten der Materialen im Produktionsprozess.

□ Die Preise zu denen die Produkte des Unternehmens verkauft werden.

S. 18: Die Unternehmensleitung möchte das Unternehmen durch den Zukauf von kleineren Unternehmen vergrößern. Für ein potenzielles Kaufobjekt stehen Ihnen die Kennzahlen aus der unten angeführten Tabelle zur Verfügung (alle Angaben in €). Ziel ist es, die finanzielle Lage des für den Kauf in Frage kommenden Unternehmens zu beurteilen. Was steht in Ihrem Bericht für die Unternehmensleitung?

Aktiva		Passiva	
Immobilien	15.000.000	Eigenkapital	16.800.000
Maschinen	14.000.000		
Fuhrpark	3.000.000		
Büro- und Geschäfts-ausstattung	1.800.000		
Warenbestad	5.000.000	Bankkredit: (langfristige Verbindlichkeiten)	21.000.000
Forderungen	3.000.000		
Kasse, Bank	5.000.000	Kurzfristige Verbindlichkeiten	9.000.000
Gesamtvermögen	46.800.000	Gesamtkapital	46.800.000

□ Mit einem Anlagevermögen von 33.800.000 € ist das Unternehmen finanziell sehr gut aufgestellt.

□ Das Anlagevermögen ist zu gering.

□ Die Liquidität 1. Grades ist zu gering.

□ Die kurzfristigen Verbindlichkeiten sind zu hoch.

S. 19: Sie sind in einer unabhängigen Finanz- und Vermögensberatung tätig. Ihr Kunde ist extrem risikoscheu, bittet Sie jedoch eine Empfehlung für den Kauf eines Wertpapiers auszusprechen. Ihr Ziel ist es, ein den Kundenwünschen entsprechendes Wertpapier zu empfehlen. Welche Eigenschaften hat das Wertpapier, das Sie empfehlen?

- □ Großer erwarteter Gewinn und große Streuung des erwarteten Gewinns.
- □ Mittlerer erwarteter Gewinn und mittlere Streuung des erwarteten Gewinns.
- □ Mittlerer erwarteter Gewinn und kleine Streuung des erwarteten Gewinns.
- □ Großer erwarteter Gewinn und kleine Streuung des erwarteten Gewinns.

S. 20: Ihr Ziel ist es, das Dispositionsverfahren für den Materialbedarf mit Hilfe einer ABC-Analyse zu bestimmen. Wie gehen Sie vor?

- □ Ich systematisiere Materialien entlang ihrer Gesamtverbrauchsmenge ihres Gesamtverbrauchswertes und ihrer Qualität.
- □ Ich systematisiere Materialien entlang ihrer Gesamtverbrauchsmenge und ihres Gesamtverbrauchswertes.
- □ Ich systematisiere Materialien entlang ihrer Bestelldringlichkeit und Qualität.
- □ Ich systematisiere Materialien nach ihrer Qualität und ihres Gesamtverbrauchwertes.

S. 21: Sie arbeiten im Bereich Produktionsplanung. Sie haben die Wahl zwischen zwei verschiedenen „Produktionsverfahren". Ziel ist es, das effizienteste Verfahren auszuwählen, ohne den Arbeitseinsatz der Mitarbeiter zu erhöhen. Welche Kriterien legen Sie zu Grunde?

- □ Ich wähle das Verfahren, das bei gegebenem Faktoreinsatz eine maximale Ausbringungsmenge erreicht.
- □ Ich wähle das Verfahren, das eine gegebene Ausbringungsmenge mit minimalem Faktoreinsatz erreicht.
- □ Ich wähle das Verfahren, das bei minimalem Faktoreinsatz maximale Ausbringungsmenge erreicht.
- □ Ich wähle das Verfahren, das bei maximalem Faktoreinsatz maximale Ausbringungsmenge erreicht.

S. 22: Sie sind in einem produzierenden Unternehmen tätig, welches an der Optimierung des Warenflusses interessiert ist. Ihr Ziel ist es, im Rahmen des Supply Chain Managements collaborative planning einzuführen. Wer wird in dieses Vorhaben einbezogen?

- □ Alle an der Produktion beteiligten Abteilungen des Unternehmens.
- □ Alle Application Manager im Bereich Produktion und Logistik.
- □ Die abnehmenden Handelsunternehmen.
- □ Die Endkunden.

S. 23: Es kommt vor, dass Verbrauchsmaterialien nicht rechtzeitig bestellt werden und somit zum Produktionszeitpunkt nicht bereitstehen. Ihr Ziel ist es, diesem Missstand entgegen zu wirken. Wie gehen Sie vor?

- □ Ich veranlasse, dass die Bestände des Materiallagers vergrößert werden.
- □ Ich lasse die Bestellungen auf einen wöchentlichen Rhythmus umstellen.
- □ Ich lasse den Mindestbestand erhöhen.
- □ Ich lasse den Meldebestand anpassen.

S. 24: Ziel des produzierenden Unternehmens, in dem Sie arbeiten ist es, durch gelungene Produktpolitik die Abnehmer Ihrer Produkte zufrieden zu stellen. Was tun Sie vor dem Hintergrund dieser Zielvorgabe?

□ Ich intensiviere den Werbemitteleinsatz auf breit gefächerten Kommunikationskanälen.

□ Ich binde den Kunden stärker in die Produktentwicklung ein.

□ Ich optimiere die produktbezogenen Verpackungen.

□ Ich optimiere den Preis, damit die Produkte günstiger werden.

S. 25: Sie sind im Bereich Marktforschung tätig. Ihr Ziel ist es, in Erfahrung zu bringen, wie sich die Endkundenteilnahme an einer Rabattaktion auf das mittelfristige und langfristige Kaufverhalten der Kunden bezüglich eines Produktes auswirkt. Welche Art von Studie planen Sie?

□ Eine Beobachtungsstudie.

□ Eine retrospektive Befragung.

□ Eine Querschnittuntersuchung.

□ Eine Längsschnittuntersuchung.

S. 26: Ziel Ihres Unternehmens ist es, verstärkt auf family branding zu setzen. Wie werden Sie dieses Ziel in der Marketingabteilung umsetzen?

□ Es wird auf familienfreundliche Produktdarstellung geachtet.

□ Es werden viele familienbezogene Produkte in das Portfolio mit aufgenommen und beworben.

□ Es werden Produkte einer Produktfamilie zusammen beworben.

□ Es werden qualitativ unterschiedliche Produkte zu einer Dachmarke zusammengefasst.

S. 27: Sie müssen sich entscheiden, wen Sie mit einem Marktforschungsauftrag des Unternehmens, in dem Sie arbeiten, betrauen. Ein wichtiges Ziel ist es, dass Informationen über die Forschung nicht frühzeitig an die Öffentlichkeit gelangen. Wen betrauen Sie mit dem Forschungsauftrag?

□ Die eigene Marktforschungsabteilung.

□ Ein externes Marktforschungsunternehmen.

□ Eine universitäre Forschungseinrichtung.

□ Eine unternehmensinterne Projektgruppe.

S. 28: Sie haben die Daten aus der unten aufgeführten Tabelle über die Preisbereitschaft der Kunden für eines Ihrer Produkte (Spezialreiniger) vorliegen. Ihr Ziel ist es, den optimalen Einzelpreis zu bestimmen. Welchen Preis bestimmen Sie für den Fall, dass die variablen Stückkosten bei 0 € liegen?

Nachfragernr.	Preisbereitschaft für Spezialreiniger
1	6,00 €
2	5,00 €
3	3,00 €
4	2,40 €

□ Ich wähle den höchsten Preis: 6,00 €.

□ Ich wähle das arithmetische Preismittel: 4,10 €.

□ Ich wähle den umsatzmaximierenden Preis: 3,00 €.

□ Ich wähle den umsatzmaximierenden Preis: 5 €.

S. 29: Sie haben die Daten aus der unten angeführten Tabelle über die Preisbereitschaft der Kunden für 2 Produkte vorliegen. Ihr Ziel ist es, zu bestimmen wer das Produkt kauft, wenn Sie einen Bündelpreis von 5,50 € wählen. Wie gehen Sie vor?

Nachfragernr.	Preisbereitschaft für Spezialreiniger	Preisbreitschaft für Wischsystem
1	6,00 €	1,00 €
2	2,00 €	5,00 €
3	5,00 €	4,00 €
4	3,00 €	2,50 €
5	2,40 €	1,80 €

□ Ich summiere die Preisbereitschaften pro Person und vergleiche diese mit dem Bündelpreis: Nachfrager 1–4 kaufen das Produkt.

□ Ich teile den Bündelpreis durch 2 und vergleiche das Ergebnis mit den einzelnen Preisbereitschaften: Nachfrager 1–4 kaufen das Produkt.

□ Ich teile den Bündelpreis durch 2 und vergleiche das Ergebnis mit der Summe der Preisbereitschaften pro Person: Keiner kauft das Produkt.

□ Ich vergleiche die jeweils höchste maximale Preisbreitschaft pro Person mit dem Bündelpreis: Nachfrager 1 kauft das Produkt.

S. 30: Es liegen Ihnen zwei Marketingstrategien vor. Wie gewinnbringend die beiden Strategien sind, hängt davon ab, ob in diesem Jahr noch ein Konkurrenzprodukt auf den Markt kommt.
Experten rechnen damit, dass mit einer 40 %igen Wahrscheinlichkeit der Wettbewerb mindestens ein Konkurrenzprodukt auf den Markt bringt. Ihr Ziel ist es, diejenige Strategie zu wählen, die den Gewinn für das Unternehmen maximiert. Wie gehen Sie auf Grundlage der unten aufgeführten Tabelle vor?

	Situation 1 Konkurrenzprodukt nicht dem Markt (W1 = .6)	**Situation 2** Konkurrenzprodukt auf dem Markt (W2 = .4)
Strategie 1: Kostenorientierung	Gewinn: 14.000 €	Gewinn: 6.000 €
Strategie 2: Qualitätsorientierung	Gewinn: 9.000 €	Gewinn: 12.000€

□ Ich ermittele den erwarteten Gewinn pro Strategie, summiere diesen über beide Situationen hinweg und wähle die Strategie mit dem höchsten erwarteten Gewinn: Ich wähle Strategie 1.

□ Ich summiere die Gewinne über die Situationen hinweg und wähle die Alternative mit dem höchsten Gewinn: Ich wähle Strategie 2.

□ Ich multipliziere den Gewinn mit der Eintrittswahrscheinlichkeit und wähle die Strategie mit dem höchsten Erwartungswert pro Situation: Ich wähle Strategie 1.

□ Ich wähle die Strategie mit der geringeren Gewinnschwankung: Ich wähle Strategie 2.

Tab. E-1: Itemkennwerte der Haupterhebung (N = 351) und Kennwerte der verlängerten Testversion (N = 35)

Nr. im Fragebogen	Nr. in Rasch-Analysen	Lösungsquote	Trennschärfe	ß	wMNSQ	T
S. 1	**1**	57 %	.36	–0.321	1.01	0.2
S. 2		25 %	.17	*Item wurde ausgeschlossen*		
S. 3		76 %	.17	*Item wurde ausgeschlossen*		
S. 4		40 %	.16	*Item wurde ausgeschlossen*		
S. 5		83 %	.14	*Item wurde ausgeschlossen*		
S. 6	**2**	44 %	.37	0.262	0.99	–0.4
S. 7	**	71 %	.27	*Item noch nicht Rasch skaliert*		
S. 8	**3**	74 %	.39	–1.148	0.97	–0.5
S. 9	**	77 %	.09	*Item noch nicht Rasch skaliert*		
S. 10	**4**	41 %	.33	0.389	1.02	0.7
S. 11	**	74 %	.22	*Item noch nicht Rasch skaliert*		
S. 12	**5**	52 %	.25	–0.122	1.05	1.6
S. 13	**6**	35 %	.41	0.651	0.97	–0.6
S. 14	**	74 %	.36	*Item noch nicht Rasch skaliert*		
S. 15	**7**	70 %	.39	–0.980	0.98	–0.4
S. 16	**8**	26 %	.32	1.084	1.01	0.2
S. 17	**9**	28 %	.23	0.992	1.03	0.6
S. 18	**10**	42 %	.35	0.326	0.99	–0.3
S. 19	**11**	72 %	.40	–1.024	0.97	–0.6
S. 20	**12**	25 %	.25	1.196	1.03	0.5
S. 21	**	71 %	.41	*Item noch nicht Rasch skaliert*		
S. 22	**13**	67 %	.35	–0.863	0.98	–0.3
S. 23	**14**	81 %	.33	–1.547	0.98	–0.2
S. 24	**	43 %	.30	*Item noch nicht Rasch skaliert*		
S. 25	**15**	34 %	.25	0.665	1.04	0.9
S. 26	**16**	67 %	.40	–0.499	0.98	–0.5
S. 27	**17**	55 %	.32	–0.183	1.01	0.5
S. 28	**18**	54 %	.37	–0.171	1.00	0.1
S. 29	**19**	72 %	.47	–0.512	0.96	–1.1
S. 30	**20**	15 %	.21	1.807***	1.02	–1.1

* Item wurde aus Haupterhebung ausgeschlossen
** Item wurde an einer Substichprobe von 35 Personen getestet
*** dieser Parameter wurde fixiert, sodass die Summe aller Parameter 0 ergibt

Tab. E-2: Bezüge der Items zum Curriculum sowie zu Arbeitsanforderungen und Itemlösungen

Inhaltsbereich	Nr.	Bezug zum Curriculum	Bezug zu Stellenanzeigen	richtige Lösung	teil-richtige Lösung
Unternehmensführung	**S. 1**	Strategien der Unternehmensführung	Begleitung von Veränderungsmaßnahmen Mitarbeiter führen und Mitarbeitermotivation	1	3
	S. 2	Konzepte des strategischen Managements (BCG-Matrix)	Internes Reporting	4	1
	S. 3	Grundkenntnisse organisatorischer Gestaltung	Mitarbeiter führen und Mitarbeitermotivation	3	2
	S. 4	Grundlagen des verhaltensorientierten Managements	Maßnahmen zur Effizienz- und Effektivitätssteigerung	3	1
	S. 5	Konzepte des strategischen Managements (mergers & acquisitions)	Strategische Steuerung und Optimierung	1	2
	S. 6	Führungssysteme	Implementieren von Prozessen und Systemen	3	1
	S. 7	Strategische Unternehmensplanung	Strategische Steuerung	1	-
	S. 8	Führungsansätze	Steuerung von Änderungsmaßnahmen	2	-
Finanz- und Rechnungswesen	**S. 9**	Methoden und Verfahren der internen Unternehmensrechnung (Kapitalwertmethode)	Aufarbeitung betriebswirtschaftlicher Kennzahlen	2	-
	S. 10	Deckungsbeitragsrechnung	Wirtschaftlichkeitsanalyse durchführen	1	3
	S. 11	Methoden und Verfahren der internen Unternehmensrechnung (Break-Even-Point)	Wirtschaftlichkeitsanalyse durchführen	1	-
	S. 12	Methoden und Verfahren der internen Unternehmensrechnung (Sensitivitätsanalyse)	Entscheidungsvorlagen erstellen	2	1
	S. 13	Informationsinteressen der planenden, leitenden und kontrollierenden Personen im Unternehmen	Target tracking zur Erreichung der Ziele	2	4
	S. 14	Methoden und Verfahren der internen Unternehmensrechnung (Soll-Ist-Vergleich)	Wirtschaftlichkeitsanalysen durchführen	4	3

(Fortsetzung Tab. E-2)

Inhaltsbereich	Nr.	Bezug zum Curriculum	Bezug zu Stellenanzeigen	richtige Lösung	teil-richtige Lösung
	S. 15	Spezielle Entscheidungsprobleme (Scoring-Modelle)	Entscheidungsvorlagen erstellen	4	2
	S. 16	Ergebniskontrolle durch Kennzahlen	Kennzahlenanalyse	1	2
	S. 17	Methoden und Verfahren der internen Unternehmensrechnung	Wirtschaftlichkeitsanalysen durchführen	2	-
	S. 18	Kosten-Erlös- und Ergebniskontrolle	Wirtschaftlichkeitsanalysen durchführen	3	4
	S. 19	Bewertung von Aktien	Finanz- und Investitionsplanung	2	-
Produktion	**S. 20**	Bereitstellungsplanung /Beschaffungslogistik	Prozesse und Systeme einführen	2	1
	S. 21	Grundlagen der Produktionswirtschaft	Prozesse und Systeme einführen	1	2
	S. 22	Produktionsplanung	Prozesse und Systeme einführen	3	1
	S. 23	Bereitstellungsplanung/Beschaffungslogistik	Prozesse und Systeme einführen	4	1
Marketing	**S. 24**	Marketing-Mix (Produktpolitik)	Produktbetreuung und Entwicklung	2	3
	S. 25	Marktforschungsmethoden	Marktforschung	4	2
	S. 26	Absatzpolitische Instrumente	Marketinginstrumente	3	4
	S. 27	Marktforschungsmethoden	Marktforschung	1	4
	S. 28	Preise bilden	Preispolitik	4	3
	S. 29	Preise bilden	Preispolitik	1	3
	S. 30	Ziele, Strategien und Maßnahmen kontrollieren	Analyse und Prognose von Absatzmärkten	1	3

Anhang F Berechnungen zur Haupterhebung

$$p\left(X_{vi}=1\right)=\frac{exp(\sum_{j=1}^{h} q_{ij}\,\theta_{vj}-\beta_i)}{1+exp(\sum_{j=1}^{h} q_{ij}-\beta_i)} \qquad \text{(Formel 5)}$$

Modellgleichung des Multidimensional Random Coefficients Multinominal Logit Modell für den Fall des Rasch-Modells (Rost, 2004)

q_{ij} = *Q-Matrix der Testmodelle*

j = latente Dimension

i = Item,

θ = *Personenparameter*

β = *Schwierigkeitsparameter*

Tab. F-1: Interkorrelationsmatrix (Pearson Korrelation, zweiseitiges Signifikanzniveau) von Lernstrategien und Testleistung

		1	2	3
[1] Leistung im situativen Test	*r*	1		
	N	351		
[2] Lernstrategie: Zusammenhangslernen	*r*	–,065	1	
	N	323	418	
[3] Lernstrategie: Wiederholung	*r*	,057	,036	1
	N	320	316	320

Tab. F-2: Ergebnisse des Wald-Tests zur Identifikation von Differential Item Functioning auf Itemebene (als kritisch eingestufte Items sind grau gekennzeichnet)

Item Nr.	**Geschlecht männlich/weiblich**			**Standort Göttingen/Anderer**			**Studiengang BWL/Wipäd**		
	Item*Geschlecht			**Item*Standort**			**Item*Studiengang**		
	est. weibl.	***z***	***p***	**est. And.**	**z**	**p**	**est. BWL**	**z**	**p**
1	–0.149	1.324	0.19	–0.019	0.116	0.91	0.013	–0.037	0.97
6	0.024	–0.181	0.86	–0.102	0.961	0.34	0.044	–0.398	0.69
7	0.341	–2.810	0.01	–0.075	0.383	0.70	0.100	–0.540	0.59
8	0.048	–0.382	0.70	0.535	4.401	0.000	–0.592	4.622	0.00
9	–0.178	1.622	0.11	–0.234	2.040	0.04	0.224	–1.806	0.07
10	0.267	–2.245	0.03	–0.046	0.517	0.61	–0.143	1.051	0.29

(Fortsetzung Tab. F-2)

Item Nr.	Geschlecht männlich/weiblich			Standort Göttingen/Anderer			Studiengang BWL/Wipäd		
	Item*Geschlecht			Item*Standort			Item*Studiengang		
	est. weibl.	*z*	*p*	est. And.	z	p	est. BWL	z	p
11	–0.278	2.227	0.03	0.024	–0.367	0.71	0.131	–0.837	0.40
12	0.026	–0.103	0.92	0.183	–1.260	0.21	–0.185	1.236	0.22
13	–0.309	2.655	0.01	0.028	–0.058	0.95	0.201	–1.629	0.10
14	0.087	–0.740	0.46	0.112	–0.911	0.36	–0.017	0.090	0.93
15	0.128	–1.177	0.24	0.028	–0.405	0.69	–0.085	0.757	0.45
16	0.078	–0.507	0.61	–0.078	0.803	0.42	0.023	–0.340	0.73
17	–0.101	0.775	0.44	0.233	–2.102	0.04	–0.223	1.763	0.08
18	0.181	–1.465	0.14	–0.163	0.872	0.38	0.121	–0.618	0.54
19	–0.291	2.618	0.01	–0.689	5.841	0.00	0.764	–5.591	0.00
20	–0.198	1.726	0.08	0.017	–0.232	0.82	0.009	–0.010	0.99
21	–0.114	1.033	0.30	–0.077	0.657	0.51	–0.078	0.680	0.50
22	0.274	–2.485	0.01	0.280	–2.501	0.01	–0.321	2.606	0.01
23	–0.055	0.442	0.66	–0.000	–0.086	0.93	0.075	–0.500	0.62
24	0.220*	–1.334	0.18	0.042*	0.064	0.95	–.061*	0.163	0.87

In der Spalte est. (für estimate) drückt ein Wert mit negativen Vorzeichen aus, um viel Logits das Item für die angegebene Subgruppe leichter ist als für die Vergleichsgruppe und vice versa für positive Werte.

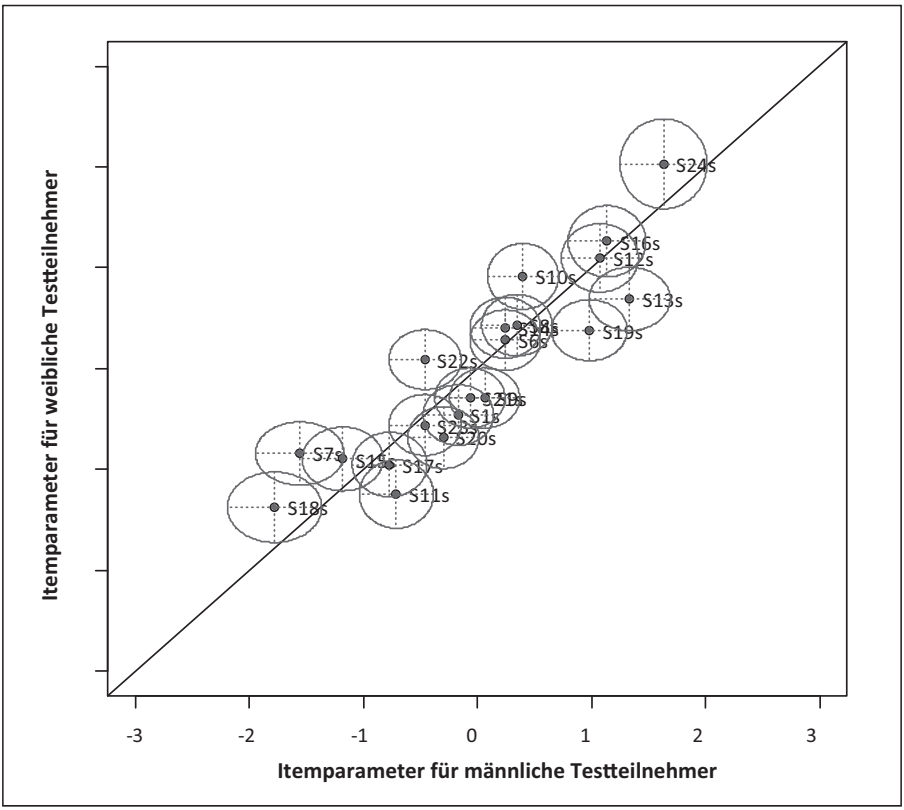


Abb. F-1: Grafischer Modelltest für männliche und weibliche Subgruppen in der Stichprobe mit 95 %-Konfidenzregion um jedes Item

Tab. F-3: OLS-Regression zur Kriterumsvalidierung nach Reduktion des Tests um „unfaire" Items

	Leistung im situativen BWL-Test			
Variable	***ß***	**stand. *ß***	***SE***	***p***
Konstante	–.467	.453		
Anzahl der besuchten BWL-Pflichtmodule	.085	.044	.115	.05
Note in den BWL- Pflichtmodulen	–.199	.088	–.130	.03
Note der Hochschulzugangsberechtigung	–.248	.100	–.143	.01
Absolvierte kaufm. Ausbildung	.578	.111	.309	.00
Geschlecht männlich	.209	.107	.112	.05
Interesse an BWL	.236	.090	.150	.01
korr. R^2	.16			
F	9.706			

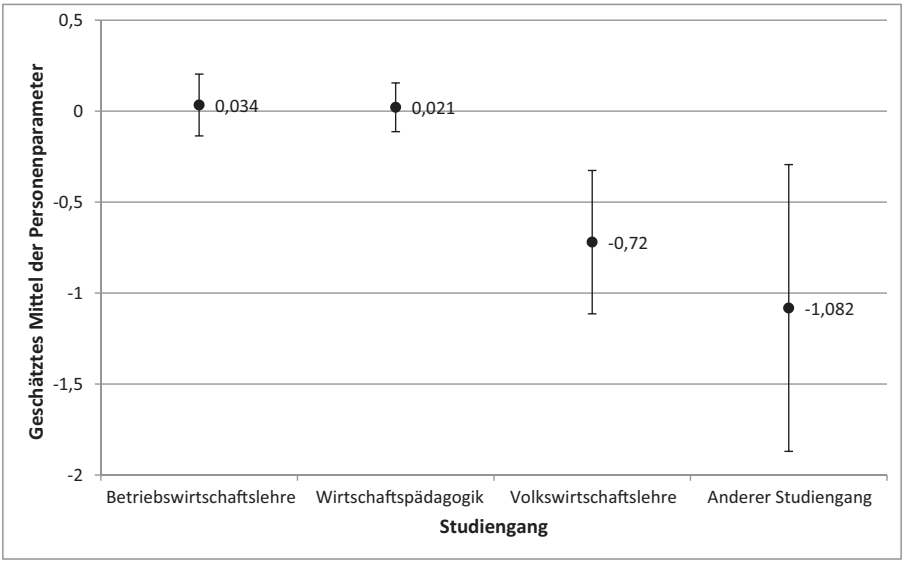


Abb. F-2: Unterschiede in der Testleistung in Abhängigkeit des Studiengangs unter Kontrolle der Note der Hochschulzugangsberechtigung (Note = 2.42) nach Reduktion des Tests um „unfaire" Items, Mittelwerte und 95 %-Konfidenzintervalle sind angegeben

Tab. F-4: Unterschiede zwischen nicht-situativen und situativen Items bezüglich der Bewertung der Facetten der Testakzeptanz sowie der curricularen Validität (ANOVA unter Kontrolle der prozentualen Lösungshäufigkeiten = 59.59)

	Aufgabentyp	***N***	***M***	***SE***	***F***	***df***	***p***
Kontrollierbarkeit*	nicht-situativ	34	4.83	.168	5.331	1	.02
	situativ	34	4.26	.168			
Messqualität*	nicht-situativ	34	3.38	.142	3.580	1	.06
	situativ	34	3.98	.142			
Augenscheinvalidität (beruflich)*	nicht-situativ	34	2.71	.159	0.719	1	.38
	situativ	34	2.96	.159			
Belastungsfreiheit*	nicht-situativ	34	4.19	.180	0.004	1	.95
	situativ	34	4.17	.180			
Curriculare Validität	nicht-situativ	34	4.61	.177	0.025	1	.88
	situativ	34	457	.177			
* Skalendokumentation kann bei Kersting (2008) entnommen werden							

Tab. F-5: Skalendokumentation für die Dimension „curriculare Validität", Skala von 1 bis 6, eigene Entwicklung

Itemtext	***M***	***SD***	***rit***
Der Test enthält Fragen, die in meinem Studium relevant sind.	4.83	1.17	.84
Die Testaufgaben erfassen Wissen und Fähigkeiten, die im Studium gebraucht werden.	4.69	1.21	.79
Die Testaufgaben sind für mein Studium relevant.	4.69	1.19	.82
Die Inhalte der Testaufgaben waren mir aus dem Studium bekannt.	4.39	1.23	.76
Der Test bildet die Inhalte aus dem Studium hinreichend ab.	3.59	1.38	.58
Die Inhalte des Tests haben mit meinem Studium nichts zu tun. [Item rekodiert]	5.19	1.23	.74
N ***M*** ***SD*** **Minimum/Maximum** **Cronbachs Alpha**	64 4.56 0.54 3.59/5.19 .91		

Tab. F-6: Skalendokumentation für die Skala „Anstrengungsbereitschaft", Skala von 1 bis 4, adaptiert nach Kunter (2002, S. 202)

Itemtext	***M***	***SD***	***rit***
Wie sorgfältig haben Sie die Aufgaben im Test bearbeitet?	2.91	0.82	.58
Wie viel Mühe haben Sie sich bei der Bearbeitung des Tests gegeben?	2.82	0.86	.73
Waren Sie bei der Bearbeitung des Tests abgelenkt [Item rekodiert]	3.13	0.88	.43
N ***M*** ***SD*** **Minimum/Maximum** **Cronbachs Alpha**	66 2.96 0.16 2.82/3.14 .74		

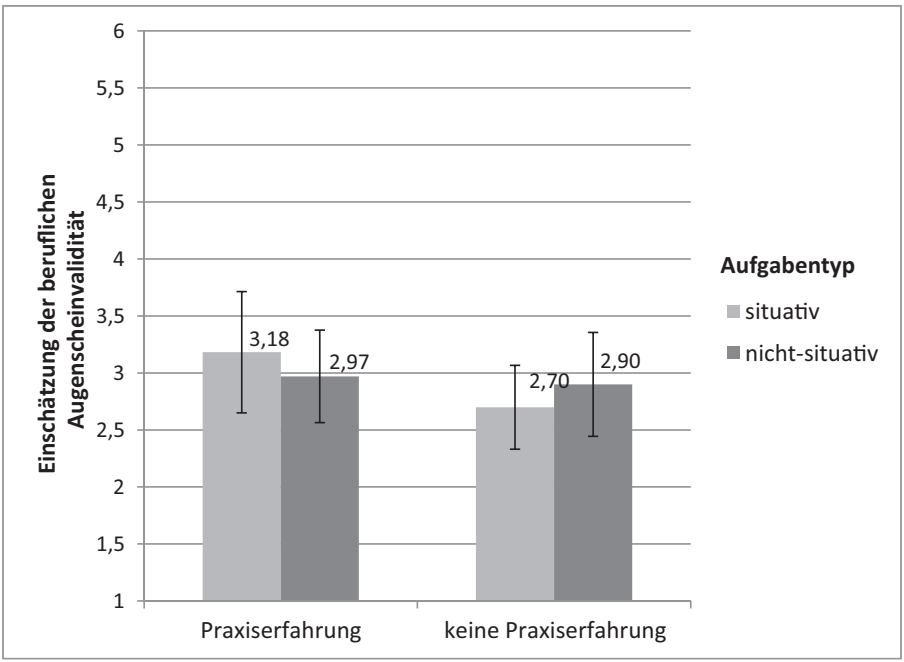


Abb. F-3: Darstellung der Einschätzung der Augenscheinvalidität in Abhängigkeit der durch Berufsausbildung, Praktika oder Nebentätigkeiten erworbenen betriebswirtschaftlichen Praxiserfahrung (95 %-Konfidenzintervall und Mittelwerte sind angegeben)

Berufsbildung, Arbeit und Innovation

Dissertationen/Habilitationen

Die Messung betriebswirtschaftlichen Wissens von Studierenden

Die Frage, was an deutschen Universitäten gelehrt und gelernt wird, rückt immer mehr in den Fokus der Öffentlichkeit. Im Gegensatz zum öffentlichen Interesse liegen aus wissenschaftlicher Perspektive zu Prozessen des Wissenserwerbs und der Wissensentwicklung im tertiären Bildungssektor nur wenige belastbare Ergebnisse vor. Um einen Beitrag zur Schließung dieser Forschungslücke zu leisten, wird in der vorliegenden Dissertation die wissenschaftlich fundierte Entwicklung und Validierung eines Tests beschrieben, der betriebswirtschaftliche Wissensbestände von Studierenden auf Bachelorniveau erfasst. Dabei werden Konzepte der Kompetenzmessung in der beruflichen Bildung und der pädagogisch-psychologischen Diagnostik an Fragestellungen der Erfassung von Lernergebnissen an Hochschulen angepasst und auf diese angewendet.
Durch die Fokussierung auf die Rolle situativer Testaufgaben ist das Buch ein Novum in der deutschen hochschulischen Bildungsforschung. Es liefert wichtige Erkenntnisse über die Struktur und die Determinanten betriebswirtschaftlichen Wissens. Darüber hinaus werden Besonderheiten der Erfassung von Lernergebnissen im Hochschulsektor herausgearbeitet, kritisch diskutiert und in den internationalen Kontext eingebettet.

Christine Caroline Jähnig
ist seit Juni 2010 wissenschaftliche Mitarbeiterin an der Professur für Wirtschaftspädagogik und Personalentwicklung der Georg-August-Universität Göttingen. Zuvor absolvierte sie ein Psychologiestudium und war als wissenschaftliche Mitarbeiterin in einem europäischen Forschungsprojekt zum Einfluss von Emotionen auf Finanzentscheidungen am Forschungszentrum Informatik am Karlsruher Institut für Technologie tätig.

ISBN 978-3-7639-5394-3

wbv