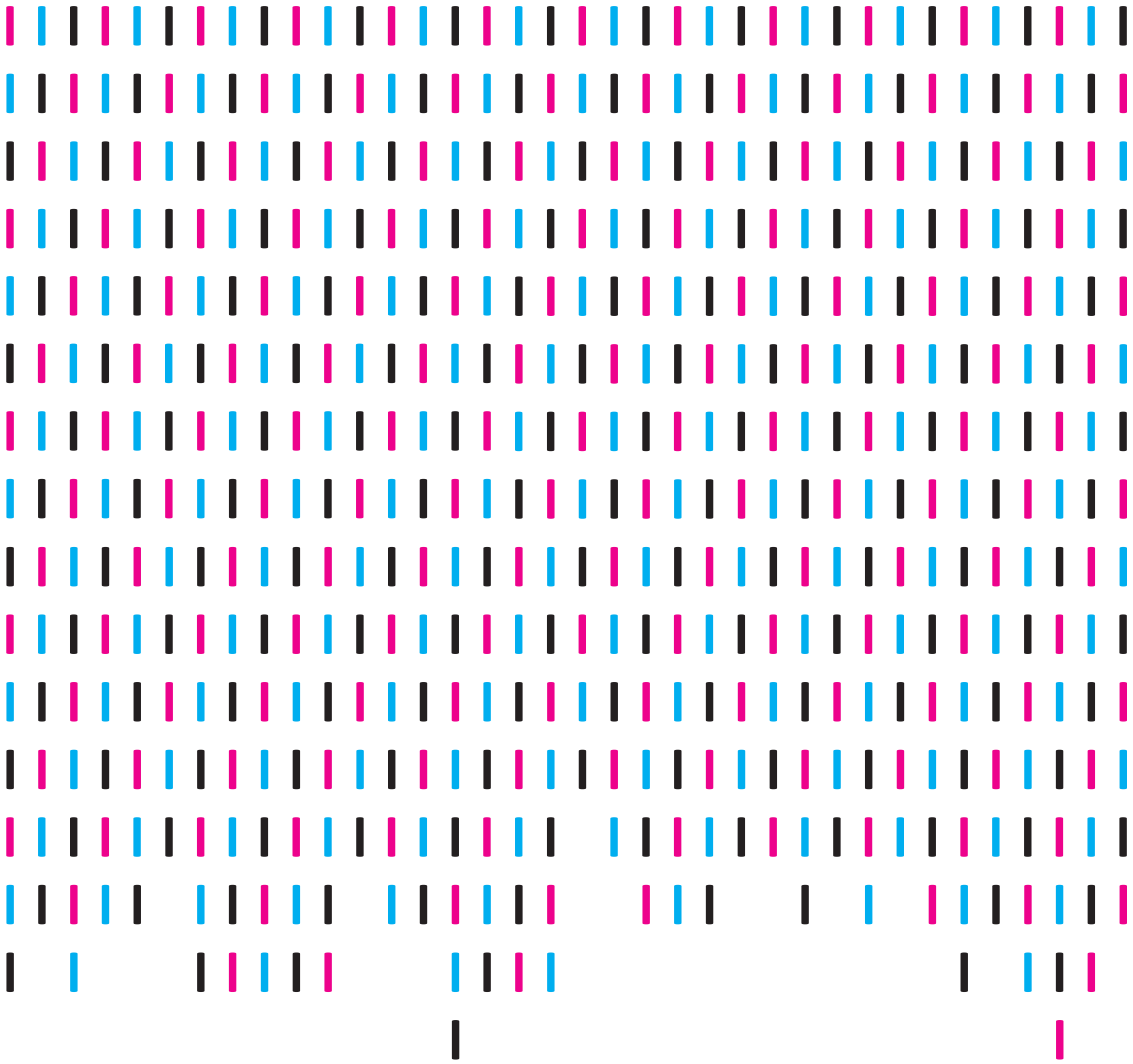

NEPS WORKING PAPERS

Oliver Lauterbach

ERFASSUNG WIRTSCHAFTS-WISSENSCHAFTLICHER FACH-KOMPETENZEN VON STUDIEREN-DEN IN STARTKOHORTE 5 DES NATIONALEN BILDUNGSPANELS – TECHNISCHER BERICHT

Aktualisierung NEPS Working Paper No. 51
Bamberg, September 2016

LIfBi
LEIBNIZ-INSTITUT FÜR BILDUNGSVERLÄUFE e.V.

NEPS
National Educational Panel Study

Working Papers of the German National Educational Panel Study (NEPS)
at the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg

The NEPS Working Papers publish articles, expertises, and findings related to the German National Educational Panel Study (NEPS).

The NEPS Working Papers are edited by a board of researchers representing the wide range of disciplines covered by NEPS. The series started in 2011.

Papers appear in this series as work in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character.

Any opinions expressed in this series are those of the author(s) and not those of the NEPS Consortium.

The NEPS Working Papers are available at **www.neps-data.de** (see section "Publications").

Contact: German National Educational Panel Study (NEPS) – Leibniz Institute for Educational Trajectories – Wilhelmsplatz 3 – 96047 Bamberg – Germany – contact@lifbi.de

Erfassung wirtschaftswissenschaftlicher Fachkompetenzen von Studierenden in Startkohorte 5 des Nationalen Bildungspanels – Technischer Bericht

Oliver Lauterbach, DZHW – Deutsches Zentrum für Hochschul- und Wissenschaftsforschung GmbH, Hannover

Kontakt:

hochschulstudium@lifbi.de

Bibliographische Angabe:

Lauterbach, O. (2015). Erfassung wirtschaftswissenschaftlicher Fachkompetenzen von Studierenden in Startkohorte 5 des Nationalen Bildungspanels – Technischer Bericht (Aktualisierung NEPS Working Paper No. 51). Bamberg: Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel.

Erfassung wirtschaftswissenschaftlicher Fachkompetenzen von Studierenden in Startkohorte 5 des Nationalen Bildungspanels – Technischer Bericht

Abstract

Das Nationale Bildungspanel (NEPS) untersucht die Entwicklung sowohl fachlicher als auch überfachlicher Kompetenzen im Lebenslauf. In Startkohorte 5 werden in dieser Teilstudie fachspezifische Kompetenzen von Studierenden der Wirtschaftswissenschaften mit einem eigens für das NEPS entwickelten Instrument erfasst. Der vorliegende Bericht gibt einen Überblick über einzelne Schritte der Testentwicklung sowie über die in der Haupterhebung erhobenen Daten und deren Skalierung. Die Inhaltsvalidität der Aufgaben wurde in Vorstudien anhand von curricularen Analysen und Expertenratings überprüft. Im Anschluss wurden die psychometrischen Eigenschaften der Aufgaben in zwei Entwicklungsstudien mit zusammen über 7.000 Probanden bestimmt. In der Haupterhebung wurden 36 ausgewählte Aufgaben verschiedener wirtschaftswissenschaftlicher Fachgebiete eingesetzt und Daten von 338 Teilnehmern erhoben, die am Ende ihres wirtschaftswissenschaftlichen Studiums standen. Anhand eines Rasch Modells wurden Item-Kennwerte, Differential Item Functioning, Dimensionalität sowie lokale stochastische Unabhängigkeit untersucht. Auch wenn einige Aufgaben nicht von allen Probanden in der vorgegebenen Testdauer erreicht wurden und durch den Einbezug verschiedener wirtschaftswissenschaftlicher Fachgebiete die dimensionale Struktur nicht eindeutig als eindimensional identifiziert werden konnte, weist der Test eine hinreichende psychometrische Qualität auf, um eine valide Erfassung wirtschaftswissenschaftlicher Kompetenz von Bachelor-Absolventen zu erlauben.

Schlagworte

Item Response Theorie, Skalierung, wirtschaftswissenschaftliche Fachkompetenz, Scientific Use File, Studierende

Abstract

The National Educational Panel Study (NEPS) aims to investigate the development of domain-specific as well as domain-general competencies across the whole life span. In starting cohort 5, subject-specific competencies of students of economics and business administration are measured by a short instrument specifically designed for this NEPS substudy. This report gives an overview of the instrument development process and describes the data of the main study and scaling results. Content validity of the test items was secured by analyses of curricula and expert ratings. Psychometric properties of the items were tested in two developmental studies with over 7,000 participants. In the main study, a selection of 36 items was used to test 338 participants who were at the end of their studies of economics or business administration. A Rasch model was used to assess item statistics, differential item functioning, dimensionality, and local item independence. Although several items were not reached by all participants, and the dimensionality of the instrument was not clearly one-dimensional because of the underlying economic subjects, the general psychometric quality of the

instrument was satisfying. Altogether, the test allows drawing valid conclusions on economic competencies of students at the end of their bachelor's degree program.

Keywords

Item response theory, scaling, economic competence, Scientific Use File, higher education students

Inhalt

1. Einleitung

Neben allgemeinen und übergreifenden Kompetenzen sollen in Startkohorte 5 des Nationalen Bildungspanels, die vom Arbeitsbereich „Hochschulstudium und Übergang in den Beruf“ (Etappe 7) mit konzipiert wird, auch fachspezifische Kompetenzen von Studierenden gemessen werden (Aschinger et al., 2011). Als größte Studierendengruppe wurden dafür zunächst die Studierenden der Wirtschaftswissenschaften ins Auge gefasst. Bislang existierte jedoch kein passendes Instrument zur Erfassung wirtschaftswissenschaftlicher Fachkompetenz von Hochschulabsolventen. Bestehende Instrumente fokussieren entweder andere Zielgruppen wie z. B. Schulabgänger oder Studienanfänger (Beck & Krumm, 1998), oder sie liegen nicht in deutscher Sprache vor. Diese Lücke zu schließen war Ziel des WiwiKom-Projekts der Universität Mainz (Zlatkin-Troitschanskaia, Förster, Brückner, Hansen & Happ, 2013; Zlatkin-Troitschanskaia, Förster, Brückner & Happ, 2014). In Zusammenarbeit mit diesem Projekt hat Etappe 7 eine Kurzform eines wirtschaftswissenschaftlichen Kompetenztests erstellt, die bei Studierenden der Startkohorte 5 in der Haupterhebung B90 eingesetzt wurde. Im ersten Teil des Berichts wird auf die Entwicklungsarbeiten und die inhaltliche Struktur des Tests eingegangen, im zweiten Teil werden die psychometrischen Eigenschaften des Tests in der Haupterhebung B90 im Jahr 2014 dargestellt.

2. Testentwicklung

Die Entwicklung fand in Kooperation mit dem Projekt WiwiKom der Universität Mainz statt, das unter anderem die Adaption der international etablierten wirtschaftswissenschaftlichen Kompetenztests EGEL (Exámenes Generales de Egreso de la Licenciatura; CENEVAL, 2011) und TUCE (Test of Understanding in College Economics; Walstad, 2007) zum Ziel hatte. Die Übersetzung der Testaufgaben der mexikanischen bzw. US-amerikanischen Originalinstrumente führte das WiwiKom-Projekt durch. Etappe 7 des Nationalen Bildungspanels übernahm Analysen der curricularen Struktur wirtschaftswissenschaftlicher Studiengänge in Deutschland, ein Online-Expertenrating der übersetzten Testaufgaben sowie Teile der beiden empirischen Entwicklungsstudien. Die Aufgaben der in der Haupterhebung B90 eingesetzten Kurzversion wurden von Etappe 7 mit Unterstützung des WiwiKom-Projekts ausgewählt.

2.1 Analyse von Modulbeschreibungen wirtschaftswissenschaftlicher Studiengänge

Um einen Überblick über die zentralen Lehrinhalte des Fachgebiets zu erlangen und die Passung der zu adaptierenden Instrumente an deutsche Curricula prüfen zu können, wurden Modulhandbücher wirtschaftswissenschaftlicher Studiengänge herangezogen. Neben den behandelten Inhalten sind in den Modulhandbüchern die zu vermittelnden Kompetenzen, Pflicht- bzw. Wahlstatus sowie häufig die zeitliche Anordnung im Studienverlauf dokumentiert.

Es wurden insgesamt 96 Bachelor-Studiengänge an 40 Universitäten und 23 Fachhochschulen erfasst, darunter betriebswirtschaftliche, wirtschaftswissenschaftliche und volkswirtschaftliche Studiengänge. Während BWL-Studiengänge häufiger an Fachhochschulen (62 %) als an Universitäten (38 %) studiert werden können, werden Studiengänge der Wirtschaftswissenschaften eher an Universitäten (75 %) und volkswirtschaftliche Studiengänge fast ausschließlich an Universitäten angeboten (94 %). In der Regel unterscheiden sich wirtschaftswissenschaftliche Studiengänge dadurch von reinen BWL- oder VWL-Studiengängen,

dass sich die Studierenden nach einem gemeinsamen Grundlagenstudium in BWL und VWL im weiteren Verlauf in einem der beiden Fächer spezialisieren können.

Zentrale Ergebnisse der Analysen sind die hohe Heterogenität der Studienangebote sowohl zwischen verschiedenen Fächern als auch zwischen den verschiedenen Studiengängen innerhalb eines Fachs. So werden einzelne Inhaltsbereiche in sehr unterschiedlichem Umfang gelehrt, weiterhin bestehen durch zahlreiche Wahlmöglichkeiten unterschiedliche Profilbildungen innerhalb eines Studiengangs. Beispielsweise schwanken die Anteile von Lehrangeboten aus dem Bereich Organisation und Management auch innerhalb der BWL-Studiengänge von unter 5 % bis zu 20 % und Inhalte aus dem Bereich Personalwesen werden nicht an allen Hochschulen angeboten.

Andererseits kann festgehalten werden, dass es hinreichende inhaltliche Schnittmengen zwischen betriebswirtschaftlich, wirtschaftswissenschaftlich und volkswirtschaftlich ausgerichteten Studiengängen gibt. Während das Fach Wirtschaftswissenschaften per se durch die Kombination betriebs- und volkswirtschaftlicher Inhalte gekennzeichnet ist, kann darüber hinaus davon ausgegangen werden, dass Absolventen eines betriebswirtschaftlichen Studiengangs grundlegende Kenntnisse der Mikro- und Makroökonomie und andersherum Absolventen der Volkswirtschaftslehre zumindest Grundkenntnisse im Bereich Marketing oder Finanzierung besitzen sollten.

2.2 Übersetzung und Adaption der Originalinstrumente

Ziel des WiwiKom-Projekts ist die Adaption international etablierter wirtschaftswissenschaftlicher Instrumente an deutsche Gegebenheiten. Eher betriebswirtschaftlich ausgerichtet ist der in Mexiko entwickelte EGEL (Exámenes Generales de Egreso de la Licenciatura), der die Bereiche Management (Administración) und Rechnungswesen (Contaduría) umfasst. Die Aufgaben wurden in Zusammenarbeit mit Praktikern der Berufsfelder entwickelt und die entsprechenden Tests sind als Eignungsprüfungen für Absolventen der Betriebswirtschaftslehre in Mexiko sehr verbreitet. Der EGEL-Administración umfasst im Original 250 Aufgaben. Da in den vorhergehenden curricularen Analysen bereits festgestellt wurde, dass Rechnungswesen in Deutschland einen wichtigen Inhaltsbereich wirtschaftswissenschaftlicher Studiengänge darstellt, wurden zusätzlich 47 Aufgaben des EGEL-Contaduría in den Itempool aufgenommen. Den Bereich der Volkswirtschaftslehre erfasst der US-amerikanische TUCE (Test of Understanding in College Economics; Walstad, 2007). Jeweils 30 Aufgaben decken die Bereiche Mikro- und Makroökonomie ab. Sowohl bei den Aufgaben des EGEL als auch bei denen des TUCE handelt es sich um Aufgaben mit geschlossenem Antwortformat (Multiple Choice), bei denen von vier Antwortmöglichkeiten jeweils nur eine richtig ist. Das Format wurde bei der Adaption beibehalten.

Die Übersetzung der Aufgaben wurde von Mitarbeitern des Fachbereichs Translations-, Sprach- und Kulturwissenschaft der Universität Mainz durchgeführt. Bereits im Übersetzungsprozess wurden Aufgaben identifiziert, die sich schlecht an deutsche Gegebenheiten anpassen ließen (beispielsweise weil sie sich auf spezifische rechtliche Vorschriften im Ausland beziehen) und deshalb aus der folgenden empirischen Prüfung ausgeschlossen wurden. Dies traf ausschließlich für Aufgaben des mexikanischen Instruments zu; der amerikanische TUCE konnte dagegen komplett übernommen werden.

2.3 Expertenrating der Testaufgaben

Das als Online-Befragung durchgeführte Expertenrating diente der Prüfung der Aufgaben hinsichtlich ihrer Eignung, die Lehrinhalte an deutschen Hochschulen abzubilden sowie die Relevanz entsprechender Kompetenzen für Studium und Beruf zu erfassen. Weiterhin sollte eine erste Einschätzung der Aufgabenschwierigkeit erfolgen und es bestand für die Experten die Möglichkeit, Übersetzungsprobleme und unklare Aufgabenstellungen in einem offenen Kommentarfeld anzumerken. In den Expertenbefragungen wurden jeweils maximal 20 Aufgaben zusammengefasst, um die Experten zeitlich nicht zu überfordern. Die Aufgaben der einzelnen Inhaltsbereiche wurden von Experten hinsichtlich (1.) der Verbreitung der Inhalte in der Lehre, (2.) der Schwierigkeit der Aufgabe, (3.) der Relevanz für das Studiengebiet und (4.) die spätere Berufstätigkeit sowie (5.) der Eignung insgesamt bewertet. Insgesamt wurden 225 Aufgaben aus dem EGEL-Test und alle 60 Aufgaben des TUCE zur Beurteilung vorgelegt.

Von April bis Juni 2012 wurden insgesamt 250 Experten zur Teilnahme am Online-Rating der Testaufgaben eingeladen. Die Experten waren Professoren und Professorinnen, die in den entsprechenden Inhaltsbereichen der Wirtschaftswissenschaften an deutschen Hochschulen lehren und dem WiwiKom-Projektteam persönlich bekannt waren oder in der Phase der Nachrekrutierung über die Internetseiten der Hochschulen recherchiert wurden. Die Einladung erfolgte per E-Mail, teilweise mit persönlicher oder telefonischer Vorankündigung. In wenigen Fällen wurden auch wissenschaftliche Mitarbeiter kontaktiert bzw. die Einladungsmail wurde von den Angeschriebenen an diese weitergeleitet.

Von 250 eingeladenen Experten haben 85 (34 %) Angaben zu den präsentierten Aufgaben gemacht. Mehr als 80 % der Items aller Inhaltsbereiche erhalten eine mindestens mittlere Bewertung hinsichtlich der Kriterien Verbreitung in der Lehre und Verbreitung im Studium. Bezogen auf die einzelnen Inhaltsbereiche schwanken die Anteile zwischen 55 % (Organisation) und 97 % (Mikro- und Makroökonomie). Aufgaben, die sich hinsichtlich der erfassten Kriterien als problematisch erwiesen haben, wurden mit Experten der jeweiligen Fachgebiete diskutiert und falls nötig überarbeitet oder ausgeschlossen. In die folgende empirische Prüfung im Rahmen der ersten Entwicklungsstudie wurden 150 Aufgaben aus dem EGEL und die kompletten 60 Aufgaben des TUCE übernommen.

2.4 Entwicklungsstudien

Es wurden zwei empirische Erprobungen der Testaufgaben an verschiedenen Hochschulstandorten durchgeführt. Über Anschreiben an die Hochschulleitungen und über persönliche Kontakte wurden Lehrende wirtschaftswissenschaftlicher Studiengänge gebeten, 45 Minuten ihrer Lehrveranstaltung für die Durchführung der Erhebung zur Verfügung zu stellen. Die Testaufgaben wurden nach einem Booklet-Design (vgl. Frey, Hartig & Rupp, 2009) in Clustern zu jeweils 9 oder 10 Aufgaben auf verschiedene Booklets mit jeweils 27 bis 30 Aufgaben verteilt. Jedes Cluster stand dabei zur Kontrolle von Ermüdungseffekten sowohl am Anfang als auch in der Mitte und am Ende eines Testhefts. Es wurden sowohl inhaltshomogene als auch ein inhaltsheterogenes Booklet eingesetzt. Die inhaltshomogenen Booklets erlauben die Überprüfung der psychometrischen Eigenschaften der Aufgaben jeweils eines Inhaltsbereichs, die inhaltsheterogenen Booklets die Berechnung der Korrelationen der Inhaltsbereiche anhand ausgewählter Aufgaben verschiedener Inhaltsbereiche. Zusätzlich zu den Testaufgaben wurden den Testpersonen Fragen zu relevanten Personenmerkmalen, Vorbildung

und Studienerfahrungen vorgelegt, um mit diesen Daten die Validierung der Testergebnisse zu ermöglichen.

2.4.1 Erste Entwicklungsstudie

Im Wintersemester 2012 fand die erste empirische Erprobung der Testaufgaben statt. Es wurden Erhebungen an 15 Universitäten und 8 Fachhochschulen durchgeführt. Insgesamt wurden in der ersten Entwicklungsstudie 220 Aufgaben verteilt auf 5 Booklets und 43 verschiedene Testhefte eingesetzt. Nach Ausschluss von Personen, die nicht zur Zielpopulation gehören (Studierende anderer Fachrichtungen und Gasthörer), liegen Daten von 3580 Studierenden vor. Die durchschnittliche Studiendauer der Teilnehmer und Teilnehmerinnen beträgt 3,8 Fachsemester. Weitere Stichprobenmerkmale sind in Tabelle 1 dargestellt.

2.4.2 Vorauswahl der Aufgaben

Die Aufgaben wurden auf Basis der klassischen Testtheorie sowie Item-Response-Theorie auf ihre Schwierigkeit, interne Konsistenz und Kriteriumsvalidität untersucht. Als Kriterien für curriculare Validität wurden der Studienfortschritt (Studienanfänger bzw. Studienanfängerinnen versus fortgeschrittene Studierende) sowie der Besuch einschlägiger Veranstaltungen bezogen auf den Aufgabeninhalt herangezogen. Zusätzlich wurden die Trennschärfe und Besetzung der Distraktoren (falsche Antworten) der Aufgaben untersucht.

Aufgaben wurden als problematisch angesehen, wenn sie (1.) extrem schwer ($p < .20$) oder leicht ($p > .80$) waren oder (2.) eine nur geringe oder negative Trennschärfe bezogen auf die anderen Aufgaben des Inhaltsbereichs aufwiesen oder wenn sich (3.) ein signifikanter Outfit (MNSQ > 1.2) bzw. ein signifikanter T-Wert (> 1.96) ergab. Weiterhin wurden Aufgaben besonders in den Blick genommen, die nicht oder negativ mit dem Studienfortschritt oder Veranstaltungsbesuch korrelieren oder deren Distraktoren positiv mit dem Gesamtscore korrelieren. Solche Aufgaben wurden entweder eliminiert oder erneut mit Experten des Fachgebiets diskutiert und überarbeitet.

Besonderes Ziel der ersten Entwicklungsstudie war es, im Hinblick auf den Einsatz eines Kurztests im Rahmen der NEPS-Studie eine Vorauswahl von Aufgaben zu treffen, die in der zweiten Entwicklungsstudie in einem gemeinsamen Korrelationsbooklet eingesetzt werden sollten. Hierfür wurden nach den berichteten Gütekriterien jeweils 10 Aufgaben aus den Bereichen Marketing, Organisation, Personal, Mikro- und Makroökonomie sowie jeweils 9 Aufgaben aus den Bereichen Finanzierung und Rechnungswesen ausgewählt. Dabei wurde berücksichtigt, dass die NEPS-Haupterhebung sich an fortgeschrittene Studierende und Absolventen richtet, weshalb für die Bewertung der Aufgabenschwierigkeit nur Daten von Studierenden ab dem vierten Fachsemester herangezogen wurden.

2.4.3 Zweite Entwicklungsstudie

Im Sommersemester 2013 fand die zweite empirische Erprobung der zum Teil überarbeiteten und neu zusammengestellten Testaufgaben statt. Es wurden Erhebungen an 18 Universitäten und 7 Fachhochschulen durchgeführt. Insgesamt wurden in der zweiten Entwicklungsstudie 205 Aufgaben eingesetzt, die auf 4 Booklets und 42 verschiedene Testhefte verteilt waren. Nach Ausschluss von Personen, die nicht zur Zielpopulation gehören (Studierende anderer Fachrichtungen und Gasthörer), liegen Daten von 3512 Personen vor. Die durchschnittliche Studiendauer der Teilnehmer und Teilnehmerinnen beträgt 3,7 Fachsemester. Weitere Stichprobenmerkmale sind in Tabelle 1 dargestellt.

Tabelle 1: Stichprobenbeschreibung Entwicklungsstudien

		ES I (WS 2012)		ES II (SoSe 2013)	
		n	%	n	%
Geschlecht	weiblich	1792	49.3	1697	48.4
	männlich	1846	50.7	1812	51.6
Fachsemester	1-2	862	24.6	1337	39.2
	3-4	1321	37.8	1264	37.1
	5-6	871	24.9	567	16.6
	7 oder höher	443	12.7	242	7.1
Studiengang	Wirtschaftswissenschaften	1463	41.4	1165	33.6
	Betriebswirtschaftslehre	1294	36.6	1312	37.8
	Volkswirtschaftslehre	220	6.2	408	11.8
	Wirtschaftspädagogik	422	11.9	359	10.4
	spezielle wirtschaftswissenschaftliche Studiengänge	138	3.9	223	6.4
	andere	-	-	-	-
Hochschultyp	Universität	3008	84.0	3005	85.6
	Fachhochschule	572	16.0	507	14.4
Wirtschaftswissenschaftliches Studium als	Hauptfach	3357	94.7	3306	95.5
	Nebenfach	169	4.8	139	4.0
	sonstiges	19	0.5	15	0.4
Angestrebter Abschluss	Bachelor	3230	91.1	3339	96.6
	Master	245	6.9	109	3.2
	Diplom/Magister	70	2.0	7	0.2
	Staatsexamen	-	-	-	-
Muttersprache	deutsch	3018	84.5	3020	86.0
	nichtdeutsch	553	15.5	490	14.0
Berufsausbildung	keine	2694	75.3	2693	77.2
	kaufmännische Berufsausbildung	728	20.4	644	18.5
	andere Berufsausbildung	154	4.3	150	4.3
gesamt		3580	100.0	3512	100.0

2.4.4 Zusammenstellung des B90-Testhefts

Für das Testheft der Haupterhebung B90 wurden jene Aufgaben erneut untersucht, die bereits in der ersten Entwicklungsstudie in die Vorauswahl genommen wurden. Die Analysen zur Dimensionalität beziehen sich auf die Aufgaben des sogenannten Korrelationsbooklets, für weitere Analysen wurde jedoch auch auf Daten zurückgegriffen, die für diese Aufgaben in anderen Booklets erhoben wurden.

Unter psychometrischen Gesichtspunkten ungünstige Aufgaben wurden nach denselben Kriterien wie bereits in der ersten Entwicklungsstudie identifiziert. Erneut war ein Schwierigkeitsgrad zwischen $p < .20$ und $p > .80$ für Studierende höherer Semester Ausschlusskriterium, wobei zusätzlich darauf geachtet wurde, möglichst unterschiedliche Schwierigkeitsniveaus in den einzelnen Inhaltsbereichen abzudecken. Aufgrund der generell eher geringen Homogenität der Inhaltsbereiche wurde eine Trennschärfe von mindestens .10 gefordert. Weitere Ausschlusskriterien waren Kennwerte nach IRT: ein signifikanter Outfit (MNSQ > 1.2) bzw. ein signifikanter T-Wert (> 1.96). Zusätzlich wurde darauf geachtet, dass innerhalb eines Inhaltsbereichs keine Aufgaben aufgenommen wurden, die verwandte Fachinhalte erfragen.

Für die Auswahl der Aufgaben wurden außerdem DIF-Analysen durchgeführt. Ziel war es, keine Aufgaben in das endgültige Testheft aufzunehmen, die deutliche Gruppenunterschiede hinsichtlich Geschlecht, Hochschultyp oder Muttersprache aufwiesen. Dies bedeutet nicht, dass nicht über alle Aufgaben hinweg solche Gruppenunterschiede festgestellt werden können, lediglich sollten keine Einzelaufgaben auf besondere Weise hinsichtlich dieser Gruppen differenzieren (beispielsweise durch geschlechtsspezifische Aufgabeninhalte oder ungebräuchliche Formulierungen, die Nicht-Muttersprachler benachteiligen). Ließ sich die Aufnahme einer Aufgabe mit DIF nicht vermeiden, wurde alternativ darauf geachtet, mit einer weiteren Aufgabe den Gruppenunterschied auszubalancieren.

Inhaltsanalysen der Aufgaben sowie der Curricula der Studiengänge haben ergeben, dass die Fachgebiete Organisation bzw. Unternehmensführung und Personalwesen als relativ verwandt angesehen werden können. Aus diesem Grund und um eine Bearbeitungszeit von maximal 45 Minuten einzuhalten, wurden diese Inhaltsbereiche für das B90-Testheft zusammengefasst. Das finale Testheft enthält insgesamt 36 Aufgaben, davon jeweils 6 aus den Bereichen (1.) Marketing, (2.) Organisation/Unternehmensführung/Personal, (3.) Finanzierung, (4.) Rechnungswesen, (5.) Mikroökonomie und (6.) Makroökonomie.

Die Dimensionalität der für die B90 ausgewählten Aufgaben wurde anhand mehrdimensionaler IRT-Modelle (1PL) überprüft. Hier konnte gezeigt werden, dass die sechs repräsentierten Inhaltsdimensionen mittel bis hoch (.53 bis .89) positiv korreliert sind. Beim Vergleich des sechsdimensionalen Modells mit einem eindimensionalen Modell konnte zwar festgestellt werden, dass das sechsdimensionale Modell die Daten signifikant besser abbildet als das eindimensionale Modell (ΔDeviance = 35.43, $p < .05$), jedoch sprechen AIC und BIC für das im Sinne der Modellkomplexität sparsamere Modell (eindimensionales Modell: AIC = 16942.52, BIC = 17120.17; sechsdimensionales Modell AIC = 16947.10, BIC = 17220.77). Insgesamt deutet dies darauf hin, dass die Bildung eines Gesamttestwerts wirtschaftswissenschaftlicher Fachkompetenz zulässig ist.

3. Haupterhebung B90

3.1 Studiendesign

Die Haupterhebung B90 wurde im Rahmen von Einzelinterviews an einem vom Studienteilnehmer zu bestimmenden Ort (vorzugsweise in der eigenen Wohnung oder am Arbeitsplatz) durchgeführt. Die eigentliche Testung wurde ergänzt um ein computerbasiertes Face-to-face-Interview (CAPI, Computer Assisted Personal Interview), in dem unter anderem aktuelle Informationen zum Studienstatus und zu inhaltlichen Schwerpunkten im Studium sowie Selbsteinschätzungen des Erfolgs in der Beantwortung der Aufgaben erhoben wurden.

Zum Einsatz kam ein papierbasiertes Testheft mit 36 Multiple-Choice-Aufgaben aus sechs wirtschaftswissenschaftlichen Fachgebieten (s. Kap. 2.4.4). Es wurde nur eine Testheftversion eingesetzt, so dass die Aufgaben immer dieselbe Position im Testheft hatten. Die Aufgaben waren zur einfacheren Bearbeitung nach Fachgebieten angeordnet. Die maximale Bearbeitungszeit für die Testaufgaben betrug 45 Minuten.

3.1.1 Einsatzstichprobe

Die Zielpopulation der Haupterhebung B90 bestand aus Teilnehmern der Startkohorte 5, die in einer der vorangehenden Befragungen angegeben haben, ein wirtschaftswissenschaftliches Fach (Betriebswirtschaftslehre, Wirtschaftswissenschaften, Volkswirtschaftslehre, Europäische Wirtschaft, Internationale BWL oder internationales Management, Wirtschaftspädagogik, Verkehrs- oder Tourismuswirtschaft) zu studieren. Da die Erhebung als Individualtestung durch Interviewer vor Ort durchgeführt werden musste und aus Kostengründen die Testung nur an wenigen Orten stattfinden konnte, wurde vom Erhebungsinstitut auf Grundlage der Wohnorte der Zielpersonen eine Substichprobe von 15 regionalen Clustern gezogen. Aus der Auswahlgesamtheit von 1701 Fällen wurden so 601 Personen für die Einsatzstichprobe der Kompetenztestung berücksichtigt (Prussog-Wagner & Aust, 2014).

3.1.2 Realisierte Stichprobe

Innerhalb von 10 Wochen Feldzeit konnten Interviews mit 342 Teilnehmern (57 % der Einsatzstichprobe) realisiert werden. Von den 338 auswertbaren Fällen[1] sind 148 männlich (44 %) und 190 weiblich (56 %). Die Mehrheit von 287 Personen (85 %) studiert aktuell, 31 Personen (9 %) haben ihr Studium abgeschlossen und 20 Personen (6 %) haben das Studium unterbrochen oder aufgegeben. 236 Personen (70 %) studieren oder studierten in einem Bachelorstudiengang, 76 (22 %) in einem Masterstudiengang und 26 (8 %) in einem sonstigen Studiengang.

Ihren letzten wirtschaftswissenschaftlichen Studienabschnitt haben 230 Personen (68 %) an einer Universität und 105 Personen (31 %) an einer Fachhochschule (inkl. Duale Hochschule) absolviert. Die Bildungsherkunft wird nicht wie sonst üblich anhand der Bücherfrage, sondern anhand der Abschlüsse der Eltern bestimmt. Von 155 Personen (46 %) besitzt mindestens ein Elternteil einen Hochschulabschluss, entsprechend haben 183 Personen (54 %) Eltern ohne Hochschulabschluss. Ein Migrationshintergrund liegt bei 53 Personen (16 %) vor: 24 Personen (7 %) sind selbst zugewandert und bei 29 Personen (9 %) ist mindestens ein Elternteil zugewandert.

[1] Vier Testhefte sind nicht auswertbar, so dass 338 Fälle für die folgenden Analysen zur Verfügung stehen.

3.1.3 Dauer und Ort des Interviews

Die gesamte Interviewdauer unter Berücksichtigung der Einführung und der Interviewerfragen betrug durchschnittlich 62.3 Minuten (SD = 13.0); ohne Einleitung und Interviewerfragen liegt die mittlere Länge der Erhebung bei 59.5 Minuten (SD = 12.7). Der größte Teil entfiel dabei mit durchschnittlich 42.0 Minuten (SD = 6.9) auf die Bearbeitung des Testhefts (Bruttobearbeitungszeit einschließlich Unterbrechungen). Das Interview fand in 257 Fällen (76 %) in der Wohnung der Zielperson statt, in 40 Fällen (12 %) in einem Café oder Restaurant, in 23 Fällen (7 %) an der Hochschule, in 10 Fällen (3 %) am Arbeitsplatz und in 8 Fällen (2 %) an einem anderen Ort.

Zu Unterbrechungen der Testung, beispielsweise durch zu versorgende Kleinkinder oder länger andauernde Ablenkung durch andere Personen kam es in 21 Fällen (6 %). Das Zeitlimit für die Testdurchführung wurde in diesen Fällen entsprechend verlängert. Dieses für Testungen vielleicht ungewöhnliche Vorgehen geschah aus der Überlegung, dass es aus Vergleichbarkeitsgründen zielführender ist, die individuelle Bearbeitungszeit des Tests zu standardisieren und nicht die Gesamtdauer des Interviews. Gewährleistet war auf jeden Fall, dass eine Nettobearbeitungszeit von 45 Minuten nicht überschritten wurde.

Zu weniger gravierenden Störungen der Interviewdurchführung ohne zeitliche Kompensation kam es in 31 Fällen (9 %). Dabei handelte es sich meistens um Gespräche (19 Fälle; 6 %), Türklingeln, kurze Telefonate oder Mitbewohner (8 Fälle; 2 %) oder technische Probleme mit dem Laptop (3 Fälle; 1 %). Der größte Teil der Interviews (287 Fälle; 85 %) verlief ohne Unterbrechungen oder Störungen.

3.1.4 Gültige und fehlende Antworten

Insgesamt haben 164 Personen (49 %) alle 36 Aufgaben des Testhefts beantwortet. Die fehlenden Antworten verteilen sich wie in Tabelle 2 dargestellt auf die einzelnen Aufgaben. Unterschieden werden drei Arten fehlendender Antworten: (1.) nicht erreichte Aufgaben, (2.) ausgelassene Aufgaben und (3.) ungültig beantwortete Aufgaben. Zwei Aufgaben wurden von mehr als 10 % der Befragten ausgelassen (bas7fin5, bas7acc3). Dies betrifft die Bereiche Finanzierung und Rechnungswesen und ist vermutlich auf den im Vergleich zu anderen Aufgaben höheren Bearbeitungsaufwand zurückzuführen. Die Anzahl nicht erreichter Aufgaben steigt gegen Ende des Testhefts. Die letzten Aufgaben aus dem Bereich Makroökonomie wurden nur von ca. 70 % der Teilnehmer beantwortet. Ungültige Antworten kamen insgesamt nur in 2 Fällen vor und können deshalb vernachlässigt werden. Es stellt sich jedoch die Frage, wie mit ausgelassenen und nicht erreichten Aufgaben umzugehen ist.

3.1.5 Umgang mit fehlenden Werten

Das Auslassen von Aufgaben kann sowohl auf einfaches Überblättern zurückgeführt werden als auch darauf, dass Aufgaben den Probanden zu schwierig oder zu aufwändig erscheinen. Während das Überblättern einen zufälligen Ausfall darstellt (Missing at random), ist das bewusste Auslassen von Aufgaben abhängig von Personen- und Aufgabenmerkmalen (und damit Missing not at random). Angesichts einer begrenzten Befragungszeit und eines hohen Anteils ausgelassener Aufgaben erscheint es plausibel, dass Aufgaben aus Effizienzüberlegungen heraus ausgelassen wurden. Das Fehlen von Aufgaben am Ende eines Tests ist dagegen nicht abhängig von Aufgabenmerkmalen, jedoch möglicherweise von der Personenfähigkeit, soweit diese die Bearbeitungsgeschwindigkeit beeinflusst. Nach Rohwer (2013) gibt es keine eindeutige Regel zum Umgang mit fehlenden Werten, sondern es ist eine

definitorische Frage, ob man Kompetenz als Problemlösefähigkeit losgelöst von einer Zeitbeschränkung oder unter der Bedingung begrenzter Bearbeitungszeit betrachtet.

Es herrscht daher auch kein einheitlicher Standard zum Umgang mit fehlenden Werten in der Kompetenzmessung. Gängige Verfahren bestehen darin, diese als falsch zu kodieren (z. B. in den PISA-Studien; Adams & Wu, 2002) oder sie bei Multiple-Choice-Aufgaben mit der statistischen Lösungswahrscheinlichkeit zu ersetzen (Lord, 1974). Dahinter steckt die Annahme, dass eine ausgelassene Aufgabe vom Probanden nicht gelöst werden konnte bzw. der Proband bei zufälliger Antwortauswahl eine Lösungswahrscheinlichkeit in Abhängigkeit der Anzahl der Antwortkategorien gehabt hätte. Neuere Studien deuten allerdings darauf hin, dass das Ignorieren fehlender Antworten zu einer genaueren Parameterschätzung führt (Custer, Sharairi & Swift, 2012; Pohl, Gräfe & Rose, 2014; Shin, 2009), auch wenn manche Autoren die Aussagekraft von Simulationsstudien in dem Kontext anzweifeln (Robitzsch, 2014a).

Um Hinweise auf den angemessenen Umgang mit fehlenden Werten in der B90 zu gewinnen, wurden sowohl die fehlenden Werte selbst (im Sinne einer Trennschärfenanalyse) als auch die Anzahl der fehlenden Werte mit dem Anteil richtiger Antworten unter den bearbeiteten Aufgaben korreliert. Die fehlenden Werte selbst gingen dabei also zur Vermeidung von Konfundierung nicht in die Bildung des Gesamttestwerts ein, sondern wurden ignoriert.

Sowohl ausgelassene Aufgaben als auch nicht erreichte Aufgaben korrelieren nicht oder nur geringfügig mit dem Anteil richtig gelöster Aufgaben (ausgelassene Aufgaben -.01, nicht erreichte Aufgaben -.07). Die Anzahl ausgelassener Aufgaben sowie die Anzahl nicht erreichter Aufgaben korreliert ebenfalls nicht oder nur geringfügig mit dem Anteil richtig gelöster Aufgaben (ausgelassene Aufgaben -.02, nicht erreichte Aufgaben -.07). Mit anderen Worten: Sowohl bezogen auf einzelne Aufgaben als auch auf die Anzahl von fehlenden Werten pro Person zeigt sich für das Auslassen von Aufgaben kein Zusammenhang mit der Personenfähigkeit. Für das Nicht-Erreichen von Aufgaben zeigt sich ein geringfügig negativer Zusammenhang mit der Personenfähigkeit.

Für die B90 ergibt sich die Schlussfolgerung, dass ausgelassene Aufgaben bei der Ermittlung der Personenfähigkeit ignoriert werden können, da ihr Vorhandensein nicht systematisch mit der Lösungshäufigkeit der bearbeiteten Aufgaben zusammenhängt. Der nur geringe Zusammenhang des Nicht-Erreichens von Aufgaben mit der Lösungshäufigkeit der bearbeiteten Aufgaben scheint ebenfalls zu erlauben, diese bei der Ermittlung der Personenfähigkeit unberücksichtigt zu lassen. Es werden daher sowohl fehlende Werte durch „omitted“ als auch fehlende Werte durch „not reached“ bei der Bildung des Summenwerts und der WLE-Schätzung ignoriert.

*Tabelle 2: Gültige Antworten und Itemkennwerte (Anmerkung: * Aufgabe bei der Bildung des Gesamttestwerts/WLE unberücksichtigt.)*

		gültige Werte											DIF			
Aufgabe	Position	gültig	% gültig	% aus-gelassen	% nicht er-reicht	% ungültig	Schwie-rigkeit b	SE	WMNSQ	T	r_{it}	r_{itc}	Geschlecht	Bildung Eltern	MH	Hoch-schultyp
bas7mar1	1	334	98.82	1.18	.00	.00	-1.74	.16	.92	-.90	.41	.35	-.61	.10	-.06	.40
bas7mar2	2	334	98.82	1.18	.00	.00	-.21	.12	1.01	.40	.36	.28	.49	.39	.30	-.46
bas7mar3	3	336	99.41	.59	.00	.00	-1.10	.14	1.12	2.00	.17	.09	.25	-.42	.32	.64
bas7mar4	4	338	10.00	.00	.00	.00	-.87	.13	1.02	.50	.31	.23	.16	-.05	.03	.34
bas7mar5	5	337	99.70	.00	.00	.30	-.09	.12	1.07	1.90	.28	.20	.27	.22	.35	-.21
bas7mar6	6	329	97.34	2.66	.00	.00	-.79	.13	1.04	.70	.33	.25	-.09	-.06	-.46	.01
bas7org1	7	329	97.34	2.37	.00	.30	-1.33	.14	1.05	.80	.26	.18	-.45	-.41	.39	.24
bas7org2	8	334	98.82	1.18	.00	.00	-.32	.13	.98	-.40	.40	.32	-.19	-.26	-.03	.32
bas7org3	9	333	98.52	1.48	.00	.00	-.74	.13	.91	-1.80	.47	.40	-.16	-.07	-.23	.07
bas7org4	10	332	98.22	1.78	.00	.00	-.95	.13	.97	-.60	.41	.34	.57	.20	-.42	.25
bas7org5	11	334	98.82	1.18	.00	.00	-1.36	.14	.97	-.40	.38	.31	.21	.03	-.35	.27
bas7org6	12	333	98.52	1.48	.00	.00	-.23	.12	1.04	1.10	.32	.24	.21	.02	.02	.30
bas7fin1	13	324	95.86	4.14	.00	.00	-1.45	.15	1.00	.10	.32	.25	-.27	.32	.69	-.01
bas7fin2	14	317	93.79	6.21	.00	.00	.11	.13	.97	-.70	.44	.36	-.21	-.24	.18	-.56
bas7fin3	15	304	89.94	1.06	.00	.00	-.14	.13	1.08	1.90	.30	.22	.55	.32	.12	-.04
bas7fin4	16	334	98.82	1.18	.00	.00	-1.84	.16	.94	-.60	.40	.34	-.47	.05	-.23	.24
bas7fin5	17	264	78.11	21.89	.00	.00	1.07	.15	.99	-.10	.33	.26	.06	-.49	-.27	.33
bas7fin6	18	311	92.01	7.99	.00	.00	-.36	.13	.99	-.30	.42	.34	-.45	-.11	.40	-.13
bas7acc1	19	322	95.27	4.73	.00	.00	-1.23	.14	.91	-1.40	.47	.40	-.26	.08	.11	.21
bas7acc2	20	304	89.94	9.76	.30	.00	.08	.13	1.03	.70	.34	.25	.29	.19	.19	-.21
bas7acc3	21	296	87.57	11.24	1.18	.00	-1.72	.17	.94	-.60	.41	.35	-.33	.05	-.06	.36
bas7acc4	22	317	93.79	4.73	1.48	.00	-.88	.14	.93	-1.20	.45	.38	-.25	-.09	-.35	-.11
bas7acc5	23	297	87.87	9.47	2.66	.00	.30	.13	.97	-.80	.42	.34	.28	-.07	.31	-.21
bas7acc6	24	288	85.21	1.65	4.14	.00	-.52	.14	.95	-1.10	.43	.36	.06	.02	-.08	.45
bas7mic1	25	316	93.49	1.18	5.33	.00	-.82	.13	.98	-.40	.39	.32	-.11	.15	.00	-.24
bas7mic2	26	309	91.42	1.48	7.10	.00	-.04	.13	1.10	2.50	.23	.14	-.20	-.23	.48	-.43
bas7mic3	27	295	87.28	3.25	9.47	.00	.38	.13	1.08	1.80	.26	.17	.01	.12	.41	.43
bas7mic4	28	294	86.98	1.18	11.83	.00	.56	.13	1.10	2.20	.22	.13	-.23	-.03	.17	-.07
bas7mic5	29	286	84.62	.89	14.50	.00	.45	.14	.97	-.70	.42	.34	.31	-.44	.10	-.23
bas7mic6*	30	280	82.84	.89	16.27	.00	.69	.14	.95	-1.10	.43	.36	.05	-.13	-.16	-1.29
bas7mac1*	31	269	79.59	2.96	17.46	.00	.62	.14	.99	-.20	.38	.30	-.25	.42	-1.13	-.25
bas7mac2	32	272	8.47	1.48	18.05	.00	.63	.14	.95	-1.00	.44	.36	-.19	.17	.37	.43
bas7mac3	33	269	79.59	.59	19.82	.00	-1.05	.15	1.01	.20	.32	.25	.14	.11	-.44	.06
bas7mac4	34	251	74.26	.30	25.44	.00	-.17	.14	1.04	.90	.31	.23	.28	-.07	-.30	.40
bas7mac5	35	230	68.05	2.66	29.29	.00	.52	.15	1.10	1.80	.23	.15	.11	.29	-.56	-.43
bas7mac6	36	234	69.23	.00	30.77	.00	-.27	.15	.96	-1.00	.44	.36	-.28	-.04	-.61	-.42

3.2 Skalierung

Es wurden IRT-Modelle mit dem Programm ConQuest (Adams, Wu & Wilson, 2012) berechnet. Für die Ermittlung der Item- und Personenparameter und die Untersuchungen zum Differential Item Functioning (DIF) wurde die Standard-Schätzmethode (Gauss-Hermite) angewendet. Für die Schätzung der mehrdimensionalen Modelle wurde die Monte-Carlo-Methode angewendet. Ergänzend wurden auch Kennwerte der klassischen Testtheorie wie beispielsweise Trennschärfen berechnet.

3.2.1 Itemparameter und Item-Fit

Tabelle 2 zeigt die Itemparameter der 36 Aufgaben. Die Item-Fits (WMNSQ) reichen von .91 bis 1.12. Die Item-Schwierigkeitsparameter reichen von -1.84 bis 1.07. Die durchschnittliche Trennschärfe (r_{it}) beträgt .36, die Trennschärfen der einzelnen Items reichen von .17 bis .47. Die durchschnittliche korrigierte Trennschärfe (r_{itc}) beträgt .28, die korrigierten Trennschärfen der einzelnen Items reichen von .09 bis .40. Drei Items (bas7mar3, bas7mic2, bas7mic4) weisen einen signifikanten T-Wert auf, der zugehörige WMNSQ liegt jedoch für alle Items unter der kritischen Grenze von 1.20.

3.2.2 Differential Item Functioning

Wie bei Testentwicklungen im NEPS üblich (vgl. Pohl & Carstensen, 2012, 2013) wurden DIF-Analysen für das Geschlecht, den Migrationshintergrund und die Bildungsherkunft durchgeführt. Da die in anderen Studien für den Bildungsherkunft verwendete Bücherfrage aus einer anderen Befragungswelle zugespielt werden müsste und dann ca. 30 % fehlende Werte aufweist, wurde stattdessen der Bildungsabschluss der Eltern herangezogen (0: kein akademischer Abschluss der Eltern, 1: mindestens ein Elternteil mit akademischem Abschluss). Aufgrund der in den Entwicklungsstudien beobachteten unterschiedlichen inhaltlichen Schwerpunktsetzung nach Hochschultyp wurden zusätzlich hierfür DIF-Analysen (Universität vs. Fachhochschule) berechnet. Ermittelt wurde DIF durch Schätzung getrennter Modelle für beide Gruppen und Ermittlung der absoluten Differenz der Logits bei separater Schätzung des Haupteffekts (Syntax: „model = item-group+item*group“). Absolute Differenzen größer .40 werden als schwacher, größer .60 als mittlerer und größer 1.00 als starker DIF betrachtet. Weiterhin wurde der Einfluss von DIF anhand von Modellvergleichen untersucht.

Schwachen DIF weisen 6 Items bezogen auf das Geschlecht, 8 Items bezogen auf den Hochschultyp, 5 Items auf die Bildungsherkunft und 6 Items bezogen auf den Migrationshintergrund auf (vgl. Tabelle 2). Mittleren DIF weisen 1 Item bezogen auf das Geschlecht, 1 Item bezogen auf den Hochschultyp und 2 Items bezogen auf den Migrationshintergrund auf. Starken DIF weisen das Item bas7mic6 bezogen auf den Hochschultyp und das Item bas7mac1 bezogen auf den Migrationshintergrund auf. Diese beiden Items werden deshalb bei der Ermittlung der Summen- und WLE-Scores für den Scientific-Use-File nicht berücksichtigt.

Das Vorhandensein von DIF wurde außerdem geprüft anhand von Modellvergleichen jeweils eines Modells nur mit Haupteffekt des Gruppierungsmerkmals und eines Modells mit zusätzlicher Schätzung eines Interaktionsterms (siehe Tabelle 3). Eine signifikant bessere Passung des Modells einschließlich Interaktionsterm zeigt sich allein für den Hochschultyp (ΔDeviance = 59.95, $p < .01$). AIC und BIC sprechen allerdings – wie auch bei allen anderen untersuchten Merkmalen – für das Modell ohne Interaktionsterm. Diese Ergebnisse deuten

darauf hin, dass das Ausmaß von DIF insgesamt keine ernsthafte Bedrohung der Testgüte darstellt.

Tabelle 3: Modellvergleiche DIF

	Deviance	Parameter	n	AIC	BIC
nur HE Geschlecht	13103.62	38	338	13179.62	13324.89
DIF Geschlecht	13058.71	74	338	13206.71	13489.62
nur HE Bildungsherkunft	13111.67	38	338	8166.08	8292.90
DIF Bildungsherkunft	13084.44	74	338	8206.19	8453.17
nur HE Migration	13108.51	38	338	13184.51	13329.79
DIF Migration	13074.68	74	338	13222.68	13505.58
nur HE Hochschultyp	13113.10	38	335	13089.93	13234.87
DIF Hochschultyp	13053.15	74	335	13104.61	13386.86

3.2.3 Reliabilität und Verteilung der Personenparameter

Der Mittelwert der Personenparameter wurde auf 0 fixiert (constraints = cases). Die Varianz der Personenparameter beträgt .59. Die WLE-Schätzer reichen von -2.36 bis 3.34, die WLE-Reliabilität beträgt 0.77. Der Mittelwert der Itemschwierigkeiten liegt bei -.47. Der grafische Vergleich der Personen- und Itemparameter zeigt, dass der Test insbesondere im mittleren und unteren Bereich der Personenfähigkeit gut differenziert. Im Bereich sehr guter Personenfähigkeit liegen nur wenige Aufgaben entsprechender Schwierigkeit vor.

3.2.4 Dimensionalität

Zur Überprüfung der Dimensionalität des Instruments wurde ein nach Inhaltsbereichen gegliedertes sechsdimensionales IRT-Modell berechnet. Das sechsdimensionale Modell bildet die Daten signifikant besser ab als das eindimensionale Modell (ΔDeviance = 60.60; $p < .001$), und der AIC fällt besser aus (siehe Tabelle 4). Der die Modellkomplexität berücksichtigende BIC fällt jedoch für das eindimensionale Modell günstiger aus.

Tabelle 4: Modellvergleiche Dimensionalität

	Deviance	Parameter	n	AIC	BIC
eindimensional	14500.14	37	338	13192.09	13333.55
sechsdimensional	14340.54	57	338	13171.50	13389.41

Die Korrelationen zwischen den Dimensionen sind durchgängig positiv und betragen zwischen .63 und .91 (siehe Tabelle 5). Während die meisten Inhaltsbereiche zwischen .80 und .90 korreliert sind, weisen die Bereiche Marketing und Organisation zwar einen hohen Zusammenhang miteinander, aber einen geringeren zu anderen Inhaltsbereichen (insbesondere zu Finanzierung und Makroökonomie) auf. Den geringsten Beitrag zur Differenzierung der Personenfähigkeiten liefern aufgrund ihrer vergleichsweise geringeren Varianz die Inhaltsbereiche Marketing und Mikroökonomie.

Im Vergleich zu anderen Kompetenzdomänen wie etwa der Lesefähigkeit (Pohl, Haberkorn & Hardt, 2014) fallen die Zusammenhänge zwischen den Subdimensionen im wirtschaftswissenschaftlichen Test teilweise geringer aus. Dies muss jedoch vor dem Hintergrund betrachtet werden, dass mit der B90 nicht eine allgemeine oder grundlegende Kompetenz von Studienanfängern, sondern das Studienergebnis einer vergleichsweise heterogenen Gruppe von Studierenden verschiedener wirtschaftswissenschaftlicher Studiengänge und Studienfortschritte erfasst wird. Dies könnte die geringer ausfallende interne Konsistenz des Instruments zumindest zum Teil erklären.

Tabelle 5: Korrelationen zwischen den Inhaltsbereichen

	Marketing	Organisation	Finanzierung	Rechnungsw.	Mikroökonomie	Makroökonomie
Marketing	*.52*					
Organisation	.91	*1.00*				
Finanzierung	.63	.63	*.85*			
Rechnungswesen	.79	.81	.89	*1.36*		
Mikroökonomie	.77	.78	.84	.88	*.43*	
Makroökonomie	.67	.67	.90	.86	.86	*.63*

Anmerkung: Varianzen sind in der Diagonale, Korrelationen unterhalb der Diagonale dargestellt.

3.2.5 Lokale stochastische Unabhängigkeit

Eindimensionale IRT-Modelle setzen voraus, dass die verwendeten Aufgaben abgesehen vom zu erfassenden Personenmerkmal voneinander unabhängig sind, also nach Herauspartialisierung des zu erfassenden Merkmals nicht miteinander oder mit anderen äußeren Merkmalen korrelieren (Embretson & Reise, 2000). Ein verbreitetes Verfahren zur Überprüfung lokaler stochastischer Unabhängigkeit ist die Q3-Statistik (Yen, 1984), die jeweils paarweise für die vorliegenden Aufgaben berechnet wird.

Zur Überprüfung der Annahme der lokalen stochastischen Unabhängigkeit wurden Q3-Statistiken mithilfe des R-Pakets „sirt“ (Robitzsch, 2014b) berechnet. Die Korrelationen der Residuen reichen von -.25 bis .17, die mittlere Korrelation der Residuen beträgt -.03. Während einige Autoren einen Grenzwert von $|Q3| \leq .20$ empfehlen (Chen & Thissen, 1997; Yen, 1994), merken Chen und Wang (2007) an, dass ein solcher Wert mit Vorsicht zu verwenden ist. Da die Q3-Statistik die Korrelation der Residuen darstellt, beschreibt ihr Quadrat den gemeinsamen Varianzanteil zweier Items (de Ayala, 2009). Mit einer gemeinsamen paarweisen Varianz von maximal etwa 6 % erscheint das Ausmaß lokaler Abhängigkeiten zwischen den Aufgaben vertretbar.

3.3 Im Scientific-Use-File berichtete Daten

Im Scientific-Use-File werden alle 36 im Testheft enthaltenen Aufgaben berichtet. Die im Multiple-Choice-Format erhobenen Daten wurden auf 0 = falsch und 1 = richtig umkodiert. Weiterhin werden die Summe richtiger Antworten und die Summe gültiger Antworten sowie Personenparameter als WLE und deren Standardfehler berichtet. Diese Kennwerte beziehen sich jeweils auf 34 Aufgaben, da 2 Aufgaben ausgelassen wurden (vgl. DIF-Analysen in Abschnitt 3.2.2). Die ConQuest-Syntax zur Berechnung der Personenparameter ist in Anhang I dargestellt, eine Übersicht über die veröffentlichten Variablen im Scientific-Use-File in Anhang II.

4. Zusammenfassung

In Etappe 7 des Nationalen Bildungspanels wurde in Zusammenarbeit mit dem WiwiKom-Projekt ein Test zur Erfassung wirtschaftswissenschaftlicher Fachkompetenz entwickelt und bei Teilnehmern der Startkohorte 5 eingesetzt. Das Instrument besteht aus 36 Einzelaufgaben, die im Multiple-Choice-Format zu beantworten sind. Vorarbeiten in Form von curricularen Analysen und Expertenratings stellen sicher, dass zentrale Inhaltsbereiche des Fachgebiets abgedeckt werden, die für alle wirtschaftswissenschaftlichen Studienfächer von Bedeutung sind. Die Zielgruppe sind Bachelor-Studierende, die sich nah am Ende ihres Studiums befinden oder dieses bereits abgeschlossen haben.

Selbstverständlich sind mit der Erfassung fachspezifischer Kompetenzen durch ein Kurzinstrument Einschränkungen verbunden. Durch die Fokussierung auf zentrale Inhalte wurde versucht, dem Dilemma der häufig sehr spezifischen Ausrichtung anwendungsorientierter Studiengänge (wie beispielsweise Logistik- oder Tourismuswirtschaft) einerseits und der eingeschränkten Vergleichbarkeit von Studierenden aller wirtschaftswissenschaftlichen Fächer andererseits zu begegnen. Vor diesem Hintergrund betrachtet ist auch die im Vergleich zu anderen Kompetenzdimensionen – wie etwa Lesen oder Mathematik – geringere interne Konsistenz des Instruments zu betrachten. Die Auswahl repräsentativer Aufgaben verschiedener wirtschaftswissenschaftlicher Inhaltsbereiche führt zwangsläufig zu einer gewissen Heterogenität des Tests. Aufgrund der zeitlichen Begrenzung der Interviewdauer und der daraus resultierenden Kürze des Instruments ist eine Auswertung von Subskalen bzw. Kompetenzen im Bereich einzelner wirtschaftswissenschaftlicher Fachgebiete allerdings weder möglich noch Ziel der Teilstudie. Solche mit der Struktur wirtschaftswissenschaftlicher Curricula verbundene Fragestellungen können nur in tiefergehenden Untersuchungen – wie beispielsweise die des Mainzer WiwiKom-Projekts – auf Basis eines größeren Itempools bearbeitet werden. Die interne Konsistenz bzw. die Korrelationen zwischen den Inhaltsbereichen scheinen jedoch trotzdem zu erlauben, einen Gesamttestwert wirtschaftswissenschaftlicher Fachkompetenz zu berichten.

Die Vorteile der Teilstudie liegen vor allem in der Möglichkeit, im Rahmen des NEPS nahezu alle relevanten Merkmale von Sozialstruktur, Bildungs- und späterer Erwerbsbiografie mit den erfassten fachspezifischen Kompetenzen verknüpfen zu können. Auch wenn sich der Test aus Aufgaben verschiedener wirtschaftswissenschaftlicher Fachgebiete zusammensetzt, erscheint die Bildung eines Gesamttestwerts wirtschaftswissenschaftlicher Kompetenz gerechtfertigt, da das eindimensionale Modell die beste Modellanpassung aufweist. Bei der Bildung des Gesamttestwerts wurde auf zwei Aufgaben verzichtet, die im Gruppenvergleich von Hochschultyp und Migrationshintergrund relativ große Schwierigkeitsunterschiede (DIF) aufwiesen. Die verwendeten 34 Items sind bezogen auf die Merkmale Hochschultyp, Migrationshintergrund und Geschlecht dagegen unauffällig, so dass der Test in dieser Hinsicht als fair einzuschätzen ist.

Literatur

Adams, R. J. & Wu, M. L. (2002). *PISA 2000 technical report*. Paris: OECD.

Adams, R. J., Wu, M. L. & Wilson, M. R. (2012). *ACER ConQuest 3.0 (computer software)*. Melbourne: ACER.

Aschinger, F., Epstein, H., Müller, S., Schaeper, H., Vöttiner, A. & Weiß, T. (2011). Higher education and the transition to work. In H.-P. Blossfeld, H.-G. Roßbach & J. von Maurice (Eds.), *Education as a lifelong process. The German National Educational Panel Study* (pp. 267-282). Wiesbaden: VS Verl. für Sozialwissenschaften.

Beck, K. & Krumm, V. (1998). *Wirtschaftskundlicher Bildungs-Test (WBT)*. Göttingen, Hogrefe.

Centro Nacional de Evaluación para la Educación Superior (2011). *Exámenes Generales para el Egreso de la Licenciatura*. Mexico, DF.

Chen, C.-T.& Wang, W.-C. (2007). Effects of ignoring item interaction on item parameter estimation and detection of interacting items. *Applied Psychological Measurement*, 31, 388-410.

Chen, W.-H. & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22, 265-289.

Custer, M., Sharairi, S. & Swift, D. (2012). *A comparison of scoring options for omitted and not-reached items through the recovery of IRT parameters when utilizing the Rasch model and joint maximum likelihood estimation*. Paper presented at the annual meeting of the National Council on Measurement in Education. Vancouver, British Columbia, April 12-16, 2012.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford.

Edele, A., Schotte, K., Hecht, M. & Stanat, P. (2012). *Listening comprehension tests of immigrant students' first language (L1) Russian and Turkish in grade 9: Scaling procedure and results*. (NEPS Working Paper No. 13). University of Bamberg, National Educational Panel Study.

Embretson, S. E. & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Frey, A., Hartig, J. & Rupp, A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice*, 28, 39-53.

Pohl, S. & Carstensen, C. H. (2012). *NEPS technical report. Scaling the data of the competence tests*. (NEPS Working Paper No. 14). University of Bamberg, National Educational Panel Study.

Pohl, S. & Carstensen, C. H. (2013). Scaling of competence tests in the National Educational Panel Study – Many questions, some answers, and further challenges. *Journal for Educational Research Online*, 5, 189–216.

Pohl, S., Gräfe, L. & Rose, N. (2014). Dealing with omitted and not reached items in competence tests – Evaluating approaches accounting for missing responses in IRT models. *Educational and Psychological Measurement*, 74, 423-452.

Pohl, S., Haberkorn, K. & Hardt, K. (2014). *NEPS Technical Report for Reading – Scaling results of Starting Cohort 5 for first-year students* (NEPS Working Paper No. 34). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Prussog-Wagner, A. & Aust, F. (2014). *Methodenbericht NEPS Startkohorte 5 - Dritte Kompetenztestung Haupterhebung Frühjahr 2014 B90*. Bonn: infas.

Robitzsch, A. (2014a). *Zu nichtignorierbaren Konsequenzen des (partiellen) Ignorierens fehlender Item Responses im Large-Scale Assessment*. Preprint.

Robitzsch, A. (2014b). *Supplementary item response theory models*. R-Paket, Version 0.45-23.

Rohwer, G. (2013). *Making sense of missing answers in competence tests* (NEPS Working Paper No. 30). University of Bamberg, National Educational Panel Study.

Shin, S.-H. (2009) How to treat omitted responses in Rasch model-based equating. *Practical Assessment Research & Evaluation*, 14. Verfügbar unter: http://pareonline.net/getvn.asp?v=14&n=1

Walstad, W. B. (2007). *Test of Understanding in College Economics. National Council on Economic Education*, New York.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8, 125-145.

Zlatkin-Troitschanskaia, O., Förster, M., Brückner, S., Hansen, M. & Happ, R. (2013). Modellierung und Erfassung der wirtschaftswissenschaftlichen Fachkompetenz bei Studierenden im deutschen Hochschulbereich. In O. Zlatkin-Troitschanskaia, R. Nickolaus & K. Beck (Hrsg.), *Kompetenz-modellierung und Kompetenzmessung bei Studierenden der Wirtschaftswissenschaften und der Ingenieurwissenschaften* (S. 108-133). Lehrerbildung auf dem Prüfstand (Sonderheft). Landau: Verlag Empirische Pädagogik.

Zlatkin-Troitschanskaia, O., Förster, M., Brückner, S. & Happ, R. (2014). Insights from a German assessment of business and economics competence. In H. Coates (Ed.), *Higher education learning outcomes assessment – International perspectives* (pp. 175-197). Frankfurt am Main: Peter Lang.

Anhang I: ConQuest-Syntax zur Berechnung der Personenparameter

```
------------------------------------------------
title B90 (ohne Items bas7mic6, bas7mac1);
let name = b90data;
datafile %name%.dat;
format id 1-4 responses 5-40;
labels << %name%.nam;
codes 1,2,3,4;
key 14322212324411211113442141421114413 3 !1;

delete !item (30, 31);

set constraints = cases;

model item;

estimate !method=gauss, iter=1000, nodes=15;

export parameters              >> %name%.par;
export reg_coefficients        >> %name%.regr;
export covariance              >> %name%.cov;
export logfile                 >> %name%.log;

show cases !estimates = wle    >> %name%.wle;
show   >> %name%.shw;
itanal >> %name%.itn;
quit;
------------------------------------------------
```

Anhang II: Übersicht über die im Scientific-Use-File enthaltenen Variablen (Codebook)

Variable	Obs	Unique	Mean	Min	Max	Label
idtarget	338	338	7.11e+07	7.11e+07	7.12e+07	Personennummer
bas7mar1 c	338	3	-.3313609	-97	1	Ökonomische Kompetenz: Marketing 1
bas7mar2_c	338	3	-.6094675	-97	1	Ökonomische Kompetenz: Marketing 2
bas7mar3_c	338	3	.147929	-97	1	Ökonomische Kompetenz: Marketing 3
bas7mar4_c	338	2	.683432	0	1	Ökonomische Kompetenz: Marketing 4
bas7mar5_c	338	3	.2366864	-95	1	Ökonomische Kompetenz: Marketing 5
bas7mar6_c	338	3	-1.931953	-97	1	Ökonomische Kompetenz: Marketing 6
bas7org1_c	338	4	-1.831361	-97	1	Ökonomische Kompetenz: Organisation-Unternehmensführung-Personal 1
bas7org2_c	338	3	-.5857988	-97	1	Ökonomische Kompetenz: Organisation-Unternehmensführung-Personal 2
bas7org3_c	338	3	-.7869822	-97	1	Ökonomische Kompetenz: Organisation-Unternehmensführung-Personal 3
bas7org4_c	338	3	-1.035503	-97	1	Ökonomische Kompetenz: Organisation-Unternehmensführung-Personal 4
bas7org5_c	338	3	-.3846154	-97	1	Ökonomische Kompetenz: Organisation-Unternehmensführung-Personal 5
bas7org6_c	338	3	-.8934911	-97	1	Ökonomische Kompetenz: Organisation-Unternehmensführung-Personal 6
bas7fin1_c	338	3	-3.263314	-97	1	Ökonomische Kompetenz: Finanzierung 1
bas7fin2_c	338	3	-5.579882	-97	1	Ökonomische Kompetenz: Finanzierung 2
bas7fin3_c	338	3	-9.284024	-97	1	Ökonomische Kompetenz: Finanzierung 3
bas7fin4_c	338	3	-.316568	-97	1	Ökonomische Kompetenz: Finanzierung 4
bas7fin5_c	338	3	-21.02071	-97	1	Ökonomische Kompetenz: Finanzierung 5
bas7fin6_c	338	3	-7.213018	-97	1	Ökonomische Kompetenz: Finanzierung 6
bas7acc1_c	338	3	-3.87574	-97	1	Ökonomische Kompetenz: Rechnungswesen 1
bas7acc2_c	338	4	-9.316568	-97	1	Ökonomische Kompetenz: Rechnungswesen 2
bas7acc3_c	338	4	-11.29586	-97	1	Ökonomische Kompetenz: Rechnungswesen 3
bas7acc4_c	338	4	-5.33432	-97	1	Ökonomische Kompetenz: Rechnungswesen 4
bas7acc5_c	338	4	-11.30473	-97	1	Ökonomische Kompetenz: Rechnungswesen 5
bas7acc6_c	338	4	-13.70118	-97	1	Ökonomische Kompetenz: Rechnungswesen 6
bas7mic1_c	338	4	-5.52071	-97	1	Ökonomische Kompetenz: Mikro 1
bas7mic2_c	338	4	-7.642012	-97	1	Ökonomische Kompetenz: Mikro 2
bas7mic3_c	338	4	-11.68935	-97	1	Ökonomische Kompetenz: Mikro 3
bas7mic4_c	338	4	-11.94083	-97	1	Ökonomische Kompetenz: Mikro 4
bas7mic5_c	338	4	-14.14497	-97	1	Ökonomische Kompetenz: Mikro 5
bas7mic6_c	338	4	-15.86391	-97	1	Ökonomische Kompetenz: Mikro 6
bas7mac1_c	338	4	-18.98817	-97	1	Ökonomische Kompetenz: Makro 1
bas7mac2_c	338	4	-18.10651	-97	1	Ökonomische Kompetenz: Makro 2
bas7mac3_c	338	4	-18.63314	-97	1	Ökonomische Kompetenz: Makro 3
bas7mac4_c	338	4	-23.80473	-97	1	Ökonomische Kompetenz: Makro 4
bas7mac5_c	338	4	-29.85503	-97	1	Ökonomische Kompetenz: Makro 5
bas7mac6_c	338	3	-28.5355	-94	1	Ökonomische Kompetenz: Makro 6
bas7_sc3	338	31	18.61243	1	33	Ökonomische Kompetenz: Summe korrekt
bas7_sc8	338	18	30.8787	7	34	Ökonomische Kompetenz: Anzahl gültig
bas7_sc1	338	197	.0000049	-2.35782	3.34492	Ökonomische Kompetenz: WLE
bas7 sc2	338	197	.4276782	.36473	1.47999	Ökonomische Kompetenz: SE(WLE)