Michelle Liedtke / Susan Seeber

Modellgeltungstests und Einflussfaktoren auf Differentielle Item Funktionen in einem computergestützten Assessment für kaufmännische Berufe

Kurzfassung: Es ist gegenwärtig nicht unumstritten, ob berufliche Handlungskompetenz mit computerbasierten Assessmentmethoden zuverlässig und valide gemessen werden kann. Im Rahmen des Verbundvorhabens CoBALIT (Competencies in the Field of Business and Administration, Learning, Instruction and Transition) wurden vor diesem Hintergrund berufsübergreifende und berufsspezifische Kompetenzen für den Ausbildungsberuf „Industriekaufmann/Industriekauffrau" rechnergestützt erfasst. Darüber hinaus wurde geprüft, inwiefern sich kaufmännische Kompetenzen, die in verschiedenen Berufen erforderlich sind, berufsübergreifend modellieren lassen. Exemplarisch wird dies bei Auszubildenden für die beiden Berufe „Industriekaufmann/Industriekauffrau" und „Kaufmann/Kauffrau für Spedition und Logistikdienstleistungen" geprüft. Im Vorfeld der Aufgabenentwicklung wurden curriculare Analysen und Tätigkeitsanalysen durchgeführt. Anschließend wurden auf der Grundlage der Daten der Haupterhebung verschiedene Modellgeltungstests und Analysen zum möglichen Vorliegen von DIF (Differential Item Functioning) durchgeführt, um zu entscheiden, ob vergleichende Aussagen über die erworbenen Kompetenzen zwischen den beiden Gruppen belastbar sind. Die Stichprobe umfasst 179 Industriekaufleute und 423 Speditionskaufleute, die im Frühjahr/Sommer 2014 im 3. Ausbildungsjahr getestet wurden. Die Ergebnisse zu den DIF-Analysen und Modellgeltungstests verweisen auf spezifische Unterschiede, die möglicherweise auf unterschiedliche Ausbildungskulturen und Lerngelegenheiten zurückzuführen sind.

Abstract: There is presently no general consensus as to the feasibility of adequately measuring practical vocational skills by way of computer-based assessment techniques. This controversy rendered the background for testing generic and vocation-specific competencies using computer-based assessment techniques, applied to for apprentices for the trades of an „industrial management assistant". The context was given by the joint CoBALIT (Competencies in the Field of Business and Administration, Learning, Instruction and Transition), embedded in the ASCOT (Technology-based Assessment of Skills and Competencies in VET) consortium. Starting out from an analysis of the pertinent curricula including lesson and training plans, course books and examination questions, it was examined to what extent overarching commercial competencies which are required for different specific commercial occupations can be projected onto a joint metric. Two groups in the area of logistics and transport were compared: trainees for the position of industrial management assistants and apprentices for the position of a „management assistant for freight forwarding and logistics". The relevant data were scaled on basis of the Item Response Theory, incorporating various goodness of model fit tests used in order to decide whether well-founded comparative statements between the two groups in terms of acquired competencies are justified. This was examined using curricular analyses, which give first indications of prioritisation within the apprenticeship, as well as so-called DIF-Analyses (Differential Item Functioning). The sample comprises 179 industrial management assistants and 423 assistants for freight forwarding and logistics. Both groups were tested in their third year of training (spring/summer of 2014). The differences encountered appear to be attributable to a fair degree different training cultures and learning opportunities.

1 Forschungsstand

1.1 Zur Kompetenzmessung in der beruflichen Bildung

Wird die internationale Kompetenzforschung betrachtet, so zeichnen sich je nach kultureller Spezifik und wissenschaftlicher Fachdisziplin unterschiedliche Ansätze ab, Kompetenzen zu definieren, zu operationalisieren und einer belastbaren Diagnostik zugänglich zu machen. An dieser Stelle sei verwiesen auf die Arbeiten von Norris (1991), Rychen & Salganik (2001) und im Besonderen auf Weinert (2001) zur Systematisierung von Kompetenzbegriffen und -konzepten. In der beruflichen Bildung erlangte der Kompetenzbegriff bereits vor den intensiven Debatten um die Kompetenzen deutscher Schülerinnen und Schüler im Zuge der ersten PISA-Studie (Deutsches PISA-Konsortium, 2001) mit dem Konzept der Handlungsorientierung (KMK, 1996) an Relevanz. Allerdings zeichnet sich bis heute eine schwer auflösbare Diskrepanz zwischen dem in der Berufs- und Wirtschaftspädagogik auf Roth (1971) und Reetz (1984) zurückgehenden Begriff der beruflichen Handlungskompetenz und dem in der empirischen Bildungsforschung weit verbreiteten, stärker kognitionspsychologisch orientierten Kompetenzansatz ab (vgl. Klieme & Hartig, 2008; Klieme & Leutner, 2006).

Der komplexe Ansatz einer beruflichen Handlungskompetenz, der auch Aspekte der beruflichen Mündigkeit und Selbstbestimmtheit, der gesellschaftlichen Teilhabe und der aktiven Gestaltung und Veränderung von Handlungsbedingungen in beruflichen und außerberuflichen Bereichen einbezieht, wird beim gegenwärtigen Stand der methodischen und psychometrischen Forschung einer Operationalisierung kaum zugänglich sein (vgl. Nickolaus & Seeber 2013, S. 169 ff.). Nicht zuletzt deshalb konzentriert sich auch die empirische Berufsbildungsforschung in der beruflichen Kompetenzdiagnostik zunächst auf die beruflichen Fachkompetenzen (vgl. dazu Baethge et al., 2006, S. 16 f.), bezieht jedoch zusehends auch andere Kompetenzdimensionen wie die der sozialen Kompetenzen ein (vgl. Monnier, Srbeny & Tschöpe 2014), und misst diese in hinreichend vielen variierenden Situationen und Aufgabenformaten. Denn auch nur unter dieser Bedingung kann von einer ‚Messung' mit einem hohen Anspruch an die Einhaltung diagnostischer Gütekriterien (vgl. dazu Messick, 1995) gesprochen werden.

Empirisch fundierte kognitionspsychologische Ansätze zur Kompetenzmodellierung gehen davon aus, dass Kompetenzen grundsätzliche Handlungsanforderungen innerhalb eines Fachs, eines Lernbereichs oder einer Domäne widerspiegeln, durch situative Anforderungen definiert und über verschiedene individuelle Ressourcen bewältigt werden (Klieme, Maag-Merki & Hartig, 2007, S. 6 ff.). Sie sind Ergebnisse von Lernprozessen und sind damit auch durch Lern- und Trainingsprozesse veränderbar (Seeber, 2011). Wissen stellt eine entscheidende Facette der Kompetenz dar und wird als zentrale Bedingung für einen Transfer auf variierende Kontexte gesehen (Minnameier, 2006; Baethge et al., 2006, S. 40). Jedoch geht das Kompetenzkonzept deutlich über kognitive Aspekte hinaus und schließt ebenso motivationale Orientierungen sowie soziale, volitionale und affektive sowie motorische Komponenten ein. Nickolaus & Seeber (2013, S. 167) sehen in der Einlösung von Validitätsansprüchen – nicht nur mit Blick auf die Berufsprofile und die in den beruflichen Curricula definierten Kompetenzanforderungen, sondern vor allem bezogen auf die prognostische Validität – nach wie vor eine ungelöste und zentrale Herausforderung. Neben der Inhalts- und Kriteriumsvalidität betont Messick (1995)

die Konstruktvalidität und schließt darin insbesondere die möglichen Konsequenzen, die aus den Testergebnissen für die Probanden resultieren können, ein. Dies ist ein Aspekt, der in der bisherigen Kompetenzdebatte deutlich vernachlässigt und nur unzureichend eingelöst wurde und wird.

Im Rahmen des BMBF-geförderten Projekts CoBALIT (www.ascot-vet.net) wird auf der Grundlage des computerbasierten Assessments ALUSIM (Achtenhagen & Winther, 2009) spezifischen Validitätsfragen intensiver nachgegangen. So ist ein wesentliches Ziel des Teilprojekts CoBALIT-Transfer, zu prüfen, inwiefern sich kaufmännische Kernkompetenzen, die in verschiedenen Berufen erforderlich sind, berufsübergreifend modellieren lassen. Exemplarisch wird dies bei Auszubildenden für die beiden Berufe „Industriekaufmann/Industriekauffrau" und „Kaufmann/Kauffrau für Spedition und Logistikdienstleistungen" geprüft. Im Vorfeld der Aufgabenentwicklung wurden curriculare Analysen und Tätigkeitsanalysen durchgeführt, in denen vor allem Aspekte der Domänenmodellierung und der kaufmännischen Kernkompetenzen, die in verschiedenen kaufmännischen Berufen essentiell für die Bewältigung der Arbeitsanforderungen sind, herausgearbeitet wurden. In diesem Beitrag werden nunmehr die Ergebnisse der Domänenmodellierung kurz vorgestellt und anschließend wird den Fragen der Konstruktvalidität im Sinne des Ansatzes von Messick (1995) nachgegangen.

1.2 Domänenmodellierung und kaufmännische Kernkompetenzen

Unter den nach BBiG und HwO anerkannten Ausbildungsberufen sind über 50 kaufmännische Berufe vertreten, die einerseits einen Überschneidungsbereich in den kaufmännischen Anforderungen, andererseits aber auch tätigkeitsspezifische Profile aufweisen (vgl. auch Reinisch & Götzl, 2013). Diese Vielzahl an spezialisiert gestalteten kaufmännischen Berufen hat in den letzten Jahren vermehrt zu Kritik geführt, insbesondere wurden in diesem Zusammenhang Fragen der beruflichen Mobilität kritisch diskutiert (BMBF, 2008, S. 23 f.; Brötz, Schapfel-Kaiser & Schwarz, 2008)[1]. Vor dem Hintergrund der demografischen Entwicklung und den Flexibilitätserfordernissen am Arbeitsmarkt wurden in den letzten Jahren die Bemühungen verstärkt, sogenannte kaufmännische Kernkompetenzen herauszuarbeiten, „die zur Berufsausübung in allen kaufmännischen Berufsfeldern erforderlich sind und damit ein Fundament kaufmännischer Aus- und Fortbildungsstandards bilden können, wenngleich je nach Beruf in unterschiedlicher Intensität" (Brötz et al., 2009, S. 19). Nach Brötz et al. (2008, S. 24 f.) bilden derartige Kernkompetenzen die gemeinsame Basis der Einzelprofile unterschiedlicher Berufe und konstituieren sie als sogenannte Berufsfamilie. Neben gemeinsamen Kompetenzbündeln und Kernaufgaben kaufmännischer Ausbildungsberufe sind jedoch auch gezielt Unterschiede zu identifizieren, um kaufmännische „Berufsfamilien" bestimmen zu können. Bisher stehen jedoch nur wenige Ansätze zur Verfügung, die einzelne kaufmännische Berufsbilder in ihrer Gesamtheit analysieren (Brötz et al., 2009, S. 33). Hervorzuheben ist in diesem Zusammenhang, dass die verlässliche Identifikation, Abbildung und Messung von Gemeinsamkeiten und Unterschieden von Berufen entsprechend gestaltete Kompetenzmodelle und Messinstrumente erfordern, die

1 Vgl. hierzu auch die Überspezialisierungsthese der KMK (2007).

berufsübergreifend eingesetzt werden können. Genau hier setzt das im Folgenden beschriebene Forschungsprojekt an.

Mit dem Beruf des Industriekaufmanns wird für die Konzeption eines Assessments ein kaufmännischer Kernberuf zugrunde gelegt, der curriculare Überschneidungen zu einer Reihe anderer kaufmännischer Berufe aufweist, aber auch spezifische Besonderheiten besitzt. Zu diesen Besonderheiten gehören u. a. die Rekrutierung von Jugendlichen mit höheren Schulabschlüssen und die Konzentration auf mittlere und größere Produktionsunternehmen. Aufgrund der Ausbildung in mittleren und großen Industrieunternehmen sind angehende Industriekaufleute kaum mit den spezifischen Arbeitsanforderungen und -bedingungen in Unternehmen des Dienstleistungssektors vertraut. Insofern stellt sich die Frage, ob beispielsweise Industriekaufleute ihre unter den spezifischen Bedingungen eines produzierenden, eher größeren Unternehmens erworbenen kaufmännischen Kompetenzen auch in einem anderen Unternehmenskontext anwenden können. Umgekehrt lässt sich aber ebenso hinterfragen, ob kaufmännische Auszubildende, die vorrangig in kleinen und mittleren Dienstleistungsunternehmen ausgebildet werden, auch die Anforderungen an kaufmännische Tätigkeiten in größeren Industriebetrieben erfüllen können.

In den bisherigen Studien zur Kompetenzmessung im kaufmännischen Sektor wurde allenfalls ansatzweise empirisch geprüft, ob sich Ziele, Inhalte und Arbeitshandlungen vom beruflichen Handlungsfeld eines bestimmten kaufmännischen Berufs auf andere Berufe übertragen lassen. Vor dem Hintergrund dieser Befunde erscheint es unabdingbar, die Übertragbarkeit von Testarrangements, die für einen spezifischen kaufmännischen Beruf entwickelt worden sind, auch auf andere kaufmännische Berufe empirisch abzusichern.

Zu einem der quantitativ bedeutsamsten kaufmännischen Berufe zählt u. a. der Beruf des Speditionskaufmanns. Dieser ist insofern interessant, da er in seinen Kompetenzanforderungen einerseits eine deutliche Schnittmenge mit jenen von Industriekaufleuten aufweist, andererseits aber im Vergleich zum Industriekaufmann auch deutliche Unterschiede in den Eingangsvoraussetzungen zeigt, und zwar sowohl in den formalen Eingangsvoraussetzungen/Schulabschlüssen als auch in den tatsächlichen Kompetenzen (vgl. Seeber, 2007). Darüber hinaus wird der Kaufmann für Spedition und Logistikdienstleistungen in Unternehmen der Logistikbranche ausgebildet, die sich deutlich in ihrer Struktur, im Organisationsaufbau und in der Gestaltung der Geschäftsprozesse von produzierenden Unternehmen unterscheiden. Insoweit sind in mehrfacher Hinsicht wichtige Erkenntnisse zu Fragen der Generalisierbarkeit und des Transfers aus dieser Forschung zu erwarten. Dies gilt auch für den jeweiligen relativen Einfluss individueller Ausgangsvoraussetzungen und betrieblicher Ausbildungsbedingungen auf die Kompetenzentwicklung.

Zu den gemeinsamen Schnittmengen in den beruflichen Anforderungen von Industrie- und Speditionskaufleuten gehören beispielsweise die Planung, Steuerung und Kontrolle von Beschaffungs- und Absatzprozessen einschließlich Marketing, die Beurteilung und Bewertung von Lagerleistungen aus unternehmerischer Perspektive sowie die Analyse und Beurteilung von Wertschöpfungsprozessen und die Erfassung und Dokumentation von Wertströmen. Es ist jedoch deutlich, dass bei prinzipiell ähnlichen Geschäftsvorgängen die Perspektiven jeweils hochgradig berufsspezifisch sind, wenn auch die kognitiven Bearbeitungsprozesse deutliche Gemeinsamkeiten aufweisen. Während beispielsweise die Beschaffungsplanung beim Industriekaufmann für das eigene Unternehmen vorgenommen wird, betrifft die

Planung beim Speditionskaufmann hauptsächlich eine angebotene Dienstleistung für ein anderes Unternehmen und erlangt über die Akquise und die erfolgreiche Abwicklung des Kundenauftrags erst mittelbar wirtschaftliche Bedeutung für das eigene Unternehmen.

Im Rahmen des BMBF-geförderten Projektes CoBALIT-Transfer wird geprüft, ob und inwiefern sich Ziele, Inhalte und Arbeitshandlungen von dem beruflichen Handlungsfeld des Industriekaufmanns[2] in einen anderen kaufmännischen Handlungsbereich, und zwar den des Kaufmanns für Spedition und Logistikdienstleistungen, übertragen lassen. Ergänzend zu den Testaufgaben für Industriekaufleute in den Bereichen Beschaffung, Absatz und Produktionsvorbereitung sind daher kaufmännische Aufgaben für den Bereich Logistik und Transport entwickelt worden, der in beiden Berufen ein wichtiges Handlungsfeld darstellt, aber für die Kaufleute für Spedition und Logistikdienstleistungen den beruflichen Kernbereich ausmacht. Die Testumgebung ALUSIM umfasst insgesamt die Bereiche Arbeitsvorbereitung, Einkauf, Vertrieb, Unternehmenskommunikation, Intrapreneurship und Logistik. Auf Basis dieser nach kaufmännischen Handlungsbereichen geclusterten Itempools kann nun analysiert werden, welche gemeinsamen Kompetenzen sich in den Berufen identifizieren lassen und inwiefern kaufmännische Anforderungen in verschiedenen wirtschaftlichen Kontexten berufsübergreifend relevant sind.

Im Rahmen dieses Artikels werden für die Prüfung der Validität des Testarrangements zwischen beiden Gruppen die kaufmännischen Aufgaben aus dem Bereich Logistik und Transport im Vordergrund stehen. Bei der Konzeption der Aufgaben wurde darauf geachtet, dass Aufgaben konstruiert wurden, die in den curricularen Vorgaben und Tätigkeitsanforderungen beider Berufe enthalten sind. Insgesamt wurden für das Assessment-Modul „Logistik“ 18 Aufgaben entwickelt, die einer handlungslogischen Struktur folgend in komplexe Arbeits- und Geschäftsprozesse eingebunden sind. Die Items wurden auf unterschiedlichem Schwierigkeitsniveau[3] entwickelt und sind durch die Auszubildenden handlungsleitend zu erschließen und zu bearbeiten. Alle Items werden über die webbasierte Unternehmenssimulation ALUSIM[4] zur Verfügung gestellt, wobei allen Probanden die gleichen Informationen und Bearbeitungstools zur Bewältigung der Anforderungssituationen zur Verfügung stehen (ERP-Software, Ablagen, E-Mail etc.).

Als Datenbasis wird der erste Haupterhebungszeitpunkt des Projektes (Februar 2014 bis Juni 2014) herangezogen. Die im Beitrag dargestellten Ergebnisse basieren auf einer Stichprobengröße von 179 Industriekaufleuten und 423 Kaufleuten für Spedition und Logistikdienstleistungen.

2 Im Interesse der besseren Lesbarkeit wird im Folgenden bei der Bezeichnung von Personen und Personengruppen ausschließlich das generische Maskulinum verwendet; es schließt beide Geschlechter mit ein.

3 Die Items wurden auf unterschiedlichen Schwierigkeitsstufen entwickelt, sodass differenzierte kognitive Prozesse und Ressourcen zur Bewältigung eines spezifischen Items notwendig werden (vgl. Winther, 2010, S. 99). Dazu werden drei verschiedene Kategorien von Itemeigenschaften herangezogen, die sich im Rahmen der Kompetenzmessung als signifikante Prädiktoren für die Itemschwierigkeit erwiesen haben. Hierzu zählen die funktionale Modellierung, die inhaltliche Komplexität und die kognitive Taxonomierung.

4 Vgl. zur Unternehmenssimulation ALUSIM z. B. Winther (2010).

2 Analysen und Ergebnisse zur Modellgeltung

Das theoriegeleitet entwickelte Messinstrument zur Erfassung des latenten Konstrukts der kaufmännischen Kompetenz im Bereich Spedition und Logistik wurde vor dem Hintergrund von Auswertungen mit probabilistischen Testmodellen entwickelt. Im Rahmen dieser Testtheorie entstand eine Vielzahl von Testmodellen, die sich zur Skalierung und Analyse von Daten eignen (vgl. z. B. Rost, 2004). Aufgrund der Struktur des entwickelten Tests, der sowohl dichotome als auch ordinale Items enthält, wurde zunächst das eindimensionale ordinale Rasch-Modell (Partial Credit Modell) nach Masters (1982) geprüft, bevor weitere Testmodelle zu Analysen herangezogen werden. Das eindimensionale ordinale Rasch-Modell berücksichtigt, dass nicht nur die Antwortformate „richtig" und „falsch" zugrunde gelegt werden können, sondern auch teilrichtige Lösungen innerhalb einer Aufgabe bewertbar sind. Von den 18 Aufgaben des Moduls weisen vier Aufgaben (Item 5, 7, 8, 15) eine solche gestufte Lösungsstruktur auf. Während sich bei dichotomen Items für jedes Item nur ein Schwierigkeitsparameter ergibt, zerlegt das Partial Credit Modell die ordinalen Items mit verschiedenen Antwortkategorien in unterschiedliche Schwellenschwierigkeiten (Wright & Masters, 1982).

Vor jeder Anwendung eines solchen Testmodells ist zu prüfen, ob die erhobenen Daten die Modellannahmen, in diesem Fall die zentralen Eigenschaften des Rasch-Modells, erfüllen (Rost, 2004, S. 345). Hierzu wurde eine Reihe von Modellgeltungstests entwickelt. Auch wenn es bis heute keinen festgelegten Kanon gibt, welche zentralen Annahmen zwingend zu prüfen sind, lassen sich jedoch aus dem aktuellen Forschungsstand (vgl. z. B. Bühner, 2011; Rost, 2004; Strobl, 2012) Empfehlungen ableiten, welche Tests durchgeführt werden sollten, bevor bei einer Schätzung das Rasch-Modell zur Anwendung kommt. Diese zentralen Modellannahmen und die dazugehörigen Modellgeltungstests werden im Folgenden näher erläutert.

2.1 Eigenschaften des Rasch-Modells und dessen Überprüfung

Als zentrale Eigenschaften des Rasch-Modells, die sich mit entsprechend etablierten Testverfahren überprüfen lassen, gelten u. a. die Eindimensionalität, die lokale stochastische Unabhängigkeit und die spezifische Objektivität (vgl. z. B. Rost, 2004; Koller, Alexandrowicz & Hatzinger, 2012). Vor dem Hintergrund heterogener kaufmännischer Anforderungen, selbst innerhalb eines Handlungsbereichs, stellen insbesondere die Kriterien der spezifischen Objektivität und der Eindimensionalität eine entscheidende Herausforderung dar. Nur wenn diese Eigenschaften erfüllt sind, handelt es sich um ein faires Testinstrument, das sich gleichermaßen für Industriekaufleute als auch für Kaufleute für Spedition und Logistikdienstleistungen zur Kompetenzmessung eignet. Verstöße gegen diese Kriterien zeigen dementsprechend Unterschiede zwischen den Berufsgruppen auf, die dann Überlegungen zu anderen angemessenen psychometrischen Modellen für die Interpretation der Daten erfordern. Entsprechend dieser Ausführungen ist in einem ersten Schritt die Passung des ordinalen eindimensionalen Rasch-Modells zu prüfen, um etwaige Unterschiede zwischen den Berufsgruppen festzustellen und darauf aufbauend ggf. ein Mischverteilungsmodell zu prüfen, das diesen potentiellen Unterschieden gerecht werden kann. Im Folgenden werden knapp die Eigenschaften des eindimensionalen Rasch-Modells erläutert und anschließend die Ergebnisse der empirischen Überprüfung diskutiert.

Eindimensionalität

Als eine wichtige Anforderung, die an ein Testinstrument zu stellen ist, gilt, dass die Beantwortung der entwickelten Items vorrangig auf die Fähigkeitsausprägung des zu untersuchenden latenten Konstrukts zurückzuführen ist (Koller et al., 2012, S. 4, 15). In unserem Beispiel wäre dies die kaufmännische Kompetenz im Bereich Spedition und Logistik. Entsprechend dieser Forderung müssen alle Items die gleiche Dimension erfassen und somit als homogen zu bezeichnen sein. Folglich ist bereits bei der Testkonstruktion darauf zu achten, dass Einflüsse, die eine entscheidende Rolle in Bezug auf die Eindimensionalitätsannahme einnehmen, möglichst gering zu halten sind. Zur Prüfung der Homogenität von Items eines Messinstrumentes können zum einen Abweichungsmaße für einzelne Items herangezogen werden (Rost, 2004, S. 351). Zum anderen kann auch die Bildung von potenziell heterogenen Itemgruppen als Modelltest dienen, indem die Items entsprechend gruppiert werden. Basis dieser Überlegungen ist eine Hypothese darüber, welche Itemgruppen möglicherweise unterschiedliche Persönlichkeitseigenschaften messen. Im einfachsten Fall stellt sich dies in Form von zwei Testhälften dar. Die Idee, die einem darauf beruhenden Modellgeltungstest zugrunde liegt, ist die, dass für beide Testhälften getrennt die Personenparameter ermittelt werden und geprüft wird, ob die Messwerte bis auf Zufallsschwankungen identisch sind.

Zur Überprüfung der Eindimensionalitätsannahme, auf Basis zuletzt beschriebener Vorgehensweise, wurde der sogenannte Martin-Löf-Test eingesetzt. Dieser hat jedoch nicht die geschätzten Personenparameter zum Gegenstand, sondern bezieht sich auf deren erschöpfende Statistiken, d. h. auf die Summenscores für beide Testhälften. Der Signifikanztest ist ein modifizierter Likelihoodquotiententest (Rost, 2004, S. 352). Er beruht auf den bedingten Likelihoods beider Testteile. Ein signifikantes Ergebnis weist darauf hin, dass die Items nicht homogen sind und somit eine grundlegende Annahme des Rasch-Modells verletzt ist.

In dem vorliegenden Beispiel wurde der Martin-Löf-Test mithilfe des Programms R durchgeführt. Als Teilungskriterium wurde der Median herangezogen, der als Standardteilungskriterium gilt (Koller et al., 2012, S. 95). Die Ergebnisse sind der folgenden Tabelle 1 zu entnehmen.

Tab. 1. Ergebnisse des Martin-Löf-Tests

Splitkriterium: Itemscoremedian	
Group 1	Group 2
Items: 1, 4, 6, 7, 9, 10, 11, 13, 16	Items: 2, 3, 5, 8, 12, 14, 15, 17, 18
Log-Likelihood: -1982.189	Log-Likelihood: -2522.458
Overall Rasch-Model: Log-Likelihood: -5309.692; LR-value: 95.378; Chi-square df: 139; p: 0.99	

Das Ergebnis zeigt, dass der Likelihoodquotient (-5309.692) zur Prüfung der Itemhomogenität bei df = 139 Freiheitsgraden nicht signifikant ist. Folglich kann die Itemhomogenität als zentrale Eigenschaft des Rasch-Modells bestätigt werden.

Lokale stochastische Unabhängigkeit

Lokale stochastische Unabhängigkeit zwischen Items bedeutet, dass für eine Person mit einer bestimmten Fähigkeitsausprägung die Wahrscheinlichkeit ein Item zu lösen nur von dem Item selbst abhängt und somit unabhängig von der Lösungswahrscheinlichkeit eines anderen Items ist (Koller et al., 2012, S. 17). Die Annahme der lokalen stochastischen Unabhängigkeit wäre z. B. verletzt, wenn die Lösung einer Aufgabe die richtige Beantwortung einer vorhergehenden Aufgabe voraussetzt (Rost, 2004, S. 69). Bei Nichterfüllung dieser Eigenschaft des Rasch-Modells würde sich dies in hohen Korrelationen zwischen den Items widerspiegeln (Inter-Itemkorrelation) (Koller et al., 2012, S. 17). Insbesondere bei der Aufgabenkonstruktion ist somit zu berücksichtigen, dass jede Aufgabe unabhängig von den anderen Aufgaben im Test zu lösen ist (Strobl, 2012, S. 18). Zu bedenken ist in diesem Zusammenhang auch, dass Items, die lokal stochastisch abhängig sind, keine zusätzliche Information für die Bestimmung der Fähigkeitsausprägung liefern (Koller et al., 2012, S. 18). Auch die Unabhängigkeit von Personen ist für die Gültigkeit des Rasch-Modells von Bedeutung (Strobl, 2012, S. 19). Somit darf die Wahrscheinlichkeit, mit der eine bestimmte Person fähig ist eine Aufgabe zu lösen, nicht systematisch davon abhängen, ob eine andere Person dazu in der Lage ist diese Aufgabe zu lösen. Für die Testdurchführung ist daher zentral, dass es Personen beispielsweise nicht möglich ist, während der Testung voneinander abzuschreiben, um eine Verletzung der Annahme der lokalen stochastischen Unabhängigkeit zu vermeiden. Zur Überprüfung der Annahme der lokalen stochastischen Unabhängigkeit auf Itemebene wurden mithilfe der Statistiksoftware SPSS die Korrelationskoeffizienten paarweise zwischen den einzelnen Items berechnet. Da alle Inter-Itemkorrelationen Werte unter 0,3 aufweisen und somit als gering einzustufen sind, kann die Forderung nach lokaler stochastischer Unabhängigkeit als erfüllt angesehen werden (Bühl, 2008, S. 269). Dieses Ergebnis ist als erwartungskonform einzustufen, da bereits bei der Testkonstruktion darauf geachtet wurde, dass die Items unabhängig voneinander zu lösen sind.

Spezifische Objektivität

Spezifische Objektivität bedeutet, dass die Schätzung der Fähigkeitsparameter unabhängig davon ist, an welchen Aufgaben diese verglichen werden (Strobl, 2012, S. 20). Umgekehrt gilt dies auch für den Vergleich von zwei Aufgaben, welcher unabhängig von der Wahl der Person sein soll. Folglich ist es irrelevant, welche Personen aus einer definierten Stichprobe gezogen werden, um zwei Items bezüglich ihrer Schwierigkeit zu vergleichen (Koller et al., 2012, S. 19). Umgekehrt ist es aber ebenso irrelevant, welche Items aus dem Itempool gezogen werden, um zwei Personen in Bezug auf ihre Fähigkeit zu vergleichen. Ein zentraler Aspekt der subjektiven Objektivität auf Personenseite ist die Subgruppeninvarianz (Koller et al., 2012, S. 20). Die Annahme der Subgruppeninvarianz ist erfüllt, wenn sich die Schätzung der Personenfähigkeiten über unterschiedliche Subgruppen von Personen, wie im vorliegenden Fall zwischen Industriekaufleuten und Kaufleuten für Spedition und Logistikdienstleistungen, nicht verändert. Eine Verletzung der Annahme der Subgruppeninvarianz liegt vor, wenn die Items bzw. einzelne Items für unterschiedliche Personengruppen unterschiedlich schwierig sind, da beispielsweise je nach Personengruppe unterschiedliche Fähigkeiten bei der Beantwortung der Items angesprochen werden. Die Überprüfung der Subgruppeninvarianz kann anhand eines internen Teilungskriteriums wie dem Median

der Personenscores oder anhand eines externen Teilungskriteriums (z. B. Geschlecht, Ausbildungsberuf) erfolgen. Letzteres wird in der Literatur als die Überprüfung von Differential Item Functioning bezeichnet und wird in Abschnitt 3 ausführlich beleuchtet. Zur Prüfung der Subgruppeninvarianz anhand eines externen Teilungskriteriums kann der Anderson-Likelihood-Test herangezogen werden, der die Modellgültigkeit global, d. h. für alle Items simultan, prüft (Koller et al., 2012, S. 19, 67). Rost (2004, S. 347) spricht in diesem Zusammenhang auch von einem Test auf Personenhomogenität. Als eines der am häufigsten verwendeten internen Teilungskriterien für die Personenstichprobe gilt der Summenscore der Personen, sodass dieser auch hier zur Teilung der Stichprobe herangezogen wird, um die Itemparameterschätzungen in zwei Scoregruppen miteinander vergleichen zu können. Die Grundidee dieses globalen Modellgeltungstests ist, dass die Likelihood für den gesamten Datensatz und auch die Likelihoods getrennt für die zwei Gruppen berechnet werden (Koller et al., 2012, S. 67; Rost, 2004). Mittels des Likelihoodquotienten, der als approximativ verteilte χ^2-Prüfgröße interpretiert werden kann, erfolgt eine Signifikanzprüfung (Bühner, 2011, S. 532). Die Ergebnisse des durchgeführten Anderson-Tests mit dem Splitkriterium Median sind unten stehender Tabelle zu entnehmen:

Tab. 2. Ergebnisse des Anderson-LR-Tests

Andersen-LR-Test (Splitkriterium: Median)	
LR-value:	94.16
Chi-square df:	23
p-value:	0

Der ermittelte Wert der Teststatistik von 94,16 zur Prüfung der Subgruppeninvarianz auf globaler Ebene ist bei 23 Freiheitsgeraden signifikant, sodass eine zentrale Annahme des Rasch-Modells verletzt ist.

Dieses Ergebnis zeigt, dass das entwickelte Testinstrument für verschiedene Subgruppen unterschiedlich schwierig ist und hiermit eine zentrale Annahme des Rasch-Modells als verletzt gilt. Da es sich bei dem Anderson-Test jedoch um einen globalen Modelltest handelt, kann noch nicht festgestellt werden, für welche Personengruppen sich die unterschiedlichen Schwierigkeiten ergeben. Um dies zu analysieren, erfolgt in Abschnitt 2.2 die Prüfung eines Mixed Rasch-Modells. Dies lässt als eines der zentralen Mischverteilungsmodelle unterschiedliche Schwierigkeiten für verschiedene Personengruppen zu (Rost, 2004). Um darüber hinaus auf Itemebene zu prüfen, welche Items unterschiedliche Schwierigkeiten aufweisen, erfolgt darauf aufbauend die Überprüfung von Differential Item Functioning in Abschnitt 3.

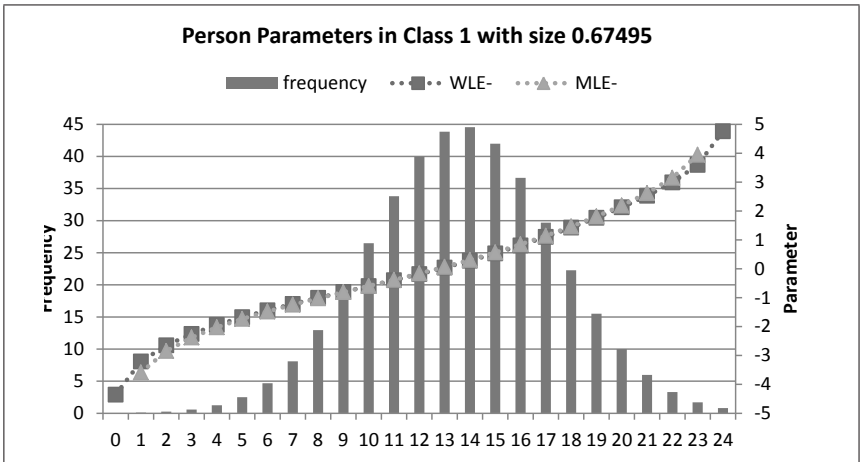
2.2 Ergebnisse der Mixed Rasch Analysen

Die zentrale Funktion des Mixed Rasch-Modells ist es, Personen so zu klassifizieren, dass innerhalb jeder Klasse eine quantitative Personenvariable mittels des Rasch-Modells gemessen wird (Rost, 2004, S. 174 f.). Das Mixed Rasch-Modell stellt eine Kombination des Rasch-Modells und der latenten Klassenanalyse dar. Zentrale Annahme ist, dass innerhalb jeder latenten Klasse das Rasch-Modell gilt. Im Vergleich zum normalen Rasch-Modell, in dem angenommen wird, dass dieselben Itemparameter

für alle Personen in der befragten Population gelten und somit die Itemschwierigkeiten für alle getesteten Personen konstant sind, erlaubt das Mixed Rasch-Modell unterschiedliche Itemschwierigkeiten für verschiedene Personengruppen. Somit wird die restriktive Annahme konstanter Itemschwierigkeiten, die oft dazu führt, dass das Rasch-Modell für einen Datensatz verworfen werden muss, umgangen. Die Klassen sind im Voraus jedoch nicht bekannt und werden bestimmt, indem Personen mit sich maximal unterscheidenden Antwortmustern gesucht und in verschiedene Teilstichproben eingeteilt werden (Bühner, 2011, S. 509 ff.; Rost, 2004). Mithilfe von Modellgeltungstests kann anschließend geprüft werden, welches Mixed Rasch-Modell mit welcher Anzahl von Klassen die beste Anpassung an die Daten liefert und ob das Ergebnis mit der Anzahl theoretisch angenommener, qualitativ unterschiedlicher Klassen, wie im vorliegenden Fall der Klasse der Industriekaufleute und der Klasse der Kaufleute für Spedition und Logistikdienstleistungen, übereinstimmt.

Zur Berechnung des Mixed Rasch-Modells wurde das Programm Winmira (von Davier, 2001) verwendet, das speziell für solche Analysen entwickelt wurde. Dem Projektziel entsprechend wird die Passung eines ordinalen Mixed Rasch-Modells mit zwei Klassen geprüft. Im Rahmen der Modelltests wird diese Lösung gegen das Mixed Rasch-Modell mit drei Klassen getestet. Die folgenden Analysen[5] zeigen die Ergebnisse der Berechnung getrennt nach Klassen:

Klasse 1



WLE estimates:	Mean = 0.304 Var = 0.902 stdev = 0.950
Marginal error variance:	= 0.270 stdev = 0.520
Anova reliability:	= 0.770
Andrichs reliability:	= 0.701
WLE:	Warm's modified likelihood estimates
MLE:	Standard maximum likelihood estimates
Raw-Score:	Mean = 13.698 Stdev = 3.605

Abb. 1. Ergebnisse für Klasse 1

5 Vgl. zur Interpretation des Outputs einer Berechnung eines Mixed Rasch-Modells mit der Software Winmira von Davier (2001) sowie Bühner (2011).

Die obere Abbildung verdeutlicht, dass der ersten Klasse rund 67 % der Befragten zugeordnet werden. Gleichzeitig zeigt sie die erwarteten Häufigkeiten von Personen mit der jeweiligen Anzahl an richtigen Antworten (rawscore). Folglich können erwartungsgemäß 0,04 % der Personen keine einzige richtige Antwort (rawscore = 0) geben. Darüber hinaus ist zu erwarten, dass 0,1 % der Personen eine richtige Antwort geben (rawscore = 1). 39,98 % der Probanden dieser Klasse erreichen voraussichtlich 12 Punkte und können somit die Hälfte des Tests lösen. Nur 0,81 % erreichen erwartungsgemäß mit 24 Punkten die volle Punktzahl. Die Schätzung der Personenparameter erfolgt sowohl mit dem MLE- als auch mit dem WLE-Schätzer. Im Rahmen der Ergebnisdarstellung sollte der WLE dem MLE vorgezogen werden, da dieser auch die Schätzung der Personenparameter für die extremen Summenwerte von 0 und 24 Punktwerten ermöglicht (vgl. Bühner, 2011, S. 565; Rost, 2004). Bei der Interpretation dieses Parameters wird deutlich, dass eine der ersten Klasse zugeordnete Person mit keiner richtigen Antwort eine Fähigkeit von -4,356 aufweist. Die durchschnittliche Fähigkeit der Personen, die dieser Klasse angehören, gibt der „WLE-estimate mean“ an. In diesem Fall liegt die durchschnittliche Fähigkeit aller Personen bei 0,304. Die durchschnittliche Anzahl richtiger Antworten, die dem Wert „raw-score mean“ entnommen werden kann, beträgt ca. 13,7.

Der im Anhang abgebildeten Tabelle 1 (s. Anhang 1) können die erwarteten Häufigkeiten der Antwortkategorien für die 18 Items des Logistiktests entnommen werden. Erwartungsgemäß können ca. 65 % der Personen dieser Klasse die erste Frage richtig beantworten. Folglich handelt es sich um ein vergleichbar einfaches Item, was auch aus dem Schwierigkeitsparameter mit einem negativen Wert von -0,40179 hervorgeht. Zu einer der schwierigeren Fragen zählt hingegen z. B. Aufgabe 7 (vgl. Anhang 1). Abbildung 2 verdeutlicht das Ergebnis in grafischer Form, da die Antwortwahrscheinlichkeiten auf die jeweiligen 18 Items getrennt nach der latenten Klasse dargestellt sind.

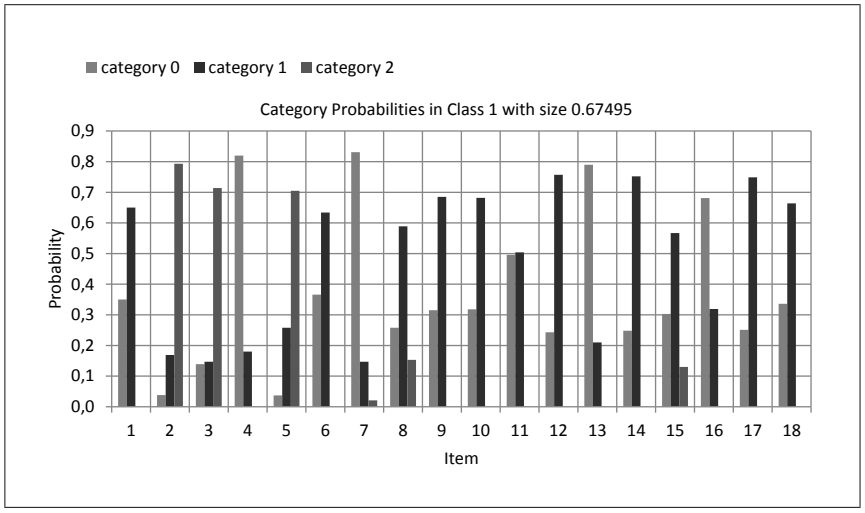


Abb. 2: Antwortwahrscheinlichkeiten für die Items des Moduls „Logistik“ für Klasse 1

Auch der Item-Fit kann getrennt für die Klassen analysiert werden (s. Tab. 3). Dazu wird im Rahmen des Mixed Rasch-Modells der sogenannte Q-Index herangezogen, der angibt, wie wahrscheinlich die Antwortmuster eines Items unter den gegebenen Modellparametern sind (Bühner, 2011, S. 573, 595; von Davier, 2001, S. 75 ff.). Als Normalwert für den Q-Index kann der Bereich zwischen 0,1 und 0,3 angesehen werden, in dem sich im vorliegenden Beispiel alle Items bewegen.

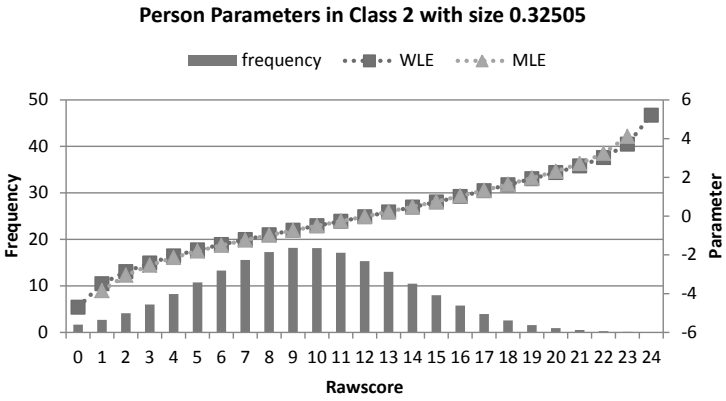
Tab. 3: Item-Fit-Werte des Q-Index für Klasse 1

Itemlabel	Q-Index
AUF1	0.2967
AUF2	0.2909
AUF3	0.2031
AUF4	0.2668
AUF5	0.2706
AUF6	0.2688
AUF7	0.2744
AUF8	0.2143
AUF9	0.2006
AUF10	0.1997
AUF11	0.2676
AUF12	0.1873
AUF13	0.2893
AUF14	0.1920
AUF15	0.2027
AUF16	0.2642
AUF17	0.1832
AUF18	0.2055

Klasse 2

Der nachstehenden Abbildung ist zu entnehmen, dass ca. 33 % der Befragten der Klasse 2 zugeordnet werden. Es wird erwartet, dass 1,67 % der Personen keine richtige Antwort (rawscore = 0) geben können, während 0,06 % der Personen voraussichtlich mit 24 Punkten die volle Punktzahl erzielen. Eine der zweiten Klasse zugeordnete Person mit keiner richtigen Antwort weist eine Fähigkeit von -4,7 auf. Die durchschnittliche Fähigkeit aller Personen der Klasse 2 liegt bei -0.637. Die durchschnittliche Anzahl richtiger Antworten beträgt in dieser Klasse ca. 9,6. Diese Klasse enthält im Vergleich zu Klasse 1 folglich Personen, die im Hinblick auf ihre Leistungsfähigkeit als schwächer einzustufen sind.

Der Tabelle der erwarteten Häufigkeit der Antwortkategorien (s. Anhang 2) kann entnommen werden, dass rund 1,5 % der Personen dieser Klasse die erste Frage richtig beantworten können. Somit ist diese Frage innerhalb dieser Klasse als die schwierigste anzusehen, wohingegen sich Item 14 als leichteste Aufgabe ergibt. Grafisch wird dieses Ergebnis durch Abbildung 4 verdeutlicht. Darüber hinaus liegt auch der Q-Index als Item-Fit-Index für alle Items im Normbereich.



WLE estimates:	Mean = -0.637 Var = 1.307 stdev = 1.143
Marginal error variance:	= 0.301 stdev = 0.549
Anova reliability:	= 0.813
Andrichs reliability:	= 0.770
WLE:	Warm's modified likelihood estimates
MLE:	Standard maximum likelihood estimates
Raw-Score:	Mean = 9.562 Stdev = 4.166

Abb. 3: Ergebnisse für Klasse 2

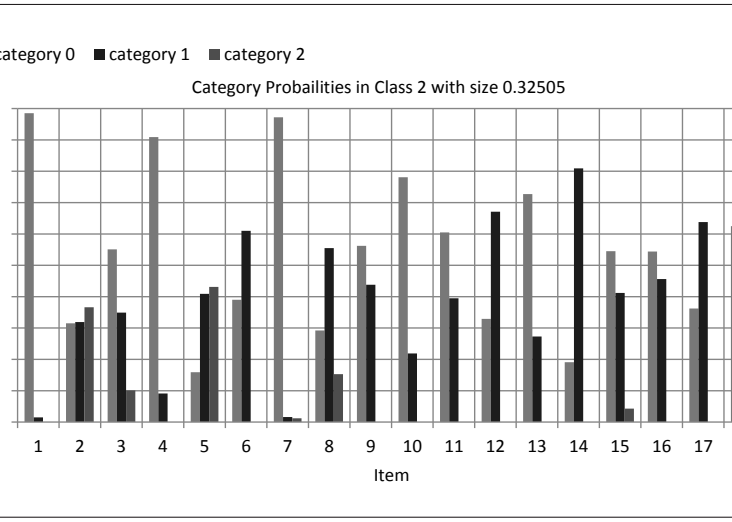


Abb. 4: Antwortwahrscheinlichkeiten für die Items des Moduls „Logistik" für Klasse 2

Neben den Item-Fit-Werten geben auch Personen-Fit-Werte Auskunft über die Modellpassung (Bühner, 2011, S. 579, 599). So lässt sich die Anpassungsgüte des

Mixed Rasch-Modells an die Antwortmuster der Befragten beider Klassen anhand der „person fit index descriptives“ (s. Tab. 4) bewerten. Alle Personenfitindizes stellen z-Werte dar, wobei Werte kleiner als -1,96 oder größer als +1,96 als kritisch anzusehen sind. Werte innerhalb dieser Intervalle weisen auf eine gute Modellanpassung hin. Im Beispiel kann folglich von einer guten Modellanpassung ausgegangen werden.

Tab. 4: Personenfitindizes für Anpassungsgüte des Mixed Rasch-Modells

Person fit index descriptives	
mean	0.0424498
st. dev.	1.0208077
skewness	-0,2042913
kurtosis	0.0826391

Die Beurteilung der Modellgüte auf globaler Ebene kann anhand eines parametrischen Bootstrappings durchgeführt werden (s. Tab. 5) (Bühner, 2011, S. 600 f.; von Davier, 2001, S. 34 ff.). Hierbei sind vorrangig die Cressie-Read-Prüfgröße (CressieRead) und der Pearson χ^2-Wert beziehungsweise deren p-Werte (p-values) zu beachten. Sollten sich signifikante Prüfwerte ergeben, ist das untersuchte Modell zu verwerfen. Da sich in diesem Beispiel keine signifikanten Werte ergeben ($\alpha = 0{,}05$), kann das Mixed Rasch-Modell grundsätzlich angenommen werden, was eine gute Anpassung der erwarteten an die beobachteten Antwortmuster der beiden Klassen bescheinigt.

Tab. 5: Parametric Bootstrap estimates for Goodness of Fit für Anpassungsgüte des Mixed Rasch-Modells

P. Bootstrap estimates for Goodness of Fit	CressieRead	Pearson X^°2
Z:	2.073	1.315
P (X > Z):	0.019	0.094
Mean:	105502.150	2897431.957
Stdev:	31628.730	3823080.735
p-values (emp. PDF):	0.063	0.088

Wird das berechnete Mixed Rasch-Modell nach einem Modelltest wie im vorliegenden Fall beibehalten, ist im weiteren Verlauf noch zu prüfen, inwiefern das hier berechnete Mixed Rasch-Modell mit 2 Klassen eine höhere Erklärungskraft hat als vergleichbare Modelle, wie hier die 3 Klassen Lösung. Dies kann anhand eines Vergleichs der Informationskriterien AIC, BIC, CAIC (s. Tab. 6) festgestellt werden (Rost, 2004). Deutlich wird, dass das Mixed Rasch-Modell mit 2 Klassen am besten auf die Daten passt, da die Werte der drei Informationskriterien am niedrigsten sind.

Tab. 6: Vergleich der Informationskriterien zur Passung eines Mixed Rasch-Modells mit zwei und mit drei Klassen

	AIC	BIC	CAIC
Mixed Rasch-Modell mit zwei Klassen	13983.61	14313.63	14388.63
Mixed Rasch-Modell mit drei Klassen	14001.76	14498.99	14611.99

Die gemeinsamen Eigenschaften der Personen innerhalb einer Klasse müssen im Rahmen der Mixed Rasch Analyse jedoch post-hoc ermittelt werden (Rost, 1990). So können vorherige Vermutungen über die potentiellen Klassen abgesichert werden. In der vorliegenden Analyse stellt dies jedoch kaum eine Herausforderung dar, denn bereits die Größe der Klassen lässt vermuten, dass es sich bei diesen um die beiden Subgruppen Industriekaufleute und Speditionskaufleute handelt. So setzt sich das N von 602 Auszubildenden aus ca. 1/3 Industriekaufleuten und 2/3 Speditionskaufleuten zusammen. Um die Interpretation der Klassen abzusichern, wurde darüber hinaus eine nach Beruf getrennte Skalierung durchgeführt. Werden die WLE-Schätzer für die Personenfähigkeit miteinander verglichen, wird deutlich, dass sich die Personenparameter der Klasse 1 und diejenigen, die sich im Rahmen der Skalierung ausschließlich für Kaufleute für Spedition und Logistikdienstleistungen ergeben, weitgehend entsprechen. Für Klasse 2 und die Skalierung der Testdaten für die Gruppe der Industriekaufleute zeigt sich ein vergleichbares Bild. Somit kann das Mixed Rasch-Modell nicht nur aus testtheoretischer Sicht, sondern auch inhaltlich als passendes Modell für die Daten angesehen werden.

3 Analysen zum Differential Item Functioning

Wie bereits in Kapitel 2 erwähnt, beschreibt Differential Item Functioning (DIF) die unterschiedlichen Itemfunktionsweisen in Abhängigkeit einer Stichprobensubgruppe (Osterlind & Everson, 2009). Da bereits der Anderson-Test im Rahmen der Modellgeltungsprüfung (vgl. Kapitel 2) bestätigt hat, dass sich Unterschiede zwischen Subgruppen ergeben und diese Gruppen mithilfe des Mixed Rasch-Modells identifiziert werden konnten, sind nun Analysen auf Itemebene erforderlich. So ist zu prüfen, welche Items DIF aufweisen und was die Items auszeichnet, die unterschiedliche Lösungswahrscheinlichkeiten für die beiden Ausbildungsberufe haben. Für eine derartige Modellierung von DIF sind unterschiedliche Methoden anwendbar. Eine der verbreitetsten und im Folgenden eingesetzte Vorgehensweise ist eine zweischrittige Methode, in der zuerst DIF-Analysen durchgeführt werden, um die entsprechenden DIF-Parameter in einem nächsten Schritt mittels verschiedener Prädiktoren auf Itemebene per linearer Regression vorherzusagen (z. B. Scheuneman & Gerritz, 1990; Jurecka, 2010). Grundlegende Arbeiten auf diesem Gebiet verfassten Scheuneman & Gerritz (1990) sowie darauf aufbauend Jurecka (2010), deren Vorgehensweisen und Ausführungen den Orientierungsrahmen für die im Folgenden dargestellten Analysen bilden.

3.1 Differential Item Functioning im Logistiktool des CoBALIT-Tests

In Anlehnung an Zumbo (2007, S. 224) kann DIF folgendermaßen definiert werden: „DIF was the statistical term that was used to simply describe the situation in which persons from one group answered an item correctly more often than equally knowledgeable persons from another group.“ Hieraus folgt, dass die Mitglieder zweier oder mehr Gruppen eine unterschiedliche Wahrscheinlichkeit aufweisen, ein Item korrekt zu lösen, obwohl sie sich in Bezug auf die zu messende latente Fähigkeit auf dem gleichen Niveau befinden (Embretson & Reise, 2000). Die sich ergebenden unterschiedlichen Lösungswahrscheinlichkeiten sind in diesem Zusammenhang

folglich ausschließlich durch die Gruppenzugehörigkeit bedingt (Ackerman, 1992; Roussos & Stout, 1996). Solche gruppenspezifischen Faktoren könnten im Rahmen des CoBALIT Projektes beispielsweise darin begründet sein, dass innerhalb der beiden Ausbildungsberufe bestimmte Ausbildungsinhalte unterschiedlich intensiv in der Berufsschule und im Betrieb behandelt wurden. Insgesamt ist zu berücksichtigen, dass Items, die Differentielle Item Funktionen (DIF) aufweisen, unter Umständen nicht oder nur eingeschränkt zum fairen Vergleich von Testleistungen verwendet werden können (Holland & Wainer, 1993). Signifikante DIF-Werte sind nicht nur problematisch in Bezug auf die Testfairness, sondern stellen strenggenommen auch eine Verletzung der Modellannahme der spezifischen Objektivität im Rasch-Modell dar, wie sie bereits in Kapitel 2 nachgewiesen werden konnte und hier noch einmal auf Itemebene zu betrachten ist (vgl. Abschnitt 2). Zu betonen ist jedoch, dass es kaum ein Testinstrument gibt, das keinen DIF aufweist und das aktuell noch keine allgemein etablierten Konventionen für die Beurteilung des Ausmaßes von DIF und akzeptable Toleranzgrenzen vorliegen (Osterlind & Everson, 2009). Gleichzeitig ist eine andere, selten herangezogene Interpretation, dass DIF nicht mit einer Verringerung der Kriteriumsvalidität gleichzusetzen ist, sondern sich im Rahmen der Analysen lediglich unterschiedliche Profile von Stärken und Schwächen der Gruppen in Bezug auf das gleiche Konstrukt zeigen (Scheuneman & Gerritz, 1990). Diese werden durch die Gruppenzugehörigkeit manifestiert, beispielsweise durch differentielle Lerngelegenheiten in den unterschiedlichen Systemen (z. B. Scheuneman & Gerritz, 1990; Klieme & Baumert, 2001). DIF kann somit als diagnostisches Instrument verstanden werden, welches der Analyse solcher differentieller Stärken und Schwächen dient und ist damit auch ein Maß für die Validität eines Testinstruments.

Die in diesem Beitrag vorzustellenden DIF-Analysen auf der Basis des Item-Response-Modells wurden mithilfe von ConQuest (Wu, Adams, Wilson & Haldane, 2007) durchgeführt, um die beiden Subgruppen, bestehend aus den zwei Ausbildungsberufen, zu vergleichen. Die Ergebnisse der Berechnungen sind entsprechende DIF-Parameter, die eine Schätzung des Schwierigkeitsunterschieds dieser Items (in Logits) für die Subgruppen darstellen (Wu et al., 2007). Anhand des Standardfehlers, der für jeden DIF-Parameter errechnet wird, ist festzustellen, ob die Differentielle Item Funktion signifikant ist. Ist der DIF-Parameter eines Items mindestens doppelt so groß wie der dazugehörige Standardfehler, liegt zwischen den Subgruppen ein signifikanter Unterschied hinsichtlich der Itemschwierigkeit vor. Die Ergebnisse der DIF-Analyse sind folgender Tabelle zu entnehmen:

Tab. 7: Ergebnisse der DIF Analysen zwischen Speditionskaufleuten (0) und Industriekaufleuten (1)

Term 2: (-)Ausbildungsgang								
Variables			UNWEIGHTED FIT			WEIGHTED FIT		
Ausbildungsgang	ESTIMATE	ERROR	MNSQ	CI	T	MNSQ	CI	T
0	0.505	0.071						
1	-0.505*							
An asterisk next to a parameter estimate indicates that it is constrained Separation Reliability Not Applicable Chi-square test of parameter equality = 49.96, df = 1								

Die Tabelle zeigt die Werte für die berufsspezifischen Unterschiede in der Fähigkeitsschätzung, wobei der Ausbildungsberuf „Kaufmann für Spedition und Logistikdienstleistungen“ mit 0 codiert wurde und der Ausbildungsberuf „Industriekaufmann“ mit 1 (Wu et al., 2007). Ein negatives Vorzeichen vor dem Bildungsgang „Industriekaufmann“ zeigt, dass diese Auszubildenden um 1.01 Logits schlechter im Test abschneiden als die Kaufleute für Spedition und Logistikdienstleistungen. Die Parameterschätzung ist ca. 7-mal größer als die Standardfehlerschätzung, sodass die Differenz zwischen den Ausbildungsberufen signifikant ist. Der Chi-Quadrat-Wert von 49,96 mit einem Freiheitsgrad entspricht diesem Befund.

Die nachstehende Tabelle 8 stellt darüber hinaus die Interaktion zwischen den einzelnen Items und dem Ausbildungsberuf/Bildungsgang dar. Aus den Kennwerten wird ersichtlich, dass insgesamt 10 der 18 Items DIF in Bezug auf den Ausbildungsberuf aufweisen (Wu et al., 2007). Während beispielsweise für die Auszubildenden des Ausbildungsberufs „Kaufmann für Spedition und Logistikdienstleistungen“ von der Schwierigkeit des Items 1 der Wert 1.141 abgezogen werden muss, ist dies für die Industriekaufleute genau umgekehrt. Hier muss entsprechend bei Item 1 der Wert 1.141 zu der Schwierigkeit addiert werden. Folglich ist das Item deutlich leichter für Auszubildende des erstgenannten Ausbildungsberufs. Würden folglich beispielsweise alle Items so wie Item 1 angelegt sein, würde das bedeuten, dass sich die geschätzte mittlere Punktzahl, die die Kaufleute für Spedition und Logistikdienstleistungen im Test erreichen, um 2.282 erhöht, für die Industriekaufleute entsprechend sinkt. Ein vergleichbarer Effekt zeigt sich auch für die Items 3 und 10. Im Gegensatz dazu ist festzustellen, dass die Items 4, 6, 8, 13, 14, 15, 16 leichter für Industriekaufleute zu lösen sind als für Kaufleute für Spedition und Logistikdienstleistungen. Allerdings wird deutlich, dass DIF hier bei den meisten Items nicht so stark ausgeprägt ist, wie dies bei den Items der Fall ist, die die Kaufleute für Spedition und Logistikdienstleistungen bevorteilen. Beispielweise würde sich die geschätzte mittlere Punktzahl, die Industriekaufleute im Test erzielen, nur um 0.172 erhöhen, wenn alle Items so funktionieren würden, wie das Item 5. Insgesamt ergibt sich die Signifikanz dieser aufgezeigten Ergebnisse bei diesen 10 DIF-Items dadurch, dass der Estimate für jedes betroffene Item wertmäßig mehr als doppelt so groß ist wie der Standardfehler. Gleichzeitig ist der signifikante Chi-Quadrat-Test (291.39, df = 17) ebenfalls ein Beweis für das Vorliegen von Differential Item Functioning innerhalb des entwickelten Testinstruments (Wu et al., 2007).

Der Befund des Bildungseffekts kann über die vorgestellten Berechnungen hinaus auch mithilfe der Item Characteristic Curves (ICCs) veranschaulicht werden, da sich diese für die DIF-Items und die getesteten Subgruppen unterscheiden (Strobl, 2012). Dargestellt sind die Item-Charakteristik-Kurven nachstehend für Item 1, das auf eine Lademeterberechnung abzielt. Die Kaufleute für Spedition und Logistikdienstleitungen (obere Line) haben, trotz ansonsten gleicher Fähigkeit, eine deutlich höhere Wahrscheinlichkeit das Item zu lösen als Industriekaufleute (untere Linie) (vgl. auch ähnliche Analysen und Interpretationen bei Wu et al., 2007, S. 84). Der Unterschied in der Lösungswahrscheinlichkeit wird durch die Verschiebung der ICCs auf der y-Achse deutlich (vgl. nachstehende Abbildung 5).

Nicht zu vernachlässigen ist jedoch auch, dass 8 Items keinen DIF aufweisen, was vor dem Hintergrund der Zielsetzung des Projektes so zu interpretieren ist, dass die Übertragbarkeit eines Testarrangements, das für einen spezifischen kaufmännischen Beruf entwickelt worden ist, auf einen anderen kaufmännischen Beruf mit

Tab. 8: Ergebnisse der DIF-Analysen auf Itemebene zwischen Speditionskaufleuten (0) und Industriekaufleuten (1)

Term 3: item*Ausbildungsgang										
Variables						UNWEIGHTED FIT			WEIGHTED FIT		
item		Ausbildungs-gang		ESTIMATE	ERROR	MNSQ	CI	T	MNSQ	CI	T
1	Lademeter	1	0	-1.141	0.152						
2	Transportdok.	1	0	-0.127	0.087						
3	Lenkzeit	1	0	-0.483	0.085						
4	Transportkalk.	1	0	0.316	0.135						
5	Maut	1	0	-0.086	0.100						
6	Lagerbestand	1	0	0.401	0.103						
7	Zoll	1	0	-1.134	0.605						
8	ABC-Analyse	1	0	0.554	0.087						
9	Wasserwege	1	0	-0.111	0.103						
10	IncotermFCL	1	0	-0.768	0.116						
11	Dokumente	1	0	0.181	0.102						
12	Kilometersatz	1	0	0.121	0.108						
13	Incoterms	1	0	0.485	0.114						
14	Just-in-Time	1	0	0.645	0.118						
15	Ruhezeiten	1	0	0.018	0.108						
16	Frachtführer	1	0	0.607	0.102						
17	Frachtbrief	1	0	0.172	0.106						
18	Verkehrsträger	1	0	0.035*							
An asterisk next to a parameter estimate indicates that it is constrained/Separation Reliability = 0.897 Chi-square test of parameter equality = 291.39, df = 17, Sig Level = 0.000											

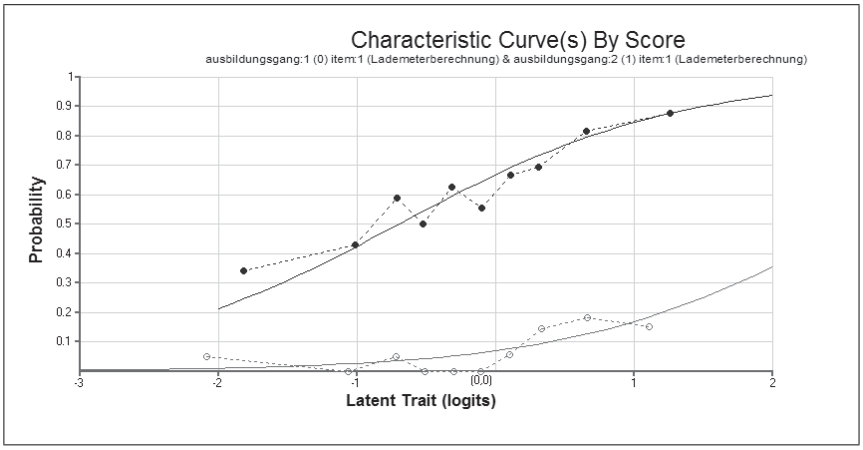


Abb. 5: Item-Charakteristik-Kurven für Speditionskaufleute (obere Linie) und Industriekaufleute (untere Linie) am Beispiel-Item 1

Einschränkungen gelingen kann und sich durchaus gemeinsame Kompetenzen in den betrachteten kaufmännischen Berufen identifizieren lassen. Um zu überprüfen, welche berufsübergreifenden Kompetenzen wie relevant sind, sind weitere Analysen notwendig, die im Rahmen dieses Beitrags nicht thematisiert werden. Vielmehr sollen im Folgenden die Unterschiede der Items und Berufsgruppen im Vordergrund stehen, um Erklärungen für DIF ableiten zu können.

3.2 Erklärung von DIF zwischen den untersuchten kaufmännischen Berufen

Im Folgenden werden auf Basis der Ergebnisse der DIF-Analyse ausgewählte Items mit signifikanten Abweichungen näher beschrieben und analysiert, um so unter Post-hoc-Betrachtung eine systematische Begründung für die DIF-Effekte abzuleiten. Neben diesem Vorgehen kann zur Erklärung eine Befragung der Auszubildenden zu den in der Schule und im Betrieb behandelten Inhalten herangezogen werden, die als Zusatzerhebung in das Projekt integriert wurde. Der eingesetzte Zusatzfragebogen enthielt genau die Themenbereiche, die auch Bestandteil des Assessments waren, das die Auszubildenden zu bearbeiten hatten. Über die deskriptiven Itemanalysen hinaus werden in einem nächsten Schritt die potentiellen Prädiktoren für DIF als unabhängige Variablen in eine lineare Regression aufgenommen, um zu überprüfen, ob sich diese tatsächlich als signifikante Indikatoren für DIF erweisen.

3.2.1 Deskriptive Itemanalysen

Die Ergebnisse der DIF-Analyse haben gezeigt, dass Speditionskaufleute auf der Ebene des Gesamttests signifikant besser abschneiden als Industriekaufleute. Es zeigte sich jedoch auch, dass für Kaufleute für Spedition und Logistikdienstleistungen nicht jedes Item signifikant einfacher ist. Im Gegenteil, es gibt sogar mehr Items, die Industriekaufleute begünstigen, wobei der Vorteil zugunsten der Industriekaufleute bei den entsprechenden Aufgaben nicht so stark ausgeprägt ist, wie dies bei jenen wenigen Items der Fall ist, die zugunsten der Kaufleute für Spedition und Logistikdienstleistungen ausfallen.

Ziel ist es nunmehr, zu untersuchen, wodurch bestimmte Items gekennzeichnet sind, die besonders gut von Kaufleuten für Spedition und Logistikdienstleistungen bzw. umgekehrt besser von Industriekaufleuten gelöst werden und somit berufsspezifische Differentielle Item Funktionen hinsichtlich der Schwierigkeitsparameter aufweisen. Hierzu werden im Folgenden zwei Items näher beschrieben, die als exemplarisch für das gesamte Logistiktool gelten können.

Wie bereits im vorherigen Kapitel verdeutlicht, erweisen sich drei der Logistikitems als einfacher lösbar für die Kaufleute für Spedition und Logistikdienstleistungen. Hierzu zählen Item 1, 3, und 10. Im Rahmen der offenen Aufgabe 1 sollen die Auszubildenden auf Basis einer vorgegebenen Ladeliste berechnen, ob der Laderaum eines Sattelzuges, der für einen Transport nach Großbritannien eingesetzt wird, vollständig ausgelastet ist. Im Zentrum steht somit die Prüfung, ob es möglich ist, noch weitere Waren in diesem Sattelzug nach Großbritannien zu transportieren. Vor dem Hintergrund der Befragung zur unterrichtlichen und betrieblichen Behandlung, deren Ergebnisse in nachstehender Tabelle dargestellt sind, ist das Ergebnis der DIF-Analyse nicht überraschend. So geben 98,3 % der befragten Speditionskauf-

leute an, sich mit Inhalten zum Thema Disposition von Frachtraum und Lademitteln in der Schule beschäftigt zu haben. Ein vergleichbares Ergebnis zeigt sich für die Ausbildung im Betrieb. Bei den Industriekaufleuten stellt es sich anders dar, so geben 54 % der befragten Auszubildenden dieses Berufes an, derartige Inhalte nie in der Schule thematisiert zu haben, während 69,7 % gleiches für den betrieblichen Teil der Ausbildung betonen.

Tab. 9: Subjektive Einschätzungen der Auszubildenden über die Behandlung itemspezifischer Themen in Schule und Betrieb

	Kaufleute für Spedition und Logistikdienstleistungen		Industriekaufleute	
	Behandlung in der Schule (%)	Behandlung im Betrieb (%)	Behandlung in der Schule (%)	Behandlung im Betrieb (%)
Item 1	98,3	96,4	56	30,3
Item 3	99,1	92,4	28,1	15,9
Item 4	100	95,4	28,4	22,7
Item 6	99,6	75,6	65,9	26,1
Item 8	78,7	68,3	98,5	39,2
Item 10	99,6	87,4	94	34,4
Item 13	99,6	87,4	94	34,4
Item 14	99,6	84,3	84,6	59,3
Item 15	99,1	92,4	28,1	15,9
Item 16	98,7	92,3	29,5	27

Obwohl sich zeigt, dass das entwickelte Instrument auf Ebene des Gesamttests einfacher für Kaufleute für Spedition und Logistikdienstleistungen ist, erweisen sich insgesamt 7 Items als leichter lösbar für die Industriekaufleute. Dies betrifft unter anderem Aufgabe 4. Hier sind die Auszubildenden für eine erste Kalkulation eines Transports von Kassel nach Antalya zuständig. Die Schwierigkeit liegt nicht nur in der Kalkulation an sich, sondern auch in der Auswahl der zu berücksichtigenden Kostenpositionen, die aus dargebotenen Mauttabellen und vergleichbaren Dokumenten zu entnehmen sind. Das Ergebnis der DIF-Analyse ist auf den ersten Blick überraschend, da die Industriekaufleute mehrheitlich angeben, dieses Thema weder in der Schule noch im Betrieb behandelt zu haben. Dagegen kreuzen 100 % der Kaufleute für Spedition und Logistikdienstleistungen an, diesen Inhalt in der Schule thematisiert zu haben. Für den betrieblichen Teil der Ausbildung ergibt sich ein Wert von 95,4 %. Obwohl dieses Ergebnis zunächst widersprüchlich erscheint, lässt sich die Bevorteilung der Industriekaufleute durch die komplexe Rechnung erklären, die im Rahmen dieser Aufgabe zu leisten ist. So sind die Industriekaufleute in Bezug auf die Eingangsvoraussetzungen, mit der sie in die Ausbildung einmünden, als kognitiv leistungsfähiger zu bezeichnen. Die folgende Tabelle verdeutlicht, dass Industriekaufleute höhere Eingangsvoraussetzungen aufweisen und folglich auch mehr Abiturienten in den Klassen vertreten sind.

Zudem zeigt sich auch in Bezug auf die mathematischen Fähigkeiten, dass die Speditionskaufleute unabhängig vom Schulabschluss mehrheitlich über mathematische Fähigkeiten im befriedigenden Bereich verfügen, während die Industriekauf-

leute deutlich häufiger Noten im guten bis sehr guten Bereich aufweisen. Zugute kommt den Industriekaufleuten darüber hinaus, dass sie innerhalb ihrer Ausbildung durchaus Kalkulationen durchführen müssen, auch wenn diese nicht explizit einen Transport betreffen. Zudem sind sie inhaltlich sehr wohl mit den einzelnen Kostenbestandteilen wie beispielsweise der Maut vertraut.

Tab. 10: Schulische Abschlüsse der beiden Ausbildungsgruppen

Schulabschluss	Industriekaufleute	Speditionskaufleute
Keine Angabe	11,7	7,3
Förderschulabschluss	0,6	0
Hauptschulabschluss	1,7	2,4
Mittlere Reife	41,3	43,7
Fachhochschulreife	15,1	22,2
Ohne Abschluss/Sonstiger	0	0,9
Abitur	29,6	23,4
Gesamt	100,0	100,0

Resümierend für das Logistiktool ist festzustellen, dass die Kaufleute für Spedition und Logistikdienstleistungen die speditionsspezifischeren Aufgaben besser lösen als die Industriekaufleute, es jedoch eine Reihe von Items gibt, deren Inhalte beiden Berufsgruppen ähnlich vertraut sind. In diesem Falle scheinen die Industriekaufleute vielfach von ihren höheren Eingangsvoraussetzungen und ausgeprägten mathematischen Fähigkeiten zu profitieren, sodass sich bei einer Reihe von Aufgaben des Logistiktools ein erklärbarer Vorteil für Industriekaufleute ergibt. Allerdings ist zu berücksichtigen, dass sich nicht für jedes Item eine eindeutige Interpretation des Ergebnisses der DIF-Analyse aufgrund von Itemeigenschaften oder Eingangsvoraussetzungen der jeweiligen Auszubildenden ableiten lässt.

3.2.2 Lineare Regression zur Analyse von DIF

Der letzte Schritt der Itemanalysen zielt darauf ab, DIF mithilfe der deskriptiv herausgearbeiteten Ergebnisse zu erklären. Hierzu wird eine lineare Regression mit dem Programm SPSS berechnet. Die im Rahmen der DIF-Analyse berechneten DIF-Parameter fungieren in diesem Zusammenhang als die abhängige Variable, die anhand von verschiedenen Itemmerkmalen vorhergesagt werden soll. Der Einfluss der Itemmerkmale wird hierbei einzeln geprüft.

Die Ergebnisse (s. nachstehende Tabelle 11) zeigen, dass lediglich das Merkmal „Behandlung in der betrieblichen Ausbildung der Speditionskaufleute“ als Maß für die betrieblichen Arbeits- und Lerngelegenheiten einen signifikanten Einfluss auf die DIF-Parameter und somit Erklärungskraft hat. Hier sind die Beta-Koeffizienten wie folgt zu interpretieren: Die Tatsache, dass der Inhalt eines Items Bestandteil der betrieblichen Ausbildung der Kaufleute für Spedition und Logistikdienstleistungen war, senkt den DIF-Parameter und folglich die Itemschwierigkeit um 0.45 Logits für diese Berufsgruppe. Dieses Ergebnis ist, vor dem Hintergrund, dass das Logistiktool innerhalb des Assessments Aufgaben für den Bereich der Spedition und Logistik-

dienstleistungen beinhaltet, nicht überraschend. Erwartungskonform ist ebenso, dass die Thematisierung im schulischen Unterricht keinen signifikanten Einfluss hat, da die Aufgaben einer handlungslogischen Struktur folgend in komplexe Arbeits- und Geschäftsprozesse eingebunden sind, die den Schülern vornehmlich aus der betrieblichen Praxis bekannt sein dürften. Insgesamt kann mithilfe des benannten Prädiktors 28,5 % der Varianz der DIF-Parameter zwischen den Ausbildungsberufen aufgeklärt werden.

Tab. 11: Ergebnisse der linear logistischen Regression zur Erklärung von Item-DIF zwischen den beiden kaufmännischen Berufsgruppen

Koeffizienten[a]

Modell	Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
	B	Standardfehler	Beta	t	sig.
1 (Konstante)	1,124	,459		2,449	,026
BehandlungBetrieb_SK	-,451	,178	-,534	-2,528	,022

R^2 = 0,285
a = abhängige Variable (Item)

Sicherlich hätten zur Modellierung von DIF weitere inhaltliche Prädiktoren einbezogen werden können, z. B. die konkreten Lerngelegenheiten in Schule und Betrieb, auch außerhalb curricularer Besonderheiten, um unterschiedliche Lösungswahrscheinlichkeiten bei ansonsten gleicher mittlerer Fähigkeit zwischen verschiedenen kaufmännischen Berufen zu erklären.

In jedem Fall verweisen die Ergebnisse darauf, dass es durchaus gelingen kann, über kaufmännische Kernkompetenzen berufsfachliche Kompetenzen in den verschiedenen kaufmännischen Berufen zu messen, allerdings erfordert dies – neben sehr sorgfältigen Analysen der Lerngelegenheiten – eine intensive Auseinandersetzung mit der psychometrischen Qualität der Aufgaben und Testmodelle, die einerseits den Kern über eine gemeinsame Metrik abbilden und andererseits Aussagen über spezifische Kompetenzprofile zulassen.

Literatur

Achtenhagen, F. & Winther, E. (2009). Konstruktvalidität von Simulationsaufgaben: Computergestützte Messung berufsfachlicher Kompetenz – am Beispiel der Ausbildung von Industriekaufleuten. Abschlussbericht für das Bundesministerium für Bildung und Forschung. Abgerufen am 06. März 2015 von http://www.bmbf.de/pub/Endbericht_ BMBF09.pdf

Ackerman, T. A. (1992). A Didactic explanation of item bias, item impact and item validity from a multidimensional perspective. Journal of Educational Measurement, 29 (1), 67–91.

Baethge, M., Achtenhagen, F., Arends, L., Babic, E., Baethge-Kinsky, V. & Weber, S. (2006). Berufsbildungs-Pisa. Machbarkeitsstudie. München: Franz Steiner.

Brötz, R., Peppinghaus, B., Schapfel-Kaiser, F. & Brings, C. (2009). Gemeinsamkeiten und Unterschiede kaufmännisch-betriebswirtschaftlicher Berufe (GUK) – Ausgangspunkte und Ziele des Forschungsprojekts. In R. Brötz & F. Schapfel-Kaiser, Anforderungen an kaufmännisch-betriebswirtschaftliche Berufe aus berufspädagogischer und soziologischer Sicht (S. 19–43). Bielefeld: Bertelsmann.

Brötz, R., Schapfel-Kaiser, F. & Schwarz, H. (2008). Berufsfamilien als Beitrag zur Stärkung des Berufsprinzips. BWP – Berufsbildung in Wissenschaft und Praxis, 4, 23–26.

Bühl, A. (2008). SPSS 16: Einführung in die moderne Datenanalyse (11. Aufl.). München: Pearson Studium.

Bühner, M. (2011). Einführung in die Test- und Fragebogenkonstruktion (3. Aufl.). München: Pearson Studium.

Bundesministerium für Bildung und Forschung (2008). Berufsbildungsbericht 2008. Abgerufen am 04. März 2015 von http://www.bmbf.de/pub/bbb_08.pdf

Deutsches PISA-Konsortium (Hrsg.) (2001). PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. Opladen: Leske und Budrich.

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

Jurecka, A. (2010). Zum Zusammenhang von Differentiellen Item Funktionen und Testkultur. Dissertation, Johann Wolfgang Goethe-Universität, Frankfurt am Main. Abgerufen am 28. September 2014 von http://publikationen.ub.uni-frankfurt.de/frontdoor/index/index/docId/20296

Holland, P. W. & Wainer, H. (1993). Differential Item Functioning. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hubley, A. M. & Zumbo, B. D. (2011). Validity and the Consequences of Test Interpretation and Use. Abgerufen am 07. März 2015 von file:///C:/Users/Admin/Downloads/Hubley_Zumbo_2011.pdf

Klieme, E. & Baumert, J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. European Journal of Psychology of Education, 16, 385–402.

Klieme, E. & Hartig, J. (2008). Kompetenzkonzepte in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs. Zeitschrift für Erziehungswissenschaft, Sonderheft 8. VS Verlag für Sozialwissenschaften, 11–29.

Klieme, E. & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG, Zeitschrift für Pädagogik, (52) 6, 2006, 876–903.

Klieme, E., Maag-Merki, K. & Hartig, J. (2007). Kompetenzbegriff und Bedeutung von Kompetenzen im Bildungswesen. In J. Hartig & E. Klieme, Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzmodelle (S. 5–15). Herausgegeben vom Bundesministerium für Bildung und Forschung (BMBF). Band 20. Bonn/Berlin 2007.

Kultusministerkonferenz (1996). Handreichungen für die Erarbeitung von Rahmenlehrplänen der Kultusministerkonferenz (KMK) für den berufsbezogenen Unterricht in der Berufsschule und ihre Abstimmung mit Ausbildungsordnungen des Bundes für anerkannte Ausbildungsberufe. Abgerufen am 06. März 2015 von http://www.kmk.org/doc/publ/handreich.pdf

Kultusministerkonferenz (2007). Erklärung der Kultusministerkonferenz gegen die Überspezialisierung in der dualen Berufsausbildung. Abgerufen am 04. März 2015 von http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2007/2007_02_28-Ueberspezialisierung-duale_Berufsausbildung.pdf

Koller, I., Alexandrowicz, R. & Hatzinger, R. (2012). Das Rasch Modell in der Praxis. Eine Einführung in eRm. Wien: Facultas.

Masters, G. N. (1982). A rasch model for partial credit scoring. Psychometrika, 47 (2), 149–174.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50 (9), 741–749.

Minnameier, G. (2006). Aspekte von ‚Fachkompetenz' – Kognitive Leistungen im Umgang mit Wissen. In G. Minnameier, & E. Wutttke (Hrsg.), Berufs- und wirtschaftspädagogische

Grundlagenforschung. Lehr-Lern-Prozesse und Kompetenzdiagnostik (S. 391–405). Festschrift für Klaus Beck. Frankfurt am Main: Peter Lang.

Monnier, M., Srbeny, C. & Tschöpe, T. (2014). Messung sozialer Kompetenzen am Beispiel Medizinischer Fachangestellter. berufsbildung, 146, S.10–12.

Nickolaus, R. & Seeber, S. (2013). Berufliche Kompetenzen: Modellierungen und diagnostische Verfahren. In A. Frey, U. Lissmann, & B. Schwarz (Hrsg.), Handbuch Berufspädagogische Diagnostik (S. 155–180). Weinheim und Basel: Beltz.

Norris, N. (1991). The trouble with competences. Cambridge Journal of Education, (21) 3, 1–11.

Osterlind, S. J. & Everson, H. T. (2009). Differential item functioning (2. Aufl.). Thousand Oaks, CA: Sage Publications.

Reetz, L. (1984). Wirtschaftsdidaktik: Eine Einführung in Theorie und Praxis wirtschaftsberuflicher Curriculumentwicklung und Unterrichtsgestaltung, Bad Heilbrunn.

Reinisch, H. & Götzl, M. (2013). Berufsgruppenbildung im Bereich kaufmännisch-betriebswirtschaftlicher Berufe aus historischer Sicht. BWP – Berufsbildung in Wissenschaft und Praxis, 3, 20–23.

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.

Roussos, L. & Stout, W. (1996). A multidimensionality-based DIF Analysis paradigm. Applied psychological measurement, 20, 355–371.

Rychen, D. S. & Salganik H. L. (Eds.) (2001). Defining and Selecting Key Competencies. Seattle, Toronto, Bern, Göttingen: Hogrefe.

Scheuneman, J. D. & Gerritz, K. (1990). Using Differential Item Functioning Procedures to Explore Sources of Item Difficulty and Group Performance Characteristics. Journal of Educational Measurement, 27 (2), 109–131.

Seeber, S. (2007). Berufsspezifische Fachleistungen in ausgewählten Berufen des Bereichs Wirtschaft und Verwaltung am Ende der Berufsausbildung. In R. Lehmann & S. Seeber (Hrsg.), ULME III. Untersuchung von Leistungen, Motivation und Einstellungen der Schülerinnen und Schüler in den Abschlussklassen der Berufsschulen (S. 107–157). Hamburg: Behörde für Bildung und Sport.

Seeber, S. (2011). Zur Messung beruflicher Kompetenzen auf der Grundlage der Item-Response-Theorie. In S. Bohlinger & G. Münchhausen (Hrsg.), Validierung von Lernergebnissen – Recognition and Validation of Prior Learning (S. 319–346). Bertelsmann: Bielefeld.

Strobl, C. (2012). Das Rasch-Modell. Eine verständliche Einführung für Studium und Praxis. München: Rainer Hampp Verlag.

Von Davier, M. (2001). WINMIRA 2001. Abgerufen am 14. November 2014 von http://208.76.84.140/~svfklumu/wmira/winmiramanual.pdf

Weinert, F. E. Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), Leistungsmessungen in Schulen (S. 17–31). Weinheim, Basel 2001.

Winther, E. (2010). Kompetenzmessung in der beruflichen Bildung. Bielefeld: Bertelsmann.

Wright, B. D. & Masters, G. N. (1982). Rating scale analysis. Chicago: MESA Press.

Wu, M. L., Adams, R. J., Wilson, M. & Haldane, S. A. (2007). ACER ConQuest. Version 2.0. Generalised Item Reponse Modelling Software. Camberwell, Victoria: ACER press.

Zumbo, B. D. (2007). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. Language Assessment Quarterly, 4 (2), 223–233.

Anhang 1

Tab. 1: Erwartete Häufigkeiten der Antwortkategorien der Items des Moduls „Logistik“ und Itemschwierigkeiten für Klasse 1

Expected category frequencies and item scores and threshold parameters:

Item label	Item's		Relative category frequencies			Treshold parameters	
	Score	Stdev	0	1	2	Item label	Item location
AUF1	0.65	0.48	0.350	0.650		AUF1	-0.40179
AUF2	1.76	0.51	0.038	0.169	0.793	AUF2	-1.60935
AUF3	1.58	0.72	0.139	0.147	0.714	AUF3	-0.79027
AUF4	0.18	0.38	0.820	0.180		AUF4	1.98182
AUF5	1.67	0.54	0.037	0.258	0.705	AUF5	-1.52220
AUF6	0.63	0.48	0.366	0.634		AUF6	-0.32481
AUF7	0.19	0.44	0.831	0.147	0.021	AUF7	2.53721
AUF8	0.89	0.63	0.258	0.589	0.153	AUF8	0.60778
AUF9	0.68	0.46	0.315	0.685		AUF9	-0.57882
AUF10	0.68	0.47	0.318	0.682		AUF10	-0.56408
AUF11	0.50	0.50	0.496	0.504		AUF11	0.27636
AUF12	0.76	0.43	0.243	0.757		AUF12	-0.97920
AUF13	0.21	0.41	0.790	0.210		AUF13	1.77045
AUF14	0.75	0.43	0.248	0.752		AUF14	-0.95031
AUF15	0.83	0.63	0.303	0.567	0.130	AUF15	0.80355
AUF16	0.32	0.47	0.681	0.319		AUF16	1.14463
AUF17	0.75	0.43	0.251	0.749		AUF17	-0.92743
AUF18	0.66	0.47	0.336	0.664		AUF18	-0.47354
SUM	13.7						

Anhang 2

Tab. 2: Erwartete Häufigkeiten der Antwortkategorien der Items des Moduls „Logistik“ und Itemschwierigkeiten für Klasse 2

Expected category frequencies and item scores and threshold parameters:

Item label	Item's		Relative category frequencies			Treshold parameters	
	Score	Stdev	0	1	2	Item label	Item location
AUF1	0.02	0.12	0.985	0.015		AUF1	3.91977
AUF2	1.05	0.82	0.315	0.319	0.366	AUF2	-0.67910
AUF3	0.55	0.67	0.551	0.349	0.101	AUF3	0.50453
AUF4	0.09	0.29	0.909	0.091		AUF4	1.96858
AUF5	1.27	0.72	0.159	0.409	0.431	AUF5	-1.36761
AUF6	0.61	0.49	0.390	0.610		AUF6	-1.14318
AUF7	0.04	0.25	0.972	0.016	0.012	AUF7	2.22352
AUF8	0.86	0.65	0.292	0.555	0.153	AUF8	-0.21908
AUF9	0.44	0.50	0.562	0.438		AUF9	-0.30515
AUF10	0.22	0.41	0.781	0.219		AUF10	0.85591
AUF11	0.39	0.49	0.605	0.395		AUF11	-0.09902
AUF12	0.67	0.47	0.329	0.671		AUF12	-1.47859
AUF13	0.27	0.45	0.727	0.273		AUF13	0.52638
AUF14	0.81	0.39	0.191	0.809		AUF14	-2.49035
AUF15	0.50	0.58	0.545	0.412	0.043	AUF15	0.96860
AUF16	0.46	0.50	0.544	0.456		AUF16	0.39040
AUF17	0.64	0.48	0.362	0.638		AUF17	-1.29241
AUF18	0.67	0.47	0.625	0.675		AUF18	-1.50239
SUM	9.56						

Anschrift der Autorinnen: Michelle Liedtke, M. Ed. in Wirtschaftspädagogik, Wissenschaftliche Mitarbeiterin an der Professur für Wirtschaftspädagogik und Personalentwicklung, Platz der Göttinger Sieben 5, 37073 Göttingen, E-Mail: michelle.liedtke@wiwi.uni-goettingen.de.
Prof. Dr. Susan Seeber, Universität Göttingen, Professur für Wirtschaftspädagogik und Personalentwicklung, Platz der Göttinger Sieben 5, 37073 Göttingen, E-Mail: susan.seeber@wiwi.uni-goettingen.de.