Referierte Beiträge

Christian Spoden / Andreas Frey / Raphael Bernhardt / Susan Seeber / Aileen Balkenhol / Birgit Ziegler

Differenzielle Domänen- und Itemeffekte zwischen Ausbildungsberufen bei der Erfassung allgemeiner schulischer Kompetenzen von Berufsschülerinnen und Berufsschülern

Kurzfassung: Vor dem Hintergrund, dass sich schulische Grundqualifikationen als erklärungsmächtig für den Ausbildungserfolg erwiesen haben, spielt die Erfassung von Kompetenzen in den Bereichen Lesen, Mathematik und Naturwissenschaften bei Berufsschülerinnen und Berufsschülern eine große Rolle. Im Zusammenhang mit der Entwicklung eines adaptiven Testinstrumentes zur Erfassung dieser Kompetenzen wird in dem vorliegenden Beitrag der Frage nach domänen- und item spezifischen Vorteilen zweier Gruppen von Ausbildungsgängen (kaufmännisch-verwaltende vs. gewerblich-technische Berufe) nachgegangen und diese zunächst im Rahmen des sogenannten Differential Item Functioning (DIF) und anschließend auf Basis einer qualitativen Analyse durch Inhaltsexperten untersucht. Die Ergebnisse verdeutlichen, dass zwar mittlere Leistungsunterschiede zwischen den Berufsgruppen vorliegen, jedoch nur einzelne DIF-Effekte bestehen, welche bei inhaltlicher Analyse jedoch unproblematisch erscheinen. Diese Befunde werden diskutiert hinsichtlich ihrer Implikationen für die diagnostische Verortung von Berufsschülerinnen und –schülern in Bezug auf allgemeine schulische Grundqualifikationen.

Abstract: Considering basic competencies as powerful predictors for achievements in apprenticeship, the assessment of student competencies in reading, mathematics and science has to be regarded as important in vocational education and training. Within the course of the development of three adaptive tests for the assessment of such competencies, domain- and item-specific effects between two groups of instruction courses (commercial and administrative jobs vs. technical and industrial jobs) are investigated in the current article both in terms of statistical differential item functioning (DIF) analyses and additional qualitative analyses by content matter experts. Results indicate differences in mean competence levels between the courses but few DIF effects exist and none of them seems to challenge the process of test development. These results are discussed concerning implications for the assessment of basic competencies.

1. Erfassung allgemeiner schulischer Grundkompetenzen von Berufsschülerinnen und Berufsschülern

Allgemeine schulische Grundkompetenzen in den Bereichen Mathematik, Naturwissenschaften und Lesen stellen neben berufs- oder domänenspezifischen Kompetenzbereichen, welche Fähigkeiten und Fertigkeiten in Bezug auf eine erfolgreiche Bewältigung authentischer und beschäftigungsspezifischer Arbeitsaufgaben beschreiben, und berufsübergreifenden Kompetenzen (Kommunikations-, Selbstorganisations- und Teamfähigkeit, genauso wie Selbstwirksamkeitsüberzeugung, Verständnis von Organisationen, Betriebsabläufen und Arbeitsmärkten) einen dritten Bereich beruflicher Handlungskompetenz dar (vgl. Baethge, 2010). Im Zusammenhang mit der ULME-III-Studie beschreiben Lehmann, Seeber und Hunger (2007) allgemeine Grundkompetenzen (bzw. allgemeine Grundbildung) mit Bezug auf OECD and Statistics Canada (1995)

> „…als grundlegende Voraussetzungen für die Gestaltung der persönlichen Handlungsoptionen im Alltag sowie für die aktive Teilhabe am beruflichen und gesellschaftlichen Leben …" (S. 16)

und gruppieren als solche basale Fähigkeiten und Fertigkeiten im Umgang mit Textgattungen inklusiver diskontinuierlicher Texte (Tabellen, Grafiken etc.) und im Umgang mit Zahlen und Rechenaufgaben (vgl. Lehmann, Seeber & Hunger, 2007).

Allgemeine schulische Grundkompetenzen haben sich als erklärungsmächtige Prädiktoren für den Erfolg im dualen Ausbildungssystem erwiesen (z. B. Lehmann & Seeber, 2007; Nickolaus, Geissel, & Gschwendtner, 2008; Nickolaus et al., 2010; Nickolaus & Norwig, 2009; Seeber, 2007a, 2007b; Seeber & Lehmann, 2011). Beispielsweise haben sich bei der oben angesprochenen ULME-III-Studie (vgl. Seeber & Lehmann, 2007) in der Domäne Lesen metakognitive Strategien zur Texterschließung, bei denen Schülerinnen und Schüler Strategien zur Texterfassung mit Pseudo-Noten bewertet haben, als bedeutsamer Prädiktor für die berufliche Fachkompetenzentwicklung herausgestellt, etwa in kaufmännischen Berufen (Seeber, 2007A) oder bei medizinischen Fachangestellten (Seeber, 2007B). Die zu Beginn der Berufsausbildung erfasste mathematische Kompetenz (Stoffgebiete Algebra, Arithmetik und Geometrie; teilweise ergänzt um einen Fachleistungstest „Texte und Tabellen") konnte in einer Untersuchung bei Büro- und Werbekaufleuten und Kaufleuten im Einzelhandel im Vergleich zur allgemeinen Intelligenz, erfasst mit einem nonverbalen Test zum schlussfolgernden Denken, als besonders erklärungsmächtiger Prädiktor des jeweiligen berufsbezogenen Fachleistungstest identifiziert werden (Seeber, 2007A). Zusätzlich zu diesen Befunden in den Domänen Mathematik und Lesen liegen in der Domäne Naturwissenschaften inzwischen beispielsweise beim Ausbildungsberuf der Mechatroniker Belege für einen substanziellen Zusammenhang zwischen dem physikalischen und technischen Fachwissen am Beginn der Ausbildung und den beruflichen Fachkompetenzen in der Mitte und am Ende der Ausbildung vor (vgl. Maier et al., im Druck). Im Zusammenhang mit diesen und anderen Studien wurde jedoch ein Defizit bisheriger Untersuchungen deutlich: Wie beispielsweise von Redder et al. (2010) oder Balkenhol (2015) in Bezug auf die Domäne Lesen festgestellt, existieren keine besonders geeigneten Testinstrumente zur Erfassung allgemeiner schulischer Grundkompetenzen in beruflichen Anforderungskontexten. Die Entwicklung entsprechender Instrumente würde eine wichtige Forschungslücke füllen, ermöglichen sie es doch, strukturelle Zusammenhänge

zwischen allgemeinen Kompetenzen und den zu Beginn aufgeführten beiden berufsbezogenen Kompetenzbereichen systematisch zu untersuchen, sodass eine Einschätzung ihrer Bedeutung für die Entwicklung berufsspezifischer Kompetenz in verschiedenen Anwendungsgebieten der beruflichen Bildung möglich ist (Baethge, 2010, S. 32). Angesichts der hohen Heterogenität der Berufe und Ausbildungsziele stellen allgemeine schulische Grundkompetenzen zudem einen der wenigen Bereiche beruflicher Bildung dar, indem auch berufsübergreifend vergleichende Aussagen zur Bedeutung von Lernvoraussetzungen und Kompetenzentwicklungen möglich sind (vgl. hierzu auch Baethge, 2012, S. 128–131).

Die Entwicklung von Testinstrumenten zur Erfassung allgemeiner Kompetenzen im berufsbildenden Bereich ist allerdings von mindestens zwei Herausforderungen begleitet. Eine erste Herausforderung stellt die Abstimmung der Testinstrumente auf das jeweilige berufstypische Leistungsspektrum dar; diese ist in bisherigen Untersuchungen mit verfügbaren Instrumenten oftmals nur unzureichend gelungen. Beispielsweise sind beim Einsatz von Instrumenten aus dem Kontext von PISA, welche auf 15-jährige Schülerinnen und Schüler abzielt, und insbesondere dem *Programme for the International Assessment for Adult Competencies* (PIAAC; z. B. OECD, 2013), welches im mathematischen Bereich lediglich Rechenfertigkeiten berücksichtigt, Deckeneffekte für von Abiturientinnen und Abiturienten und leistungsstarken Realschülerinnen und Realschülern bevorzugt gewählten Berufen zu erwarten. Bei der Verwendung von Instrumenten der *Third International Mathematics and Science Study Advanced* (TIMSS Advanced; Mullis, Martin, Robitaille, & Foy, 2009), die auf die Diagnostik mathematischer Leistungen im gymnasialen Bereich ausgerichtet sind, könnten hingegen Bodeneffekte für Berufe mit leistungsschwächeren Jugendlichen auftreten. Die Instrumente der ULME-Studie (Untersuchung der Leistung, Motivation und Einstellungen zu Beginn der beruflichen Ausbildung; vgl. Lehmann et al., 2005) und der LAU-Studie (Aspekte der Lernausgangslage und der Lernentwicklung; Behörde für Schule und Berufsbildung, 2011, 2012) für die Jahrgangsstufen 9, 11 und 13 könnten auf das Leistungsspektrum der Jugendlichen bei Einmündung in unterschiedliche Berufs- bzw. Ausbildungsbereiche verhältnismäßig gut abgestimmt werden, allerdings weisen diese Instrumente den Nachteil langer Testzeiten auf. Kürzungen der Tests gehen mit einer Verringerung der Messgüte einher (vgl. die niedrige Reliabilität der gekürzten Version des ULME-I-Tests Mathematik I bei Bankkaufleuten in Rosendahl & Straka, 2011). Eine Möglichkeit, dieser Herausforderung relativ langer Testzeiten bei niedriger Differenzierungsfähigkeit an den Rändern der Kompetenzverteilung zu begegnen, stellt das computerisierte adaptive Testen (CAT) dar. Beim CAT orientiert sich die Auswahl der zur Bearbeitung vorgelegten Items am vorherigen Antwortverhalten des untersuchten Individuums. Hochleistungsfähigen Testpersonen, die zu Beginn viele Items richtig lösen, werden im Testverlauf zunehmend schwierigere Items präsentiert, weniger leistungsfähigen Probanden leichtere Items. Da die Schwierigkeit der zu bearbeitenden Items also an die individuelle Leistungsfähigkeit angepasst (adaptiert) wird, ermöglicht CAT zum einen eine hohe Differenzierungsfähigkeit über einen sehr breiten Leistungsbereich (z. B. Frey & Ehmke, 2007), zum anderen kann die Messeffizienz erheblich gesteigert werden (z. B. Frey, 2012), sodass im Vergleich zur herkömmlichen sequentiellen Darbietung von Testitems in fester Reihenfolge (wie sie etwa bei typischen Papier-und-Bleistift-Tests eingesetzt wird) bei gleicher Messpräzision deutlich weniger Items vorgelegt werden müssen. Da die Schülerinnen und Schüler im berufsbildenden

Bereich eine besonders heterogene Population darstellen, sind diese Möglichkeiten zur differenzierten Messung über einen breiten Leistungsbereich bei gleichzeitiger Verringerung der Testzeit hoch bedeutsam.

Eine zweite Herausforderung bezieht sich auf die berufsübergreifenden Einsatzmöglichkeiten der Testinstrumente. So sollte die Erfassung allgemeiner Grundqualifikationen nicht allein auf das Leistungsspektrum einzelner Ausbildungsgänge abgestimmt sein, sondern im Idealfall die Erfassung und den Vergleich der entsprechenden Kompetenzen berufsübergreifend gewährleisten. Eine Voraussetzung für eine breite Anwendbarkeit des Instrumentes im Bereich der beruflichen Bildung ist, dass die Testinstrumente keine Items beinhalten sollten, die einzelne Ausbildungsgänge systematisch bevorzugen; die Testinstrumente sollten also *fair* hinsichtlich der verschiedenen Ausbildungsgänge konstruiert sein.

Im Zusammenhang mit der Entwicklung eines solchen CAT für die Erfassung allgemeiner Kompetenzen bei Berufsschülern (Bernhardt et al., 2013) ist es daher die Zielsetzung des vorliegenden Beitrags, domänen- und itemspezifische Leistungsunterschiede zwischen Ausbildungsgängen im Rahmen des Forschungsprojekts *Messung allgemeiner Kompetenzen – adaptiv* (MaK-adapt) hinsichtlich der Fairness der erstellten Testinstrumente zu untersuchen. Hierzu kommen statistische Methoden des sogenannten differential item functioning (DIF; z. B. Camilli, & Shepard, 1994; Clauser, & Mazor, 1998; Osterling & Everson, 2009; Zumbo, 1999) zum Einsatz, welche um eine inhaltliche Prüfung der zuvor statistisch identifizierten Items ergänzt werden. Der Beitrag gliedert sich wie folgt: Im nachfolgenden zweiten Kapitel wird erläutert, unter welchen Umständen die statistische Identifikation von Leistungsunterschieden als Hinweis auf mangelnde Testfairness interpretiert werden sollte. Im dritten Kapitel wird das methodische Vorgehen zur Untersuchung domänen- und itemspezifischer Leistungsunterschiede bei der Entwicklung von drei adaptiven Tests zur Erfassung allgemeiner schulischer Grundkompetenzen im Rahmen des Forschungsprojekts MaK-adapt dargelegt und dabei zwei Arbeitsschritte zur Analyse dieser Effekte beschrieben: Ein erster quantitativer Arbeitsschritt zur Detektion von DIF und ein zweiter, eher qualitativ ausgerichteter Arbeitsschritt zur inhaltlichen Prüfung der zuvor statistisch identifizierten Items. Die Ergebnisse dieser zwei Arbeitsschritte werden im vierten Kapitel dargestellt. Hieraus ergeben sich sowohl konkrete Implikationen für die Testentwicklung im Projekt als auch eine allgemeine methodisch-konzeptionelle Empfehlung zum Umgang mit DIF-Items.

2. DIF-Effekte zwischen verschiedenen Ausbildungsberufen als Herausforderung für die Testfairness

Um unterschiedliche Testleistungen zwischen verschiedenen Gruppen von Berufsschülerinnen und Berufsschülern auf Kompetenzunterschiede in den untersuchten Bereichen (Lesen, Mathematik, Naturwissenschaften) zurückführen zu können, muss sichergestellt sein, dass die verwendeten Testitems einzelne Gruppen nicht systematisch bevorzugen. Unterschiede in der Lösungswahrscheinlichkeit eines Items zwischen zwei oder mehr Gruppen müssen natürlich nicht zwangsläufig systematische Bevorzugung anzeigen und damit auf mangelnde Testfairness hinweisen. Wird eine Mathematikaufgabe beispielsweise von 60 % der Gruppe angehender Bankkaufleute, aber nur von 40 % der Gruppe angehender KFZ-Mechaniker richtig gelöst, könnte dieser Unterschied in der Lösungswahrscheinlichkeit durchaus

Ausdruck der wahren Unterschiede in der durchschnittlichen mathematischen Kompetenz sein. Dies würde dann keineswegs auf mangelnde Testfairness hindeuten. Problematisch ist es vielmehr, wenn auch nach statistischer Kontrolle etwaiger allgemeiner Kompetenzunterschiede zwischen interessierenden Gruppen, die Lösungswahrscheinlichkeit einzelner Items noch stark unterschiedlich ausfällt und diese Unterschiede durch schwierigkeitsbestimmende Merkmale des Items erklärt werden können, die nicht direkt konstruktrelevant sind.

Zumbo (1999, S. 12) unterscheidet in diesem Zusammenhang die folgenden Begrifflichkeiten: Mit der Bezeichnung *Impact* (oder auch Test Impact) werden Unterschiede zwischen zwei oder mehr Gruppen in der mittleren Testleistung beschrieben. Als *Item Impact* wird der Unterschied in der Lösungswahrscheinlichkeit bei einem einzelnen Item bezeichnet, der durch Gruppenunterschiede in der mittleren Testleistung, den Test Impact, erklärt werden kann. *DIF* hingegen bezeichnet den Gruppenunterschied in der Lösungswahrscheinlichkeit bei einem einzelnen Item, welcher *nicht* auf den Test Impact zurückgeführt werden kann; jede DIF-Analyse wird also unter Berücksichtigung von Unterschieden in der mittleren Testleistung durchgeführt. Unter *Item-Bias* wird schließlich ein Gruppenunterschied in der Lösungswahrscheinlichkeit bei einem einzelnen Item verstanden, der nicht durch den Test Impact erklärt werden kann (also DIF beinhaltet) und zusätzlich auf spezifische, besondere Charakteristika des Test-Items oder der zugrunde liegenden Testsituation zurückzuführen ist. Entscheidend ist dabei, dass es sich um Charakteristika handelt, die sich nicht aus den Anforderungen des eigentlichen Testinhalts ergeben; man spricht daher auch von konstruktirrelevanter Varianz.

Die Definition konstrukt-(ir-)relevanter Testinhalte ist durchaus nicht trivial. Lässt sich beispielsweise ein DIF-Effekt zwischen kaufmännisch-verwaltenden und gewerblich-technischen Ausbildungsgängen bei einem Item der Domäne Mathematik mit technischer Anwendung finden, so muss anhand der zugrundeliegenden Konstruktdefinition entschieden werden, ob die technische Anwendung im Sinne konstruktirrelevanter Varianz zu interpretieren ist (weil das Item vielleicht eher der Kompetenzdefinition einer Domäne Technik denn jener der Mathematik zuzuordnen ist) oder – vor dem Hintergrund, dass technische Anwendungen eines der zentralen Einsatzgebiete der Mathematik darstellen – im Sinne konstruktrelevanter Varianz. Für mathematische Anwendungen im kaufmännischen Bereich gilt dies gleichermaßen.

Der oben aufgeführten Definition von Item-Bias ist zu entnehmen, dass bei vorliegendem DIF eine notwendige, aber keine hinreichende Voraussetzung für die Identifikation von Item-Bias erfüllt ist. Konsequenterweise stellt es eher kein angemessenes Vorgehen dar, Items bei der Entwicklung eines Testinstruments allein aufgrund statistischer Kennwerte auszusortieren ohne inhaltliche Kriterien zu prüfen. Um den zu Beginn beschriebenen zwei Herausforderungen bei der Erfassung allgemeiner Kompetenzen von Berufsschülern zu begegnen, soll im Zusammenhang mit der Entwicklung eines CAT die Problematik der Fairness gegenüber von zwei Hauptgruppen der kaufmännisch-verwaltend und gewerblich-technisch ausgerichteten Ausbildungsberufe – neun der zehn in Deutschland am stärksten besetzten Ausbildungsberufe (Statistisches Bundesamt, 2014) können einer dieser beiden Kategorien zugeordnet werden – untersucht werden. Erwähnt sei in diesem Zusammenhang, dass für eine besondere Notwendigkeit zur Untersuchung von DIF-Effekten beim CAT nach Zwick (2003) drei Gründe sprechen:

1. Die Testinstrumente sind kürzer als beim konventionellen Testen, DIF-Effekte bei einzelnen Items fallen daher bei der Schätzung der Schülerkompetenz stärker ins Gewicht als beim konventionellen Testen.
2. Die Sequenz vorgegebener Items hängt von jeder einzelnen Item-Antwort ab; die Vorgabe eines zu schwierigen oder zu leichten Items kann also die Vorgabe weiterer zu schwieriger oder zu leichter Items provozieren.
3. Die Administrierung eines Tests am Computer kann gegenüber der konventionellen Testvorgabe weitere mögliche DIF-Quellen (Computer-Bekanntheit, differenzielle Präferenzen für eine Computer-Administration etc.) beinhalten, bezüglich derer sich Personengruppen (z. B. auch Berufsgruppen, die unterschiedlich viel Zeit mit dem Computer arbeiten) unterscheiden können.

Konkret ist es daher die Zielsetzung des vorliegenden Beitrags zu analysieren, (1) ob domänen- und itemspezifische Kompetenzunterschiede zwischen Schülerinnen und Schülern kaufmännisch-verwaltend und gewerblich-technisch ausgerichteter Ausbildungsgänge bei der Erfassung allgemeiner schulischer Grundkompetenzen von Berufsschülern auftreten, und (2) ob diese Unterschiede wahre Leistungsdifferenzen abbilden oder aber auf schwierigkeitsbestimmende Item-Merkmale zurückgeführt werden müssen, die nicht als konstruktrelevant betrachtet werden können. Bei der Analyse dieser domänen- und itemspezifischen Unterschiede werden statistische Methoden zur Analyse von DIF genutzt und um eine qualitative Item-Analyse durch Inhaltsexpertinnen ergänzt.

3. Methode

Überblick über die Analysen: Es existieren zwei grundsätzliche Verfahrensweisen um wahre Domänen- und Itemeffekte zwischen Personengruppen von konstruktirrelevanter Varianz abzugrenzen und somit um Item-Bias zu identifizieren, statistische Methoden und die Einschätzung von einem oder mehreren Inhaltsexpertinnen und -experten (Zumbo, 1999). Beide Verfahrensweisen sollen in dieser Studie zum Einsatz kommen, wobei in einem ersten Arbeitsschritt problematische Items im Rahmen einer DIF-Analyse statistisch identifiziert werden, in einem sich anschließenden zweiten Arbeitsschritt ausgewählte DIF-Items exemplarisch mit Hilfe einer qualitativ ausgerichteten Analyse von Inhaltsexpertinnen im Hinblick auf schwierigkeitsbestimmende, konstruktirrelevante Itemmerkmale und somit im Hinblick auf Item-Bias analysiert werden.
Stichprobe: Die nachfolgend beschriebenen Analysen stützen sich auf die Kalibrierungsdaten des Forschungsprojekts MaK-adapt im Rahmen der vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Forschungsinitiative *Technology-based Assessment of Skills and Competencies in VET* (Technologieorientierte Kompetenzmessung in der beruflichen Bildung, *ASCOT;* Bundesministerium für Bildung und Forschung, 2014) und wurde als Kooperationsprojekt zwischen der Georg-August-Universität Göttingen, der Technischen Universität Darmstadt und der Friedrich-Schiller-Universität Jena konzipiert und umgesetzt. Die Stichprobe umfasste Testdaten von $N = 1224$ Berufsschülerinnen und Berufsschülern (33 % weiblich, 91 % Muttersprache Deutsch, 60 % Abschluss der mittleren Reife) aus den Bundesländern Hessen, Niedersachsen und Thüringen. Unter diesen befanden sich 3 % im ersten Ausbildungsjahr, 20 % im zweiten Ausbildungsjahr, 68 % im dritten Ausbildungsjahr

Tab. 1: Zuordnung von Ausbildungsberufen zur Gruppe der kaufmännisch-verwaltend beziehungsweise der technisch-gewerblich ausgerichteten Ausbildungsberufe

	technisch-gewerblich	kaufmännisch-verwaltend
Elektroniker/in	191	0
Elektroniker/in für Automatisierungstechnik	64	0
Goldschmied/in	15	0
Hauswirtschafter/in	0	18
Industriekauffrau/-mann	0	170
Informatiker/in	30	0
Kauffrau/mann	0	376
KFZ-Mechatroniker/in	100	0
Maler/in	4	0
Maurer/in	1	0
Mechaniker/in	167	0
Mechatroniker/in	70	0
Techniker/in	2	0
Technischer Assistent	14	0
Verkäufer/in	0	2
gesamt	658	566

und 9% im vierten Ausbildungsjahr. Da für verlässliche Parameterschätzungen bei DIF-Analysen relativ große Stichproben in den analysierten Gruppen benötigt werden, wurden für die hier dargestellten Analysen die Daten von zwei Gruppen von Ausbildungsberufen, kaufmännisch-verwaltend ausgerichtete Berufe und gewerblich-technisch ausgerichtete Berufe, analysiert. Tabelle 1 zeigt die Zuordnung von Ausbildungsberufen zu diesen beiden Gruppen. Die Ausbildungsberufe der restlichen 408 Probanden konnten keiner der beiden Gruppen zugeordnet werden, sodass diese Probanden von der vorliegenden Analyse ausgeschlossen wurden.

Instrumente: Das Projekt MaK-adapt zielt auf die Entwicklung und Anwendung von drei computerbasierten Messinstrumenten zur Erfassung schulisch erworbener Basiskompetenzen ab, welche für berufliches Lernen und eine aktive Gestaltung der Berufsbiografie und des Lebensalltags relevant sind. Konkret wurden im Projekt die Kompetenzbereiche Lesen, Mathematik und Naturwissenschaften erfasst (Bernhardt et al, 2013): Bei der Erfassung von Lesen wurde weitestgehend der Definition von *Reading Literacy* bei PISA (z. B. OECD, 2009) gefolgt und Lesekompetenz als die Fähigkeit zum Verstehen und Nutzen multipler Darstellungen in schriftlichen Dokumenten angesehen. Diese schriftlichen Dokumente können dann externe Repräsentationen wie etwa Texte, Bilder, Diagramme, Tabellen oder anderes enthalten. In den konstruierten Items wurden drei Klassen von Textformaten berücksichtigt: deskriptionale kontinuierliche Texte, depiktionale Texte (Bilder, Grafen, Tabellen) und Mischformate der zwei zuvor genannten Formate. Da berufliches Lesen ganz überwiegend Lesen mit Handlungsabsicht ist (Ziegler, Balkenhol, Keimes & Rexing 2012), wurden alle Items mit einem Hinweis

zum konkreten Leseziel versehen. Zudem wurden bei der Item-Konstruktion die kognitiven Anforderungsbereiche (1) Identifizieren, (2) Integrieren und (3) Generieren abgegrenzt (vgl. Ziegler, Balkenhol, Keimes & Rexing 2012; Balkenhol 2015).

Der Test zur Erfassung mathematischer Kompetenz orientiert sich stark an der theoretischen Rahmenkonzeption von PISA 2009 (vgl. Frey, Heinze, Mildner, Hochweber & Asseburg, 2010; oecd, 2009) und entspricht somit dem dort angewandten Konzept der *Mathematical Literacy*. Bei der Zusammenstellung des Mathematiktests wurden daher bewusst unterschiedliche Kontexte (private Situationen, bildungsbezogene und berufliche Situationen, gesellschaftliche Situationen, wissenschaftliche Situationen) berücksichtigt. Die theoretische Rahmenkonzeption unterscheidet vier mathematische Inhaltsbereiche (Quantität, Veränderung und Beziehungen, Raum und Form, Unsicherheit) und drei Kompetenzcluster (Reproduktion, Verbindungen, Reflexion), welche das kognitive Anspruchsniveau eines Items abbilden (vgl. Frey et al., 2010).

Zur Erfassung der naturwissenschaftlichen Kompetenz wurden die theoretischen Rahmenkonzeptionen von TIMSS (Garden & Orpwood, 1996), PISA und der HarmoS-Studie (Konsortium HarmoS Naturwissenschaften+, 2009) herangezogen. Durch Adaption an die Dimensionen der letztgenannten HarmoS-Studie konnten Items zu vier Anwendungsgebieten konstruiert werden: (1) Leben und Gesundheit, (2) Erde, Planeten, Umwelt und natürliche Ressourcen, (3) Stoffe und Stoffveränderungen und (4) Bewegung, Kraft und Energie. Die vier Anwendungsgebiete orientieren sich an naturwissenschaftlichen Fächern (Biologie, Erdkunde, Chemie und Physik), entsprechen diesen aber nicht gänzlich, sondern weisen oft Überschneidungen auf. Die kognitiven Verarbeitungsvorgänge innerhalb der Themengebiete können in Anlehnung an die TIMSS Studien zudem drei Kategorien zugeordnet werden: (1) Verstehen einfacher und komplexer Informationen, (2) Konzeptualisieren und Analysieren sowie (3) Nutzen naturwissenschaftlicher/s Evidenzen und Wissen für Entscheidungen und Handlungen in komplexen Lebenssituationen.

Der Itempool in den drei Domänen konnte wesentlich aus freigegebenen Items der groß angelegten Vergleichsstudien (PISA; TIMSS) oder dem *International Adult Literacy Survey* (IALS; Statistics Canada & OECD, 2005), den Erhebungen zu den nationalen Bildungsstandards (Blum, Drüke-Noe, Hartung, & Köller, 2006), den Vergleichsarbeiten in der 8. Jahrgangsstufe (VERA; Institut zur Qualitätsentwicklung im Bildungswesen, n. d.) und dem Projekt ULME (Lehmann et al., 2005) für den beruflichen Kontext angepasst und computerisiert werden. Im Bereich Lesen wurden eine ganze Reihe von Items neu entwickelt, um die Testanforderungen stärker auf das Lesen in beruflichen Kontexten auszurichten. Neben einer inhaltlichen Prüfung durch die Testentwickler, wurden bei der Auswahl der Testitems aus der Gesamtmenge der verfügbaren Items übliche Kriterien für die Kalibrierung eines Itempools für das CAT berücksichtigt (vgl. z. B. Wise & Kingsbury, 2000; Thompson & Weiss, 2011):

– Hohe Anzahl an Items in allen Schwierigkeitsbereichen, um Probanden sicher diagnostisch verorten zu können.
– Möglichkeit zur automatischen Auswertung (*Scoring)* der eingesetzten Items als korrekte oder falsche Antwort.
– Abbilden der intendierten dimensionalen Struktur (hier also die Dimensionen Lesen, Mathematik und Naturwissenschaften) mit Hilfe des verwendeten Itempools.

– Passung der Items zum Testmodell der *Item-Response-Theory* (z. B. De Ayala, 2009; Embretson & Reise, 2000), auf der CAT-Verfahren basieren (vgl. auch den nachfolgenden Absatz).

Der unter Berücksichtigung dieser Kriterien zusammengestellte Itempool umfasste 73 Items zum Lesen und jeweils 133 Items zur Mathematik und zu den Naturwissenschaften. Diese Items wurden bei einer Kalibrierungsstudie der oben beschriebenen Stichprobe von $N = 1632$ Berufsschülerinnen und Berufsschülern zur Bearbeitung vorgelegt. Die vorrangigen Ziele der Kalibrierungsstudie bestanden in der Identifikation defizitärer Items und in der Schätzung von Item-Schwierigkeiten unter Nutzung eines Modells der Item-Response-Theorie. Die geschätzten Item-Schwierigkeiten werden bei den zu entwickelnden computerisierten adaptiven Tests benötigt.

Bei der Kalibrierungsstudie wurde jeder Testperson ein zufällig ausgewähltes Testheft mit jeweils 33 Items aus dem Itempool zur Bearbeitung vorgelegt. Die Verteilung der Items auf die Testhefte erfolgte auf Basis eines unvollständigen balancierten Testheftdesigns (vgl. Frey, Hartig & Rupp, 2009) mit zwei Ebenen, bei dem die Kompetenzbereiche Lesen, Mathematik und Naturwissenschaften vollständig permutiert worden sind und die Items der drei Kompetenzbereiche gleichmäßig auf alle Positionen im Testheft verteilt wurden. Insgesamt ergaben sich durch das Design 798 verschiedene Testhefte wobei jeder Schüler ein Testheft mit 12 Mathematikitems, 12 Naturwissenschaftsitems und 9 Leseitems erhielt. In den hier beschriebenen Analysen werden 71 Leseitems, 125 Mathematikitems und 130 Naturwissenschaftsitems analysiert, deren Passung sich mit den Annahmen des Rasch-Modells (Itemfit) bei der Skalenanalyse als hinreichend erwiesen hat.

Statistische Analysen zur Detektion von DIF: Die Daten wurden zunächst mit Hilfe von statistischer Standard-Software aufbereitet bevor sie mit der Software ConQuest 3.0.1 (Adams, Wu, & Wilson, 2012) mit dem unidimensionalen Rasch-Modell skaliert wurden. Der hier vorgestellten Untersuchung waren bereits DIF-Analysen in Bezug auf das Geschlecht vorausgegangen, Items mit systematischem DIF zwischen den Geschlechtern waren ebenfalls aussortiert worden. Zudem waren auch Mittelwertsdifferenzen bei den drei Kompetenzskalen zwischen Schülerinnen und Schülern mit deutscher und mit ausländischer Muttersprache deskriptiv untersucht worden, wenngleich wir – insbesondere aufgrund der stark unterschiedlichen Stichprobengrößen – auf systematische DIF-Analysen in Bezug auf dieses Merkmal verzichtet haben. Zur Analyse von DIF zwischen den zuvor beschriebenen Gruppen von Ausbildungsberufen wurden ebenfalls mit ConQuest 3.0.1 Mehrfacetten-Rasch-Modelle der Form

$$\ln\left[\frac{P\left(X_{ijk}=1\right)}{P\left(X_{ijk}=0\right)}\right]=\theta_j-\delta_i-\lambda_k+\delta_i\lambda_k-\zeta_j$$

geschätzt, wobei X_{ijk} die dichtotom (0 = falsche Lösung; 1 = richtige Lösung) bewertete Item-Antwort von Testperson *n* aus Gruppe *k* auf Item *i*, θ_j die Kompetenz von Testperson *j*, δ_i die Schwierigkeit von Item *i*, λ_k die mittlere Fähigkeit der *k*-ten Gruppe von Ausbildungsberufen beschreibt. Damit die geschätzten Effekte nicht durch bei Testungen übliche Positionseffekte verzerrt werden, wurde zusätzlich ζ_j als Kontrollvariable in das Modell aufgenommen. Dieser Parameter repräsentiert

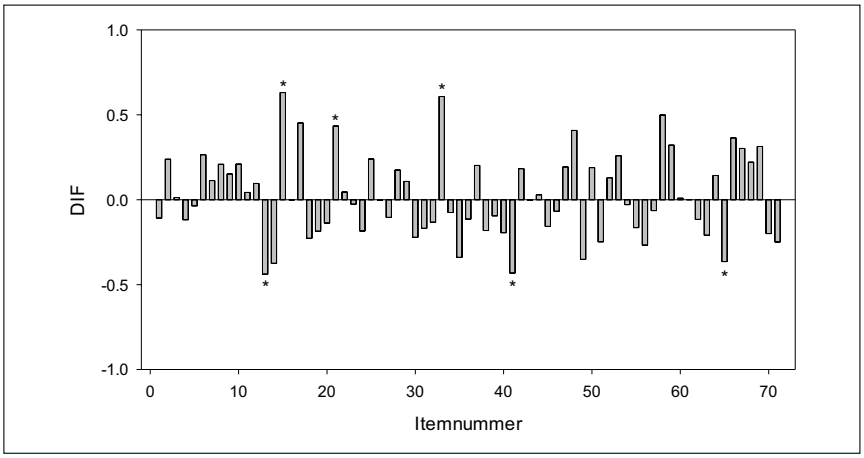
die Position (erste, zweite oder dritte Stelle) der entsprechenden Domäne in dem von Testperson *j* bearbeiteten Testheft.

Wichtig im Hinblick auf DIF ist der Interaktionseffekt $\delta_i\lambda_k$. Dieser drückt aus, inwieweit die Wahrscheinlichkeit ein Item korrekt zu beantworten nach Berücksichtigung von mittleren Kompetenzunterschieden zwischen den Ausbildungsberufsgruppen und der Position im Testheft für die untersuchten Ausbildungsberufe unterschiedlich ausfällt. Weicht dieser Parameter signifikant von 0 ab, ist dies als statistischer Hinweis auf DIF zu werten. Bei der statistischen Analyse von DIF geht man üblicherweise so vor, dass anhand des mit ConQuest berechneten Omnibustests auf Basis der χ^2-Verteilung bestimmt wird, ob die Modellfacette $\delta_i\lambda_k$ insgesamt einen von 0 abweichenden Effekt aufweist. Ist dies der Fall, kann anhand des Item-spezifischen Konfidenzintervalls (z. B. basierend auf dem doppelten Standardfehler von $\delta_i\lambda_k$) für jedes einzelne Item getestet werden, ob dessen Interaktionsterm signifikant von 0 abweicht. Zu beachten ist, dass auf Basis des Rasch-Modells lediglich sogenannte uniforme DIF-Effekte möglich sind, bei denen der Vorteil einer Gruppe über das gesamte Leistungsspektrum gleichermaßen hoch ausgeprägt ist (vgl. hierzu etwa Osterlind & Everson, 2009).

Inhaltsanalyse im Hinblick auf schwierigkeitsbestimmende, konstruktirrelevante Itemmerkmale: Im zweiten Schritt nahmen zwei Inhaltsexpertinnen (Universitätsprofessorinnen der Berufs- bzw. Wirtschaftspädagogik) eine Analyse der zuvor statistisch identifizierten Items im Hinblick auf die Frage vor, ob konstruktirrelevante aber schwierigkeitsbestimmende Itemmerkmale den DIF-Effekt erklären könnten. Dazu wurden sie gebeten, jedes der zuvor statistisch identifizierten Items hinsichtlich der folgenden drei Fragen zu begutachten:

1. Ist über mittlere Unterschiede zwischen den Ausbildungsgruppen im jeweiligen Kompetenzbereich hinaus bei dem spezifischen Item ein signifikanter Vorteil für eine Gruppe von Ausbildungsberufen zu erwarten und – wenn ja – in welche Richtung? Bei dieser Frage lagen drei Antwortoptionen vor (Vorteil von Schülern kaufmännisch-verwaltender Ausbildungsberufe, Vorteil von Schülern gewerblich-technischer Ausbildungsberufe und die Option, kein Vorteil einer der beiden Gruppen von Schülern).
2. Wie sicher (auf einer vierstufige Antwortskala mit den Antwortkategorien „sehr unsicher“, „eher unsicher“, „eher sicher“, „sehr sicher“) sind Sie bezüglich dieser Einschätzung?
3. Welche schwierigkeitsbestimmenden Merkmale könnten den Vorteil hervorrufen (offenes Antwortformat)?

Die Antworten bezüglich der ersten beiden Fragen wurden dann mit der tatsächlichen Richtung des DIF-Effektes verglichen. Auf Basis einer korrekten und sicheren Einschätzung der Richtung des DIF-Effektes und anhand der Antwort auf die dritte Frage wurde schließlich entschieden, ob die DIF-Effekte als Item-Bias zu verstehen sind oder nicht.



Anmerkungen: Positive Werte indizieren einen Vorteil der Gruppe kaufmännisch-verwaltend, negative Werte indizieren einen Vorteil der Gruppe technisch-gewerblich ausgerichteter Ausbildungsberufe. Signifikante DIF-Effekte sind durch das Symbol * gekennzeichnet.

Abb. 1: Differential item functioning zwischen technisch-gewerblich und kaufmännisch-verwaltend ausgerichteten Ausbildungsberufen im Kompetenzbereich Lesen (dargestellt als Abweichung von der durchschnittlichen Item-Schwierigkeit)

4. Ergebnisse

Differenzielle Domänen- und Item-Effekte zwischen der Gruppe kaufmännisch-verwaltender und der Gruppe gewerblich-technischer Ausbildungsberufe werden im Folgenden für die drei Kompetenzbereiche Lesen, Mathematik und Naturwissenschaften dargestellt.

DIF-Effekte in der Domäne Lesen: Der Kompetenzunterschied zwischen der Gruppe der kaufmännisch-verwaltend und der Gruppe der gewerblich-technisch ausgerichteten Ausbildungsberufe lag bei einer Standardabweichung von 0.72 in der Domäne Lesen bei 0.166 (λ_k= 0.083, *SE* (λ_k) = 0.024), zugunsten der Gruppe der kaufmännisch-verwaltenden Ausbildungsberufe. Der Omnibustest für die DIF-Effekte, die Interaktionen zwischen Item-Schwierigkeit und Gruppe der Ausbildungsberufe, fiel mit $\chi^2(71) = 95.10$ ($p < .05$) signifikant aus. Differenzen zwischen den gruppenspezifischen und den durchschnittlichen Schwierigkeitsparametern im Kompetenzbereich Lesen sind in Abbildung 1 dargestellt. Es ist ersichtlich, dass die Größe der DIF-Effekte in beide Richtungen deutlich variierte. Insgesamt wiesen sechs Items (Nr. 13, 15, 21, 33, 41, 65) signifikanten DIF auf, drei davon sind im Sinne eines Vorteils der Gruppe gewerblich-technisch ausgerichteter Ausbildungsberufe, drei im Sinne eines Vorteils der Gruppe kaufmännisch-verwaltend ausgerichteter Ausbildungsberufe zu interpretieren.

Ergebnisse der Item-Inhaltsanalyse in der Domäne Lesen: Die statistisch identifizierten Items in der Domäne Lesen wurden anschließend in einer Inhaltsanalyse weiter untersucht. Tabelle 2 zeigt die Einschätzung zum Vorteil gewerblich-technisch und kaufmännisch-verwaltend ausgerichteter Ausbildungsberufe bei dieser Item-

auswahl durch eine Inhaltsexpertin im Vergleich zur tatsächlichen Effektrichtung. Es ist ersichtlich, dass keine hohe Sicherheit über die Einschätzung hinsichtlich eines Vorteils einer der beiden Gruppen von Ausbildungsberufen bestand, wenn auch die Zuordnung bei vier von sechs Items richtig ausfiel. Definiert man als Voraussetzungen für Item-Bias hier konkret, dass (1) die Zuordnung des Vorteils einer der beiden Gruppen richtig ausfallen und (2) diese Einschätzung auch mindestens als „eher sicher" beschrieben worden sein muss, so liegen lediglich zwei problematische Items vor, jeweils eines mit einem Vorteil für die Gruppe der kaufmännisch-verwaltend beziehungsweise für die gewerblich-technisch ausgerichteten Ausbildungsberufe. Abbildung 2 zeigt diese beiden Items. Zur möglichen Begründung bezüglich des Vorteils der Gruppe der kaufmännisch-verwaltenden Ausbildungsberufe bei dem Item „Kinobesuch" merkte die Inhaltsexpertin an, dass Terminfindung und Terminvereinbarung Anforderungen darstellen, die im kaufmännisch-verwaltenden Bereich häufiger erledigt werden müssen als im gewerblich-technischen Bereich. Zur möglichen Begründung bezüglich des Vorteils der Gruppe der gewerblich-technischen Ausbildungsberufe bei der Aufgabe „Waschmaschine" wurde angemerkt, dass Auszubildende im gewerblich-technischen Bereich mit der Interpretation der Symbolik bei entsprechenden Abbildungen eher vertraut sind und mit höherer Wahrscheinlichkeit den notwendigen handwerklichen Sachverstand mitbringen. *DIF-Effekte in der Domäne Mathematik:* In der Domäne Mathematik lag der Kompetenzunterschied zwischen der Gruppe der kaufmännisch-verwaltend und der Gruppe der gewerblich-technisch ausgerichteten Ausbildungsberufe bei einer Standardabweichung von 0.79 bei 0.178 ($\lambda_k = 0.089$, $SE(\lambda_k) = 0.022$). Der Omnibustest für die DIF-Effekte fiel mit χ^2 (125) = 153.51, $p < .05$ signifikant aus. Differenzen zwischen den gruppenspezifischen und den durchschnittlichen Schwierigkeitspara-

Tab. 2: Einschätzung zum Vorteil technisch-gewerblich und kaufmännisch-verwaltend ausgerichteter Ausbildungsberufe bei statistisch detektierten Items in der Domäne Lesen

		Vorteil gemäß statistischer Analyse	
Vorteil gemäß Expertenurteil	Urteilssicherheit	Technisch-gewerblich	Kaufmännisch-verwaltend
Kein	Sehr unsicher		
	Eher unsicher		
	Eher sicher	1	
	Sehr sicher		
Technisch-gewerblich	Sehr unsicher		
	Eher unsicher		
	Eher sicher	1	
	Sehr sicher		
Kaufmännisch-verwaltend	Sehr unsicher		
	Eher unsicher	1	2
	Eher sicher		1
	Sehr sicher		

Kinobesuch L0349221

Ingo ist 15 Jahre alt. Er will während der einwöchigen Schulferien einen Kinobesuch mit zwei gleichaltrigen Freunden organisieren. Die Ferien beginnen am Samstag, den 24. März, und enden am Sonntag, den 1. April.

Nun geht es darum, eine geeignete Uhrzeit und ein geeignetes Datum für den Kinobesuch zu finden.

Ingo fragt seine Freunde, welche Tage und Uhrzeiten ihnen für diesen Besuch passen. Er hat folgende Informationen bekommen:

Frank: „Ich muss Montag und Mittwoch nachmittags von 14:30 Uhr bis 15:30 Uhr wegen meines Musikunterrichts zu Hause bleiben."

Simon: „Ich muss sonntags meine Großmutter besuchen, also ist der Sonntag ausgeschlossen. Ich habe Pokamin schon gesehen und will ihn nicht noch einmal sehen."

Ingos Eltern bestehen darauf, dass er nur einen Film sieht, der für Jugendliche seines Alters nicht verboten ist und dass er nicht zu Fuß nach Hause geht. Sie können die Jungen zu jeder Uhrzeit bis 22 Uhr abends wieder abholen.

Ingo erkundigt sich nach dem Kinoprogramm für die Ferienwoche. Er findet folgende Übersicht und entscheidet sich für den Film „Kinder im Netz".

Welcher Termin eignet sich für den Besuch am besten?

KINO TIVOLI Reservierungen unter der Nummer 01923/4230069 Infos 24h: 01924/42007100 Dienstag ist Kinotag: Alle Filme 3 Franken **Zweiwöchiges Programm ab Freitag, den 23. März**	
Kinder im Netz 113 Min Frei ab 12 Jahren. 14:00 Uhr (nur Mo-Fr) 21:35 (nur Sa/So)	**Pokamin** 105 Min. 13:40 Uhr (täglich) 16:35 Uhr (täglich)
Die Monster der Tiefe 164 Min. Frei ab 18 Jahren. 19:55 Uhr (nur Fr/Sa)	**Enigma** 144 Min. Frei ab 12 Jahren. 15:00 Uhr (nur Mo-Fr) 18:00 Uhr (nur Sa/So)
Der Menschenfresser 148 Min. Frei ab 18 Jahren. 18:30 Uhr (täglich)	**König der Wildnis** 117 Min. Ohne Altersbegrenzung 14:35 Uhr (nur Mo-Fr) 18:50 Uhr (nur Sa/So)

Welcher der folgenden Termine würde passen?

Bitte kreuzen Sie an.

[ ] Montag, 26. März

[ ] Freitag, 30. März

[ ] Mittwoch, 28. März

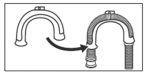
[ ] Samstag, 31. März

Waschmaschine L0679321

Es gibt mehrere Varianten, den Ablaufschlauch einer Waschmaschine anzuschließen. Zum einen der Anschluss **mit** Kunststoff-Schlauchführung, zum anderen **ohne** Kunststoff- Schlauchführung.

Welche Abbildungen (Abbildung 1 bis 4) zeigen einen Anschluss mit Kunststoff-Schlauchführung?

Abbildung 1



Kunststoff-Schlauchführung

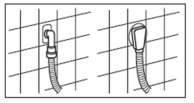


Abbildung 2

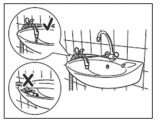


Abbildung 3

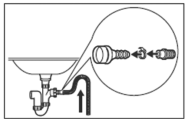
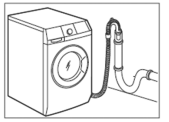


Abbildung 4



Welche Abbildungen (Abbildung 1 bis 4) zeigen einen Anschluss mit Kunststoff-Schlauchführung?
Bitte kreuzen Sie an.

[ ] Abbildung 2 und Abbildung 3

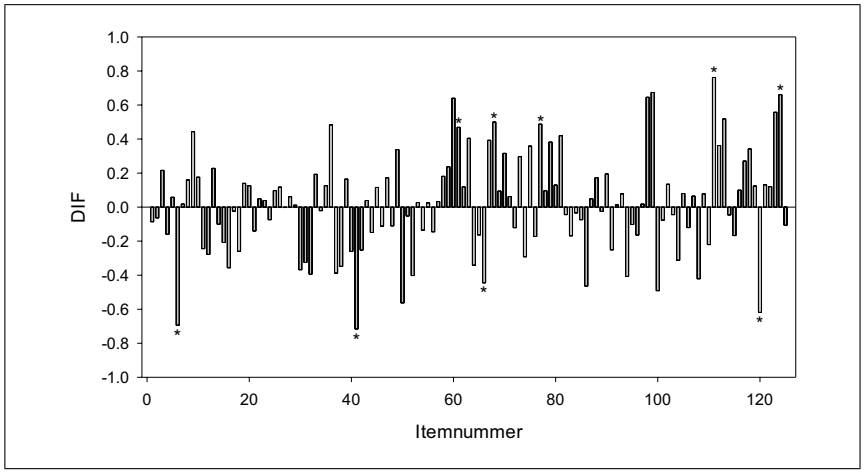
[ ] Abbildung 1 und Abbildung 4

[ ] Abbildung 3 und Abbildung 4

[ ] Abbildung 2 und Abbildung 4

Anmerkungen: Die Darstellung der Items wurde für die Abbildung hier angepasst

Abb. 2: Items mit möglichem Item Bias zum Vorteil technisch-gewerblich beziehungsweise kaufmännisch-verwaltend ausgerichteter Ausbildungsberufe im Kompetenzbereich Lesen. Bei dem linken Item ist die zweite Antwortalternative korrekt, beim rechten Item ist die vierte Antwortalternative korrekt.



Anmerkungen: Positive Werte indizieren einen Vorteil der Gruppe kaufmännisch-verwaltend, negative Werte indizieren einen Vorteil der Gruppe technisch-gewerblich ausgerichteter Ausbildungsberufe. Signifikante DIF-Effekte sind durch das Symbol * gekennzeichnet.

Abb. 3: Differential item functioning zwischen technisch-gewerblich und kaufmännisch-verwaltend ausgerichteten Ausbildungsberufen im Kompetenzbereich Mathematik (dargestellt als Abweichung von der durchschnittlichen Item-Schwierigkeit)

metern im Kompetenzbereich Mathematik sind in Abbildung 3 dargestellt. Insgesamt wiesen neun Items (Nr. 6, 41, 61, 66, 68, 77, 111, 120, 124) signifikanten DIF auf, vier davon sind im Sinne eines Vorteils der Gruppe gewerblich-technisch ausgerichteter Ausbildungsberufe, fünf im Sinne eines Vorteils der Gruppe kaufmännisch-verwaltend ausgerichteter Ausbildungsberufe zu interpretieren.

Ergebnisse der Item-Inhaltsanalyse in der Domäne Mathematik: Im Anschluss wurden die identifizierten Items in der Domäne Mathematik ebenfalls in einer Inhaltsanalyse weiter untersucht. Tabelle 3 zeigt die Einschätzung zum Vorteil gewerblich-technisch und kaufmännisch-verwaltend ausgerichteter Ausbildungsberufe bei dieser Itemauswahl durch eine Inhaltsexpertin im Vergleich zur tatsächlichen Effektrichtung. Die Einschätzung hinsichtlich eines Vorteils einer der beiden Gruppen von Ausbildungsberufen wird von der Inhaltsexpertin mit einer unterschiedlichen selbsteingeschätzten Sicherheit getroffen; die Zuordnung fiel bei drei von neun Items richtig aus. Definiert man Item-Bias auch hier wiederum durch die richtige Zuordnung des Vorteils einer der beiden Gruppen und eine mindestens als „eher sicher“ beschriebene Einschätzung, so liegt genau ein problematisches Items mit einem Vorteil für die Gruppe der kaufmännisch-verwaltend ausgerichteten Ausbildungsberufe vor. Abbildung 4 zeigt dieses Item. Zur möglichen Begründung bezüglich des Vorteils der Gruppe der kaufmännisch-verwaltenden Ausbildungsberufe bei diesem Item merkte die Inhaltsexpertin an, dass Prozentrechnung zu den Routinerechnungen im kaufmännischen Bereich gehört und aufgrund dieser Vertrautheit mit der Prozentrechnung ein Vorteil für die Kaufleute erwartet werden kann.

Tab. 3: Einschätzung zum Vorteil technisch-gewerblich und kaufmännisch-verwaltend ausgerichteter Ausbildungsberufe bei statistisch detektierten Items in der Domäne Mathematik

		Vorteil gemäß statistischer Analyse	
Vorteil gemäß Expertenurteil	Urteilssicherheit	Technisch-gewerblich	Kaufmännisch-verwaltend
Kein	Sehr unsicher		
	Eher unsicher		
	Eher sicher	1	1
	Sehr sicher		
Technisch-gewerblich	Sehr unsicher		
	Eher unsicher		1
	Eher sicher		
	Sehr sicher		
Kaufmännisch-verwaltend	Sehr unsicher		
	Eher unsicher	2	2
	Eher sicher	1	
	Sehr sicher		1

DIF-Effekte in der Domäne Naturwissenschaften: Der Kompetenzunterschied zwischen der Gruppe der kaufmännisch-verwaltend und der Gruppe der gewerblich-technisch ausgerichteten Ausbildungsberufe lag bei einer Standardabweichung von 0.59 in der Domäne Naturwissenschaften bei 0.284 ($\lambda_k = 0.142$, SE (λ_k) = 0.022) wiederum zugunsten der Gruppe der gewerblich-technischen Ausbildungsberufe. Der Omnibustest für DIF-Effekte wurde mit χ^2 (130) = 142.23, p = .219 nicht signifikant, sodass auf die Interpretation des DIF-Effektes jedes einzelnen Items sowie auf eine weitergehenden inhaltliche Analyse hinsichtlich von Item-Bias verzichtet wurde.
Zusammenfassung der Ergebnisse im Hinblick auf die Forschungsziele: Zusammengefasst lässt sich in Bezug auf die formulierten Zielsetzungen feststellen, dass in allen drei Kompetenzbereichen domänen- und item-spezifische Kompetenzunterschiede zwischen kaufmännisch-verwaltenden und gewerblich-technischen Ausbildungsgängen bei der Erfassung allgemeiner Kompetenzen von Berufsschülerinnen und Berufsschülern vorliegen, jedoch nur in der Domäne Lesen (fünf Items) und der Domäne Mathematik (neun Items) DIF-Effekte statistisch identifiziert wurden. Bei zwei Items der Domäne Lesen beurteilt die befragte Inhaltsexpertin die Effektrichtung korrekt und ist sich dabei mindestens „eher sicher". Dabei stellen die post-hoc formulierten Erklärungen für den jeweiligen DIF-Effekt zumindest kein starkes Indiz für konstruktirrelevante Varianz dar. In der Domäne Mathematik wird lediglich bei einem Item der DIF-Effekt durch eine Inhaltsexpertin mindestens „eher sicher" und korrekt hinsichtlich der Effektrichtung eingeschätzt, aus der post-hoc formulierten Erklärung leitet sich nicht unbedingt ein Hinweis auf konstruktirrelevante Varianz ab. Die Mehrzahl der statistisch identifizierten DIF-Effekte ist im Sinne konstruktrelevanter Leistungsdifferenzen zu interpretieren.

Buchführung

Bei untenstehender Quittung ist der Mehrwertsteuersatz durch einen Fleck unlesbar geworden. Der Gesamtbetrag von 19 Euro setzt sich zusammen aus dem Nettobetrag und der Mehrwertsteuer für diesen Nettobetrag.

QUITTUNG

Günther, Ulrich
PHP 5. Ein praktischer Einstieg.
3-89721-278-1 19,00

Total: 1 19,00 EUR

ec-Cash: 19,00 EUR
Zurück: 0,00 EUR

Betrag enthält 1,24 EUR MWSt:
[redacted]% = 1,24 Netto: 17,76
Bankleitzahl: 321/5800/0100
USt-Idnr.: DE 812277765
25.08.2004 11:57:52 20-2-93

Welcher Mehrwertsteuersatz ist hier verwendet worden?
Kreuzen Sie die richtige Lösung an.

[ ] 6,5 %

[ ] 7 %

[ ] 7,5 %

[ ] 16 %

[ ] 19 %

Anmerkungen: Die Darstellung des Items wurde für die Abbildung hier angepasst.

Abb. 4: Item mit möglichem Item-Bias zum Vorteil technisch-gewerblich beziehungsweise kaufmännisch-verwaltend ausgerichteter Ausbildungsberufe im Kompetenzbereich Mathematik. Die zweite Antwortalternative ist korrekt.

5. Diskussion

Es liegen eine Reihe empirischer Belege vor, welche die Bedeutung allgemeiner schulischer Grundqualifikationen wie Lesen, Mathematik und Naturwissenschaften für den Erfolg im Rahmen der beruflichen Ausbildung unterstreichen (z. B. Lehmann & Seeber, 2007; Nickolaus, Geissel, & Gschwendtner, 2008; Nickolaus et al., 2010; Nickolaus & Norwig, 2009; Seeber, 2007a, 2007b; Seeber & Lehmann, 2011). Im Rahmen der BMBF-Initiative ASCOT wurde der Zielsetzung nachgegangen, drei unidimensionale adaptive Tests zu konstruieren, mit deren Hilfe Berufsschülerinnen und Berufsschüler hinsichtlich dieser Grundqualifikationen verortet und strukturelle Zusammenhänge zwischen allgemeinen und berufsfachlichen Kompetenzdimensionen systematisch untersucht werden können. In der vorliegenden Untersuchung wurden im Rahmen einer DIF-Studie in Bezug auf den Ausbildungsberuf Items aus der Kalibrierungsstudie im Rahmen der Entwicklung der drei adaptiven Tests auf Grund statistischer Kennwerte identifiziert und – wo dies notwendig erschien, um Item-Bias zu prüfen – anschließend einer inhaltlichen Expertenbeurteilung unterzogen. Die Ergebnisse verdeutlichen, dass zwischen den beiden untersuchten Grup-

pen von Ausbildungsberufen (kaufmännisch-verwaltende vs. gewerblich-technische Berufe) substantielle Unterschiede mit Vorteilen von Schülerinnen und Schülern der kaufmännisch-verwaltenden Berufe im Lesen und Vorteilen von Schülerinnen und Schülern der gewerblich-technischen Ausbildungsberufe hinsichtlich der durchschnittlichen Kompetenz in den Bereichen Mathematik und Naturwissenschaften bestanden. Statistisch bedeutsame und inhaltlich klar auf konstruktirrelevante Merkmale zurückführbare Effekte im Sinne eines Item-Bias konnten dabei nicht identifiziert werden. Dies ist aus der Perspektive der Testkonstruktion ein erfreuliches Resultat, da somit schulische Grundqualifikationen über verschiedene Ausbildungsberufe hinweg mit dem gleichen Instrument fair erfasst und auch ihre Leistungen zueinander in Beziehung gesetzt werden können. Dementsprechend konnte das entsprechende Testinstrument (Bernhardt et al., 2013) inzwischen – der ursprünglichen Entwicklungsidee folgend – in bereits sechs Projekten der ASCOT-Initiative bei Schülerinnen und Schülern unterschiedlicher Ausbildungsgänge erfolgreich eingesetzt werden.

Die konkreten Empfehlungen für die Praxis, welche sich in diesen Projekten ableiten lassen, müssen abgewartet werden. Immerhin liegt im Sinne von Baethge (2012, S. 127) mit der Entwicklung eines validen, reliablen und objektiv auswertbaren Instrumentes zur adaptiven Messung allgemeiner schulischer Grundkompetenzen in verschiedene Ausbildungsgängen bereits eine wichtige Voraussetzung vor, um wissenschaftlich belastbare und beispielsweise auch für die Beratung von Politik und praktisch tätigen Institutionen nutzbare Aussagen tätigen zu können. Den Studien im Rahmen der ASCOT-Initiative können möglicherweise auch Aussagen dazu abgeleitet werden, ob der flächendeckende Einsatz eines qualitativ hochwertigen (adaptiven) Instrumentes zur Erfassung allgemeiner schulischer Grundkompetenzen die pädagogischen Entscheidungen von Lehrkräften unterstützen kann. Von hoher Relevanz könnte zum Beispiel eine Erweiterung des Instrumentes im Hinblick auf die von Seeber und Nickolaus (2010) gewünschte, auf adressatengerechte Entwicklungsangebote ausgerichtete Eingangsdiagnostik sein, wobei in diesem Fall nicht zuletzt Hinweise zur Ergebnisinterpretation und zur Weiterarbeit mit den Ergebnissen zu ergänzen wären.

Die vorliegende Studie leistet schließlich einen methodisch-konzeptionellen Beitrag. Wir möchten mit der Studie explizit auch kommunizieren, dass ein rein statistisches Vorgehen bei der Analyse von DIF und Item-Bias üblicherweise nicht ausreichend ist. Vielmehr sollten Items, die signifikanten DIF aufweisen, einer zusätzlichen qualitativen Beurteilung durch Inhaltsexperten unterzogen werden. Dabei ist es ratsam sowohl die Richtung von DIF einschätzen zu lassen, als auch die Sicherheit, mit der diese Einschätzung getroffen werden kann, sowie mögliche Erklärungen dafür zu erfragen. Durch die gemeinsame Betrachtung der statistischen Ergebnisse und der qualitativen Einschätzungen kann dann verlässlich beurteilt werden, ob Items schwierigkeitsbestimmende, konstruktirrelevante Merkmale aufweisen, so dass von Item-Bias auszugehen ist. Bei der Untersuchung von Testfairness im Bereich schulischer Kompetenzen kann diese inhaltliche Analyse natürlich nur unter Federführung von Inhaltsexperten aus der Fachdidaktik geschehen, welche die entsprechende Expertise einbringen, um schwierigkeitsbestimmende Merkmale der Items zu identifizieren. Um darüber hinaus sicherzustellen, dass die durch Inhaltsexperten identifizierten schwierigkeitsbestimmende Merkmale auch tatsächlich bei der Itembearbeitung seitens der Schülerinnen und Schüler wirksam sind, haben Ercikan et al. (2010) vorgeschlagen, die Methode des lauten Denkens einzusetzen und die Experteneinschätzungen mit diesen abzugleichen. Aufgrund der Post-Hoc-Anlage

der hier vorgestellten DIF-Untersuchung war dieses Vorgehen im vorliegenden Fall nicht möglich, jedoch ist grundsätzlich davon auszugehen, dass ein solches Vorgehen eine noch zuverlässigere Einschätzung ermöglicht, welche Items wirklich Bias hinsichtlich einzelner Probandengruppen aufweisen. Im Zusammenhang mit der Experteneinschätzung hat sich zudem eine aus Forschungsperspektive interessante Frage aufgetan, die auf Basis der vorliegenden Untersuchung zwar ebenfalls nicht beantwortet werden kann, in nachfolgenden Untersuchungen jedoch systematisch untersucht werden sollte. Aufgrund der überraschend geringen Übereinstimmung zwischen der statistisch identifizierten Richtung der DIF-Effekte und den Einschätzungen der Expertinnen in Bezug auf Richtung und Sicherheit eines möglichen DIF-Effektes, ergibt sich nämlich die Frage, an welchen Informationen sich Inhaltsexperten bei der Einschätzung und Bewertung schwierigkeitsbestimmender Item-Merkmalen orientieren und unter welchen Bedingungen Expertinnen und Experten diese Informationen richtig einzuschätzen vermögen.

Abschließend sei darauf hingewiesen, dass die Studie Einschränkung in Bezug auf die Stichprobengröße in dem hier verwendeten unvollständigen Testheftdesign verdeutlicht hat, welche die Identifikation von DIF auf der Ebene einzelner Ausbildungsberufe nicht ermöglichte. Tatsächlich liegt die im Datensatz verfügbare Anzahl von Itemantworten sogar bei der Gruppierung von Ausbildungsberufen zu nur zwei Berufsgruppen nahe an den Mindestanforderungen hinsichtlich der Stichprobengröße für DIF-Analysen. Eine noch differenziertere Untersuchung einzelner Ausbildungsberufe setzt demnach die Gewinnung größerer Stichproben voraus, würde allerdings auch interessante Einblicke in die relativen Leistungsvorteile und -nachteile der Ausbildungsgänge erlauben, ähnlich wie sie etwa von Klieme und Baumert (2001) im Rahmen der TIMS-Studie in Bezug auf länderspezifische Unterschiede und ihre Zusammenhänge mit kognitive Anforderungsbereichen der Mathematik anhand von DIF-Analysen berichtet wurden.

6. Literatur

Adams, R. J., Wu, M. & Wilson, M. (2012). *ACER ConQuest 3.0.1* [Computer software]. Camberwell, VIC: Australian Council for Educational Research.

Baethge, M. (2010). Ein europäisches Berufsbildungs-PISA als politisches und methodisches Projekt. In D. Münk & A. Schelten (Hrsg.), *Kompetenzermittlung für die Berufsbildung* (S. 19–36). Bielefeld: W. Bertelsmann.

Baethge, M. (2012). Large-scale Assessment in der beruflichen Bildung als Mittel zur Qualitätssicherung in der Forschung und Instrument von Politikberatung. In E. Severing & R. Weiß (Hrsg.), *Qualitätsentwicklung in der Berufsbildungsforschung* (S. 127–140). Bielefeld: W. Bertelsmann.

Balkenhol, A. (2015). *Lesen in beruflichen Handlungskontexten – Anforderungen, Prozesse und Diagnostik.* Unveröffentlichte Dissertation, Technische Universität Darmstadt.

Behörde für Schule und Berufsbildung (Hrsg.) (2011). *LAU – Aspekte der Lernausgangslage und der Lernentwicklung Klassenstufen 5, 7 und 9. Bd. 8: HANSE – Hamburger Schriften zur Qualität im Bildungswesen*. Münster: Waxmann.

Behörde für Schule und Berufsbildung (Hrsg.) (2012). *LAU – Aspekte der Lernausgangslage und der Lernentwicklung Klassenstufen 11 und 13. Bd. 9: HANSE – Hamburger Schriften zur Qualität im Bildungswesen*. Münster: Waxmann.

Bernhardt, R., Balkenhol, A., Ebermann, C., Frey, A., Seeber, S., & Ziegler, B. (2013). *Nutzung der adaptiven Tests zur Messung allgemeiner Kompetenzen im Rahmen der ASCOT-Initiative – Manual*. Jena: Friedrich-Schiller-Universität.

Blum, W., Drüke-Noe, C., Hartung, R. & Köller, O. (Hrsg.). (2006). *Bildungsstandards Mathematik konkret. Sekundarstufe I: Aufgabenbeispiele, Unterrichtsideen und Fortbildungsmöglichkeiten*. Berlin: Cornelsen Scriptor.

Bundesministerium für Bildung und Forschung (2014). *Impulse für die Ausbildung der Zukunft. Innovative Kompetenzmessung in einer dynamischen Arbeitswelt*. Berlin: BMBF.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differential item functioning test items. *Educational Measurement: Issues and Practice, 17*, 31–44.

De Ayala, R.J. (2009). *Theory and Practice of Item Response Theory.* New York: Guilford.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Hillsdale, NJ: Lawrence Erlbaum.

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of Think Aloud Protocols for Examining and Confirming Sources of Differential Item Functioning Identified by Expert Reviews. *Educational Measurement: Issues and Practice, 29*, 24–35.

Frey, A. (2012). Adaptives Testen. In H. Moosbrugger, & A Kelava (Hrsg.), *Testtheorie und Fragenbogenkonstruktion* (S. 275–293). Heidelberg: Springer.

Frey, A., & Ehmke, T. (2007). Hypothetischer Einsatz adaptiven Testens bei der Überprüfung von Bildungsstandards. *Zeitschrift für Erziehungswissenschaft, Sonderheft 8*, 169–184.

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.

Frey, A., Heinze, A., Mildner, D., Hochweber, J., & Asseburg, R. (2010). Mathematische Kompetenz von PISA 2003 bis PISA 2009. In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, W. Schneider, & P. Stanat (Hrsg.), *PISA 2009: Bilanz nach einem Jahrzehnt* (S. 153–176). Münster: Waxmann.

Garden, R. A., & Orpwood, G. (1996). Development of the TIMSS Achievement Tests. In M. O. Martin & D. L. Kelly (Eds.), *Third International Mathematics and Science Study (TIMSS) Technical Report, Volume I: Design and Development*. Chestnut Hill, MA: Boston College.

Institut zur Qualitätsentwicklung im Bildungswesen (n. d.). *VERA – Ein Überblick*. Zugriff am 19.03.2015. Verfügbar unter: https://www.iqb.hu-berlin.de/vera

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(3)*, 385–402.

Köller, O., Knigge, M., & Tesch, B. (Hrsg.) (2010). *Sprachliche Kompetenzen im Ländervergleich. Überprüfung der Bildungsstandards in den Fächern Deutsch und erste Fremdsprache in der neunten Jahrgangsstufe*. Münster: Waxmann.

Konsortium HarmoS Naturwissenschaften+ (2009). *Kompetenzmodell und Vorschläge für Basisstandards Naturwissenschaften. Kurzbericht. Provisorische Fassung (vor Verabschiedung der Standards). Stand Juli 2009, mit Ergänzungen und Korrekturen Januar 2010*. Zugriff am 13.11.2014. Verfügbar unter: http://www.edudoc.ch/static/web/arbeiten/harmos/harmoS_kurzbericht_neu.pd

Lehmann, R. & Seeber, S. (Hrsg.) (2007). *ULME III. Untersuchung von Leistungen, Motivation und Einstellungen der Schülerinnen und Schüler in den Abschlussklassen der Berufsschulen*. Hamburg: Behörde für Bildung und Sport.

Lehmann, R. H., Ivanov, S., Hunger, S. & Gänsfuss, R. (2005). *ULME I. Untersuchung der Leistungen, Motivationen und Einstellungen zu Beginn der beruflichen Ausbildung*. Hamburg: Behörde für Bildung und Sport.

Maier, A., Nitzschke, A., Nickolaus, R., Schnitzler, A., Velten, S., & Dietzen, A. (im Druck). Der Einfluss schulischer und betrieblicher Ausbildungsqualität auf die Entwicklung des Fachwissens. In M. Stock, P. Schlögl, K. Schmid, & D. Moser (Hrsg.), *Kompetent – wofür? Life Skills – Beruflichkeit – Persönlichkeitsbildung – Beiträge zur Berufsbildungsforschung* (Reihe: Innovationen in der Berufsbildung, Bd. 9). Innsbruck: StudienVerlag.

Mullis, I.V.S., Martin, M.O., Robitaille, D.F., & Foy, P. (2009). *TIMSS Advanced 2008 International Report: Findings from IEA's Study of Achievement in Advanced Mathematics and Physics in the Final Year of Secondary School.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Nickolaus, R., Geissel, B, & Gschwendtner, T. (2008). Die Rolle der Basiskompetenzen Mathematik und Lesefähigkeit in der beruflichen Ausbildung und die Entwicklung mathematischer Fähigkeiten im ersten Ausbildungsjahr. *bwp@, 14*.

Nickolaus, R., & Norwig, K. (2009). Mathematische Kompetenzen von Auszubildenden und ihre Relevanz für die Entwicklung der Fachkompetenz – ein Überblick zum Forschungsstand. In A. Heinze & M. Grüssing (Hrsg.), *Mathematiklernen vom Kindergarten bis zum Studium* (S. 205–216). Münster: Waxmann.

Nickolaus, R., Straka, G. A., Fehring, G., Gschwendtner, T., Geissel, B., & Rosendahl, J. (2010). Erklärungsmodelle zur Kompetenz- und Motivationsentwicklung bei Bankkaufleuten, Kfz-Mechatronikern und Elektronikern. *Zeitschrift für Berufs- und Wirtschaftspädagogik, Beiheft 23*, 73–87.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Redder, A.; Schwippert, K.; Hasselhorn, M.; Forschner, S. Fickermann, D. & Ehlich, K. (2010). *Grundzüge eines nationalen Forschungsprograms zu Sprachdidagnostik und Sprachförderung. ZUSE-Diskussionspapier Nr. 1*. Zugriff am 13.11.2014. Verfügbar unter: http://epub.sub.uni-hamburg.de/epub/volltexte/2011/9870/pdf/ZUSE_Diskussion001.pdf

Rosendahl, J. & Straka, G. A. (2011). *Effekte personaler, schulischer und betrieblicher Bedingungen auf berufliche Kompetenzen von Bankkaufleuten während der dualen Ausbildung. Ergebnisse einer dreijährigen Längsschnittstudie. ITB-Forschungsberichte 51/2011*. Bremen: Institut Technik und Bildung.

Seeber, S. (2007a). 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. (2007b). Berufsspezifische Fachleistungen in Ausbildungsberufen der Bereiche Gesundheit und Körperpflege In R. Lehmann & S. Seeber (Hrsg.), *ULME III. Untersuchung von Leistungen, Motivation und Einstellungen der Schülerinnen und Schüler in den Abschlussklassen der Berufsschulen* (S. 191–213). Hamburg: Behörde für Bildung und Sport.

Seeber, S. & Lehmann, R. (2011). Determinanten der Fachkompetenz in ausgewählten gewerblich-technischen Berufen. *Zeitschrift für Berufs- und Wirtschaftspädagogik, Beiheft 25*, 95–112.

Seeber, S. & Nickolaus, R. (2010). Kompetenzmessung in der beruflichen Bildung. *BWP Berufsbildung in Wissenschaft und Praxis, 1*, 10–13.

Statistics Canada & Organisation for Economic Cooperation and Development (2005). *Learning a Living. First results of the adult literacy and life skills survey*. Ottawa and Paris: OECD.

Statistisches Bundesamt (2014). *Bildung und Kultur. Berufliche Bildung 2013. Fachserie 11 Reihe 3*. Wiesbaden: Statistisches Bundesamt. Zugriff am 24.11.2014. Verfügbar unter: https://www.destatis.de/DE/Publikationen/Thematisch/BildungForschungKultur/BeruflicheBildung/BeruflicheBildung2110300137004.pdf?__blob=publicationFile

Thompson, N.A., & Weiss, D.J. (2011). A Framework for the Development of Computerized Adaptive Tests. *Practical Assessment, Research, and Evaluation, 16 (1).*

Organisation for Economic Cooperation and Development (OECD) (2014). *PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science* (Volume I, Revised edition, February 2014). Paris: OECD.

Organisation for Economic Cooperation and Development (OECD) (2013). *OECD Skills Outlook 2013: First Results from the Survey of Adult Skills.* Paris: OECD.

Organisation for Economic Cooperation and Development (OECD) (2009). *PISA 2009 Assessment Framework: Key Competencies in reading, mathematics and science*. Paris: OECD.

Organisation for Economic Cooperation and Development (OECD) & Statistics Canada (1995). *Literacy in the Information Age. Final Report of the International Adult Literacy Survey*. Paris: Organisation for Economic Cooperation and Development and the Minister of Industry, Canada.

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

Wise, S. G., & Kingsbury, G.G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicologia, 21*(1), 135–155.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Ziegler, B.; Balkenhol, A., Keimes, C., Rexing, V. (2012): *Diagnostik „funktionaler Lesekompetenz". bwp@*, *22*, 1–19.

Zwick, R. (2003). The assessment of differential item functioning in computer-adaptive tests. In W. J. van der Linden & C. A. W. Glas, (Ed.), *Computerized Adaptive Testing: Theory and Practice*. Dordrecht: Kluwer.

Anschrift der Autoren: Christian Spoden, christian.spoden@uni-jena.de, Andreas Frey, andreas.frey@uni-jena.de, Raphael Bernhardt, raphael.bernhardt@uni-jena.de, Friedrich-Schiller-Universität Jena, Institut für Erziehungswissenschaft, Professur Empirische Methoden der erziehungswissenschaftlichen Forschung, Am Planetarium 4, 07737 Jena

Susan Seeber, susan.seeber@wiwi.uni-goettingen.de, Georg-August-Universität Göttingen, Wirtschaftswissenschaftliche Fakultät, Professur für Wirtschaftspädagogik und Personalentwicklung, Platz der Göttinger Sieben 5, 37073 Göttingen,

Aileen Balkenhol, balkenhol@bp.tu-darmstadt.de, Birgit Ziegler, ziegler@bpaed.tu-darmstadt.de , Technische Universität Darmstadt, Institut für Allgemeine Pädagogik und Berufspädagogik, Professur Berufspädagogik mit dem Schwerpunkt Berufsbildungsforschung, Didaktik beruflicher Bildung und Professionalisierung von Lehrenden, Alexanderstr. 6, 64283 Darmstadt