

Kompetenzorientierte Assessments in der beruflichen Bildung – Am Beispiel der Ausbildung von Industriekaufleuten¹

KURZFASSUNG: Das systemische Verstehen von unternehmerischen Prozessen stellt eine zentrale Qualifikationsanforderung für die Aus- und Weiterbildung von Industriekaufleuten dar. Dieser Sachverhalt kommt in den Prüfungsaufgaben der Aufgabenstelle für kaufmännische Abschluss- und Zwischenprüfungen (AkA) zum Ausdruck, die unter der Überschrift „Geschäftsprozesse“ zusammengefasst werden. Mit Hilfe der Aufgaben des Prüfungsjahrganges 2008 werden zwei Fragestellungen versucht zu beantworten: Zum einen wird geprüft, inwieweit die Assessments im Bereich der kaufmännischen Berufsbildung psychometrisch valide und vergleichbar sind. Dies umfasst insbesondere eine Prüfung der Konstruktvalidität. Zum anderen wird auf Basis dieser Ergebnisse versucht, neue Formen einer fachdidaktisch begründeten und empirisch überprüfbaren Domänen- und Kompetenzmodellierung zu entwickeln. Mit Hilfe einer solchen Kompetenzorientierung von Assessmentverfahren im Bereich der Berufsbildung können – perspektivisch – sowohl individuelle Voraussetzungen als auch mögliche Förderpotentiale und notwendige Fördermaßnahmen präziser bestimmt werden.

ABSTRACT: The paper is dealing with the economic understanding of trainees in the field of Business and Administration using final examination for industrial clerks (designed by AkA). The analysis of assessments in the field of VET and Professional Development includes the evaluation of both psychometric validity and comparability as well as construct validity. Based on that analysis scientifically founded evaluation models will be developed. Simultaneously, there is a need to learn more about complex situations and their descriptive components to design competence-oriented assessment tasks. This also includes a deep insight into the modeling and scaling approach, but also a constructive discussion about theoretical assumptions regarding the competence construct that is measured in a specific domain. Based on empirical findings, the project aims to develop suitable processes and teaching assistance in VET to foster learning and teaching processes in the domain.

1 Projektziele und Forschungsdesiderate

Im Rahmen des Projekts „Kompetenzorientierte Assessments in der beruflichen Bildung“ sollen die Ausbildungsleistungen, wie sie sich in der AkA-Prüfung niederschlagen, mit Hilfe eines Domänen- und Kompetenzmodells beschrieben werden. Damit werden vor allem zwei Ziele verfolgt:

- Die Beschreibung der Ergebnisse der AkA-Prüfungen über Fähigkeits- und Schwierigkeitsparameter (mit Hilfe der Item Response-Theorie).
- Die Umsetzung der Ergebnisse in ein Modell bezüglich eines systemischen Verstehens von Geschäftsprozessen, das sowohl fachdidaktisch-inhaltliche Facetten der Domäne als auch verschiedene Kompetenzdimensionen integriert.

Den Ausgangspunkt beschreibt eine vertiefte Analyse der Prüfungsaufgaben, mit der fachdidaktisch relevante Inhaltsbereiche zu sichten und bezüglich verschie-

¹ Der Beitrag ist entstanden im Rahmen des Teilprojekts „Kompetenzorientierte Assessments in der beruflichen Bildung“ (WI 3597/1-1) des DFG-Schwerpunktprogramms „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen (SPP 1293).

dener Aufgabenanforderungen zu lokalisieren sind. Damit wird eine Grundlage dafür geschaffen, kognitive Prozesse innerhalb identifizierter Teilkompetenzen für die Industriekaufleuteausbildung beschreiben zu können. Neben den Fragen psychometrischer Kompetenzmodellierungen rücken im Rahmen von Prüfungs- und Zertifizierungsprozessen Aspekte der Fairness und der diagnostischen Relevanz in den Mittelpunkt der Betrachtung. Etwa derart, dass die Aufgaben, die für Auszubildende gestellt werden, die mit einer unterschiedlichen Vorbildung in Betrieben ausgebildet werden, die sich nach der Branche, nach der Betriebsgröße sowie nach dem technologischen Stand in erheblichem Ausmaß unterscheiden, vergleichbare Antwortchancen – im Sinne der Item- und Testfairness – für alle Prüflinge gewährleisten müssen.

Mit den AKA-Prüfungsaufgaben, die auf der Grundlage der lernfeldorientierten Curricula und der Ausbildungspläne der kaufmännischen Ausbildungsberufe gestaltet werden, wird die Forderung erhoben, berufliche Handlungskompetenz abzubilden. Diese Forderung lässt sich bislang nicht einlösen. Hierfür sprechen mindestens zwei Gründe: Die IHK-Abschlussprüfungen als summatives Assessment eines Ausbildungsganges sind (1) nicht unter dem Fokus einer beruflichen Kompetenzmessung entwickelt worden, wenngleich die Gestaltung der Bildungsgänge dem Primat der beruflichen Handlungskompetenz folgt bzw. folgen sollte, und haben (2) keine Anbindung an Entwicklungen der neueren Lehr-Lern-Theorie, so dass Kompetenzerwerbsprozesse und damit verbundene Vermittlungsprozesse nicht angemessen erfasst werden. Für den kaufmännischen Bereich kommt hinzu, dass keine standardisierten Vorbilder für die Gestaltung kompetenzorientierter Assessments mit einem inhaltlichen Fokus auf betriebswirtschaftliche Abläufe und Entscheidungen existieren; Assessments im Ökonomiebereich sind bislang vorrangig auf volkswirtschaftliche Inhalte konzentriert (vgl. u. a. Test of Economic Understanding (TEU): NATIONAL COUNCIL ON ECONOMIC EDUCATION, 1964; TEST OF ECONOMIC LITERACY (TEL): SOPER & WALSTAD 1987; WALSTAD & REBECK, 2001; sowie die deutschen Adaptationen: BECK, 2001).

Insgesamt lässt sich damit ein erstes Forschungsdesiderat auf der Ebene der Assessments selbst verorten: Die Entwicklung kompetenzorientierter Assessments sollte (1) dem Anspruch moderner Assessmentmodelle genügen (vgl. u. a. WHITE & FREDRIKSEN, 1998; WILSON & SLOANE, 2000; MISLEVY & HAERTEL, 2006), (2) auf theoretischen Modellen basieren, die sowohl die Struktur von Fähigkeiten als auch deren Graduierung beschreiben, und (3) inhaltlich den Fokus stärker auf das Verstehen von Zusammenhängen und Konzepten anstatt auf das Wissen von Fakten und Prozeduren legen – was natürlich Wissensbestände voraussetzt (vgl. ACHTENHAGEN, 2004).

Ein zweites Forschungsdesiderat fokussiert auf den Nutzen von Assessments. Hier geht es um die Zielsetzung, dass mit Hilfe der kaufmännischen Abschlussprüfungen unter einer kompetenzorientierten Perspektive arbeitsmarkt- sowie bildungsbezogene Entscheidungen Einzelner besser als bisher vorbereitet werden können. Angesprochen sind in diesem Zusammenhang Effekte pädagogisch-psychologischer Diagnostik. Für den Bereich der beruflichen Kompetenzmessung stehen dabei insbesondere die Relationen zwischen der fachlichen Domäne und damit des Inhaltsbereichs von beruflicher Kompetenz auf der einen Seite und die der Person als Merkmalsträger von Kompetenz und als Akteur in beruflichen Settings auf der anderen Seite im Vordergrund. In dieser Gegenüberstellung wird bereits

deutlich, dass berufliche Kompetenzmessung nicht auf curriculare Inhalte und Wissenskategorien begrenzt sein kann, sondern dass ebenso funktionale Aspekte der Bildung einschließlich verschiedener Verhaltenskomponenten und differenzierter Einstellungen in realen beruflichen Lebenswelten die Diskussion zu bestimmen haben (vgl. u. a. ROTH, 1971; BAETHGE, ACHTENHAGEN, ARENDS, BABIC, BAETHGE-KINSKY & WEBER, 2006; STRAKA & MACKE, 2010). Während die internationale Forschungsliteratur Hinweise für die Verknüpfung von Diagnostik und Assessment auf unterschiedlichen Ebenen (SCHEERENS, GLAS & THOMAS, 2003) sowie – noch uneinheitliche – empirische Befunde zu den Auswirkungen von Assessment-Systemen liefert (CHENG, WATANABE & CURTISET, 2004; FUHRMAN & ELMORE, 2004; zusammenfassend KLIEME & LEUTNER, 2006), können national und hier insbesondere für den Bereich der beruflichen Bildung keine Aussagen gemacht werden. Die in der beruflichen Bildung geführte vorrangig epistemologische – und dabei den internationalen Forschungsstand völlig ignorierende bzw. nicht akzeptierende – Diskussion beispielsweise zur geeigneten Methodenwahl (hier aktuell: MUSEKAMP, SPÖTTL & BECKER, 2010) oder zu Sinn und Zweck systemischer und systematischer Kompetenzmessungen auf nationaler und auch internationaler Ebene (u. a. GROLLMANN & SPÖTTL, 2008) muss den Eindruck verstärken, dass dies politisch auch nicht intendiert ist. Von weitaus entscheidenderer Bedeutung sind in diesem Zusammenhang jedoch die Fragen, welche Aktionsradien, welche Handlungs- und Entscheidungsspielräume Absolventen beruflicher Bildung auf dem Arbeitsmarkt haben – und damit die Frage nach typischen Anforderungsfeldern, aus denen sich „kompetentes“ Verhalten ablesen ließe, sowie damit verbunden die Frage nach Assessments, die berufliches Verhalten stimulieren und dadurch dieses Verhalten einer Kompetenzmessung zugänglich machen. Dabei ist es unumgänglich, dass die Auswertung und Analyse (a) dem „state of the art“ der internationalen empirischen Forschung (vgl. hierzu u. a. AERA, APA & NCME, 1999) entspricht und (b) rückgebunden an die Anforderungsprofile des Arbeitsmarktes und die Handlungsparameter der Beschäftigten erfolgt.

2 Assessment und Assessmentmodelle

Die Herausforderungen, die an Erhebungen im Bereich des Lernens und Leistens gestellt werden, nehmen mit der Entwicklung der entsprechenden technischen und statistischen Methoden kontinuierlich zu (MISLEVY, 1994; Beiträge in WILSON, 2004a): Diese sollen das Lernen begleiten, die Entwicklung des Lernenden beschreiben und Prognosen über Lernverläufe und berufliche Karrieren erlauben. Die Assessmentsituationen werden dabei zunehmend komplexer – dies gilt sowohl im Hinblick auf den Vernetzungsgrad der beobachtbaren Performanz als auch auf den Abstraktionsgrad der zugrundeliegenden Aufgabenstellung. Für deren Modellierung sind Abstimmungsprozesse erforderlich, die klar definieren, was und wie gemessen werden soll und unter welchen Zielsetzungen eine Messung zu sehen ist. Im Folgenden wird veranschaulicht, welche Gestaltungskriterien für die Entwicklung von Testinstrumenten zur Erfassung beruflicher Kompetenz theoretisch auszuarbeiten sind, um Bezugspunkte und Beurteilungsmaßstäbe sowie die praktische Relevanz des Assessments begründen und in ein Messmodell umsetzen zu können. Die zentrale Leitlinie zur Formulierung relevanter Gestaltungskriterien ist der dem Kompetenzkonstrukt innewohnende Kontextbezug. Kompetenzen als „realized abilities“

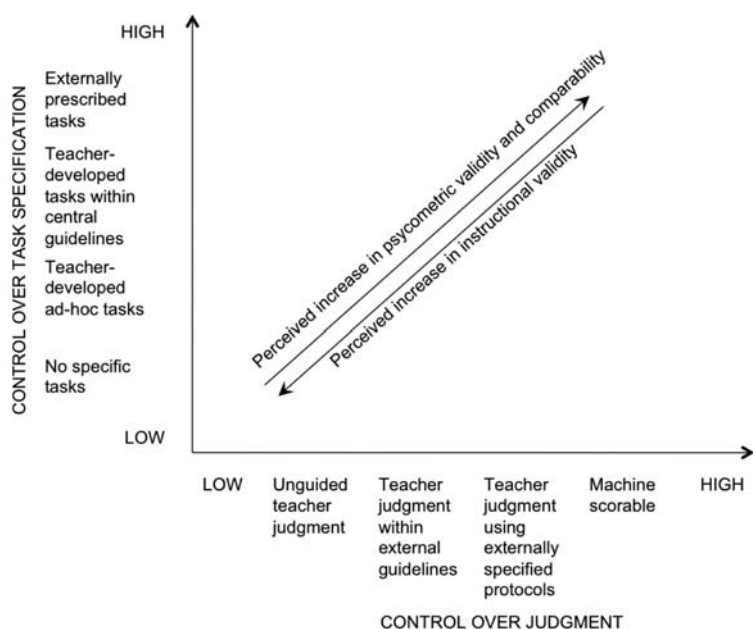
(CONNEL, CHERIDAN & GARDNER, 2003, p. 142; HARTIG, KLIEME & LEUTNER, 2008) beziehen sich demnach immer auf den Umgang mit und die Bewältigung von spezifischen Anforderungssituationen. Diese Situationen werden aus Domänen bzw. fachbezogenen Leistungsbereichen abgeleitet und bilden den Ausgangspunkt eines jeden Assessments. Entsprechend der „Standards for educational and psychological testing“ (AERA, APA & NMCE, 1999) besteht die Entwicklung eines Tests aus vier Phasen:

„(1) delineation of the purpose(s) of the test and the scope of the domain (content and skills) to be measured; (2) development and evaluation of the test specifications; (3) development, field testing, evaluation, and selection of items and scoring guides and procedures, and (4) assembly and evaluation of the test for operational use“ (ebd., p. 37).

WILSON (2004b, p. 2ff.) weist in diesem Zusammenhang und unter Bezug auf die in dem KWSK-Assessment Triangle (NRC Report: Knowing What Students Know) angesprochene Kohärenz der Entwicklung, des Einsatzes und der Verwertung von Assessments auf drei zentrale Herausforderungen hin, die den Nutzen von Assessments in Lern- und Arbeitskontexten nachhaltig beeinflussen (vgl. *term of coherence*, PELLEGRINO, CHUDOWSKY & GLASER, 2001). Er geht im Sinne von BLACK und WILIAM (u. a. 2004) davon aus, dass Assessments – und hier auch solche, die Systemvergleiche anstreben, oder solche, mit denen Bildungsbiographien zertifiziert werden – in erster Linie der Förderung und Stützung von Lern- und Arbeitsprozessen zu dienen haben (vgl. hierzu auch Beiträge in MOSS, PULLIN, GEE, HAERTEL & YOUNG, 2008). Unter dieser Bedingung ist eine erste Herausforderung, für jede Art von Assessment Inhalts- und Konstruktvalidität sicherzustellen. Nur unter der Validitätsvoraussetzung lassen sich die Ergebnisse eines Assessments direkt für den Unterricht, für die betrieblichen Arbeitsabläufe und/oder für Bildungsgangentscheidungen nutzen. Eine zweite Herausforderung stellt die Praktikabilität bzw. Nutzerfreundlichkeit von Assessments dar. Sollen die Ergebnisse eines Assessments direkt für den Unterricht oder betriebliche Abläufe nutzbar sein, müssen sowohl die Auswertungsverfahren als auch die Interpretationen der Ergebnisse transparent und einfach verständlich sein. Den Nutzern von Assessments ist zu vermitteln, wie Ergebnisse erzeugt werden und welche Aussagekraft diese Ergebnisse besitzen (vgl. *term of comprehensiveness*, PELLEGRINO, CHUDOWSKY & GLASER, 2001). Dies stellt eine Voraussetzung dafür da, dass die Ergebnisse verschiedener Assessments – von Klassenarbeiten bis hin zu internationalen Vergleichsstudien – Eingang in die aktuelle Bildungslandschaft finden. Eng damit verbunden ist die Frage, wie diese verschiedenen Assessments aufeinander bezogen werden können. Eine dritte Herausforderung ist demnach, wie Ergebnisse von Assessments für schulische und betriebliche Leistungserfassungen genutzt werden können und vice versa (top down-button up-Prinzip; vgl. *term of continuity*, PELLEGRINO, CHUDOWSKY & GLASER, 2001). Die in der Forschung entwickelten Assessmentmodelle sollen hierfür Hilfsinstrumentarien zur Verfügung stellen. Das generelle Prinzip verschiedener Assessmentmodelle (vgl. u. a. SHAVELSON, BAXTER & PINE, 1991, GUSTAFSSON & UNDHEIM, 1996; WHITE & FREDERIKSEN, 1998; WILSON & SLOANE, 2000) ist, dass sie für die Entwicklung von Assessments und für die Interpretation von Ergebnissen strukturelle Muster bzw. Beziehungsgefüge anbieten, mit denen sich Assessments unterschiedlicher inhaltlicher Färbung und Zielsetzung miteinander vergleichen lassen. Damit stellen moderne Assessmentmodelle eine (abstrakte) Struktur im Sinne normativer Zielvorgaben zur Verfügung, anhand derer

sich domänenbezogene Testinhalte beschreiben und konstruieren lassen. Ein zentrales Element eines Assessmentmodells ist demnach eine konzeptionelle Theorie darüber, wie Lernprozesse funktionieren, wie Wissen erworben und Kompetenzen aufgebaut werden (kognitiver Aspekt). Diese Theorie sollte auch Aussagen darüber treffen können, welche Zusammenhänge zwischen Lehr- und Lernprozessen bestehen, wie sich Lerninhalte vermitteln lassen und welche Variablen das Lernen fördern bzw. behindern können (Bezug zu Erkenntnissen der neueren Lehr-Lern-Theorie). Im Ergebnis sollte auf Basis dieser Heuristik ein Modell entworfen werden können, das kognitive Prozesse domänenunabhängig definiert und zugleich in Beziehung zu den spezifischen Inhalten, Arbeitstechniken und Zielen einer Domäne setzen kann. Bei der Entwicklung nationaler Bildungsstandards werden derartige Überlegungen beispielsweise genutzt, um über eine Auflistung von Prozessen und Zielen in unterschiedlichen Fachbereichen zu einer Vergleichbarkeit der kognitiven Anforderungen zu gelangen (u. a. Department of Education and Science, 1987; KMK, 2003).

Abbildung 1: Testformate und Validität (WILSON, 2004b, p. 10)



In Abstimmung mit den spezifischen Inhalten, Arbeitstechniken und Zielen einer Domäne ist zu prüfen, welche Testformate geeignet sind und wie sich ein Assessment aus verschiedenen Testformaten möglichst optimal zusammensetzt. Insbesondere die Entwicklung von Testformaten und Testitems steht dabei in dem Konflikt zwischen psychometrischer Validität und Vergleichbarkeit und der instruktionalen Validität (Abbildung 1). In der Literatur wird dieser Konflikt in der Regel über den Vergleich von standardisierten Multiple-Choice-Tests und alternativen Assessmentformen dargestellt (vgl. u. a. performance assessment: SHAVELSON, RUIZ-PRIMO & WILEY, 1999; portfolio assessment: WOLF, 1989; computer simulation: BAXTER, 1999).

Abbildung 1 verdeutlicht, dass mit zunehmender psychometrischer Validität und Vergleichbarkeit eine Reduktion der instruktionalen Validität einhergeht. Oder anders formuliert: Die in Unterrichtsprozessen üblichen, aus der Interaktion entstehenden Rückfragen oder ad hoc-formulierten Aufgaben erfüllen zwar die Kriterien der instruktionalen Validität, nicht jedoch jene der psychometrischen Validität, für die sowohl kontrollierbare Aufgabenmerkmale als auch eine hohe Kontrolle über die Bewertungskriterien vorliegen müssen. Für die Bewertung und Analyse kaufmännischer Abschlussprüfungen lässt sich dieser Zusammenhang als kritische Realitätsprüfung nutzbar machen.

3 Status Quo: Kaufmännische Abschlussprüfungen

3.1 Inhalte und curriculare Validität der IHK-Abschlussprüfung Industriekaufmann/Industriekauffrau

Die kaufmännischen Abschlussprüfungen im Ausbildungsberuf Industriekaufmann/-frau folgen nicht mehr dem Leitbild des abwicklungsorientierten Sachbearbeiters in klassischen kaufmännischen Funktionsbereichen, sondern dem des kundenorientierten Sachbearbeiters, der team-, prozess- und projektorientiert unter Verwendung aktueller Informations-, Kommunikations- und Medientechniken an der Erstellung kundengerechter Problemlösungen arbeitet (u. a. IHK AACHEN, 2008). Damit soll auch in den Prüfungen berücksichtigt werden, dass die Ausbildung einer geschäftsprozessorientierten Logik folgt. Der schriftliche Teil der Abschlussprüfungen besteht dabei aus drei Bereichen: aus praxisbezogenen Aufgaben der Bereiche (1) kaufmännische Steuerung und Kontrolle und (2) Wirtschafts- und Sozialkunde sowie aus (3) situationsbezogenen Aufgaben und Fallbeispielen des Prüfungsbereichs „Geschäftsprozesse“. Der letztgenannte Prüfungsbereich ist mit 180 Minuten Prüfungsdauer und 40 Prozent Gewicht der Prüfungsleistung der umfangreichste Teilprüfungsbereich. Der Prüfungsbereich „Geschäftsprozesse“ fasst Aufgaben aus den Gebieten Marketing und Absatz, Beschaffung und Bevorratung, Personal sowie Leistungserstellung zusammen. Im Zentrum dieses Prüfungsbereichs stehen Prozesse und komplexe Sachverhalte, anhand derer die Lernenden Geschäftsprozesse analysieren sowie Problemlösungen ergebnis- und kundenorientiert entwickeln sollen. Die zur Anwendung kommenden „situativen“ Aufgabentypen sollen praxis-, adressaten-, aktivitäts- und entscheidungsorientiert eine berufliche Situation abbilden und für die Berufsausübung wesentliche Kompetenzen abfragen (handlungsorientiertes Prüfen; AkA, 2009). Im Prüfungsbereich „Geschäftsprozesse“ werden Prüfungsaufgaben mit freier Beantwortung eingesetzt, für deren Korrekturen Lösungshinweise gegeben werden. Unter Rückgriff auf *Abbildung 1* dürfte vor diesem Hintergrund eine besondere Schwierigkeit darin liegen, psychometrische Validität und Vergleichbarkeit zu gewährleisten. Vorausgesetzt wird hier zunächst, dass (a) die Spezifika der einzelnen Aufgaben von den Prüfern vergleichbar interpretiert werden und dass (b) die Aufgaben angemessen fachlich korrekt konstruiert sind.

Die IHK-Abschlussprüfung Industriekaufmann/-frau des Prüfungsjahrgangs 2008 umfasst im Prüfungsbereich „Geschäftsprozesse“ sechs gebundene Aufgaben, von denen mit einer Gesamtpunktzahl von 38 Punkten (zwei Teilaufgaben) das Aufgabengebiet Marketing und Absatz (A1 und A2) den größten Anteil stellt. Die

drei anderen Aufgabengebiete: Beschaffung und Bevorratung (A3), Personal (A4) sowie Leistungserstellung (A5 und A6), sind gleich gewichtet.

Tabelle 1: Aufgaben des Prüfungsbereichs „Geschäftsprozesse“ im curricularen Vergleich

Aufgaben Prüfungsbereich „Geschäftsprozesse“ 2008	Anzahl	Punkte	Umfang Lernort Betrieb	Umfang Lernort Berufsschule	curriculares-Gewicht*
Marketing und Absatz	2	38	5 bis 9 Monate	160 Stunden	26,67%
Beschaffung und Bevorratung	1	20	5 bis 7 Monate	80 Stunden	13,33%
Personal	1	21	2 bis 6 Monate	80 Stunden	13,33%
Leistungserstellung	2	21	6 bis 10 Monate	280 Stunden	46,67%

* Anmerkung: Curriculares Gewicht = Anteil der Aufgabengebiete entsprechend des Rahmenlehrplans für den Ausbildungsberuf Industriekaufmann/-frau (KMK, 2002) bezogen auf 600 Zeitstunden.

In *Tabelle 1* werden die Aufgaben im curricularen Vergleich dargestellt. Von besonderem Interesse ist der Anteil der Zeitstunden entsprechend des Rahmenlehrplans für den Ausbildungsberuf Industriekaufmann/-frau (KMK, 2002) in den Lernfeldern, die inhaltlich den vier Teilaufgabengebieten entsprechen. Von den insgesamt 880 Zeitstunden des Rahmenlehrplans lassen sich 600 direkt den Aufgaben des Prüfungsbereichs „Geschäftsprozesse“ zuordnen. Entsprechend der Systemorientierung in der Betriebswirtschaftslehre stehen Prozesse der Leistungserstellung dabei zentral im Fokus – im Rahmenlehrplan zeigt sich dies in einem Umfang von 280 Zeitstunden. Auch die sachliche sowie zeitliche Gliederung des Ausbildungsrahmenplans für die Berufsausbildung zum Industriekaufmann/zur Industriekauffrau macht die besondere Stellung dieses Inhaltsbereichs deutlich.

Für die Beurteilung curricularer Validität eines Tests kann unterstellt werden, dass der Test lehrzielorientiert konstruiert wurde und die Iteminhalte die essenziellen Facetten des zu messenden Merkmals adäquat abbilden. Auf Basis logischer und fachlicher Überlegungen sollten folglich die ausgewählten Aufgaben eine repräsentative Stichprobe der Aufgaben darstellen, wie sie auch innerhalb der Instruktionsprozesse zum Einsatz gekommen sind. Dies umfasst spezifische Aufgabenmerkmale – wie Design, Präsentationsform, Hilfsmittel – sowie die spezifischen Aufgabenschwierigkeiten im Rahmen der unterrichteten Inhalte. Im Hinblick auf die *curriculare Validität* des Prüfungsbereichs „Geschäftsprozesse“ lässt sich damit in einem ersten Zugriff festhalten, dass der inhaltlichen Dominanz der Prozesse der Leistungserstellung zum Prüfungszeitpunkt 2008 nicht adäquat entsprochen wird, mit der Folge, dass Fragen der Wertschöpfung und damit des quantifizierbaren Prozesses der Leistungserstellung insgesamt unterrepräsentiert sind. Dies ist insofern von besonderer Relevanz, als im Rahmen der Vorbereitungen für ein international vergleichendes Large-Scale Assessment in der beruflichen Bildung (BAETHGE & ARENDS, 2009) Aspekte der Wertschöpfung und der Leistungserstellung

als die Bereiche des beruflichen Tätigkeitsprofils von kaufmännischen Angestellten im industriellen Sektor identifiziert wurden, die die höchste Akzeptanz und in der Bewertung internationaler Experten die höchste interne Konsistenz aufweisen. Die internationale Analyse der Arbeitsmarktanforderungen zeigt, dass die Prozesse der Leistungserstellung für einen internationalen Vergleich von besonderem Interesse wären, da mit ihnen die Internationalität der Geschäftsbeziehungen, Aspekte der Kunden- und Lieferantenbindungen und somit insgesamt ein Verständnis für betriebliche Geschäfts- und Arbeitsprozesse getestet werden können (BREUER, HILLEN & WINTHER, 2009).

Nach NUNNALLY (1978) ist curriculare Validität als Kategorie der Inhaltsvalidität an zwei Standards gebunden: (1) an das Sampling der Testitems und (2) an die Konstruktion der Items. Damit werden zwei Beurteilungsmaßstäbe notwendig: Zum einen ist der Beitrag eines Items zur Definition des zu messenden Merkmals zu beurteilen und zum anderen das Set von Items, das alle Facetten des Merkmals repräsentieren soll. Die Beurteilung fokussiert damit auf die Beziehung zwischen Operationalisierung und empirischen Evidenzen. Für den Prüfungsbereich „Geschäftsprozesse“ gilt, dass authentische, situative Arbeitsaufgaben durch die Anwendung von erworbenem Wissen, von Kenntnissen und Fähigkeiten so bewältigt werden sollen, dass die Fähigkeit gezeigt wird, eine qualifizierte, an Geschäftsprozessen ausgerichtete kaufmännische Berufstätigkeit auszuüben (Ausbildungsordnung, § 3 Abs. 2). Die einzelnen Items des Prüfungsbereichs sollten folglich einen unmittelbaren Ausschnitt aus diesem Verhaltensbereich abbilden. Eine inhaltliche Analyse der Testitems zeigt jedoch gravierende Mängel im Hinblick auf (1) die Umsetzung der Geschäftsprozessidee, (2) die Authentizität der Fälle sowie (3) die Berücksichtigung verschiedener kognitiver Anforderungen, mit denen auf das zu untersuchende Verhalten geschlossen werden kann (vgl. *Tabelle 2*; zur Kritik an „unechten Situationsaufgaben“ vgl. REETZ & HEWLETT, 2009).²

Die inhaltlichen Schwächen des Prüfungsbereichs „Geschäftsprozesse“ erfassen sowohl das Sampling der Testitems als auch deren Konstruktion. Insgesamt ist festzustellen, dass sich die Items hinsichtlich ihrer erwarteten Schwierigkeit kaum voneinander unterscheiden. Alle Items stellen auf die Aktivierung und Anwendung deklarativer und prozeduraler Wissensbestände in wohlbekannten Anforderungssituationen ab. Damit werden ausschließlich kognitive Prozesse der Reproduktion und Anwendung angesprochen. Einer Variation der Komplexität der Testitems wird auch dadurch entgegengewirkt, dass die Unternehmensbeschreibung zwar eine situative Einbindung der Testitems stützt, dass auf diese bei der Bearbeitung der Testitems jedoch in keinem Fall zurückgegriffen werden muss; Prozesse der Informationssuche oder der Identifikation lösungsrelevanter Daten werden damit nicht verlangt. Von besonderer Relevanz ist darüber hinaus, dass dem Kriterium der Authentizität kaum Beachtung geschenkt wird und eine Konstruktion von Aufgaben im Rahmen komplexer Geschäftsprozesse nicht intendiert zu sein scheint. Dies zeigt sich darin, dass ausschließlich statische Aspekte zu bearbeiten sind, die weder eine zeitliche noch eine prozessuale Dimension aufweisen und damit wenig plausibel sind. Das

2 Die Aufgaben des Prüfungsbereichs behandeln Einzelaspekte eines Geschäftsprozesses, z. B. Beschreiben einer spezifischen Preisstrategie (Aufgabe A1_1.5 des Bereichs Marketing und Absatz), Verfassen einer Mängelrüge (Aufgabe A3_3.5 des Bereichs Beschaffung und Bevorratung), Formen der Zusammenarbeit (Aufgabe A4_4.1 des Bereichs Personal) sowie Durchführung eines Kostenvergleichsverfahrens (Aufgabe A5_5.1 des Bereichs Leistungserstellung).

Tabelle 2: Beispiel für inhaltliche Schwächen des Prüfungsbereichs „Geschäftsprozesse“

(Unzureichende) Umsetzung der Geschäftsprozessidee	(unzureichende) Authentizität der Fälle	(Unzureichende) Variation der kognitiven Anforderung
Unternehmensbeschreibung Schoko-Tender AG	Ein Bezug (wenn nicht angezeigt) auf die Unternehmensbeschreibung ist in keiner Aufgabe notwendig.	Keine Informationssuche bzw. Reduktion betrieblicher Komplexität notwendig
Durchführung eines Kostenvergleichsverfahrens (statischer Teilaspekt der Wertschöpfungskette)	Fallsituation unangemessen konstruiert: Absatzrückgang, Umstellung der Produktion anlässlich eines saisonalen Ereignisses macht die Anschaffung einer neuen Maschine notwendig	Anwendung und Aktivierung prozeduraler Wissensbestände (Berechnung der Gesamtkosten) Anwendung und Aktivierung deklarativer Wissensbestände (Beschreibung des Kostenvergleichsverfahrens)

Fehlen von Testitems, in denen Entscheidungen oder Validierungen im Sinne eines Agierens in Geschäftsprozessen überprüft werden, schränkt die Variation der erwarteten Aufgabenschwierigkeiten darüber hinaus erheblich ein.

Für die Charakteristik von Aufgabenschwierigkeiten liegen im Bereich der beruflichen Bildung erste Befunde vor. So lassen sich für den kaufmännischen und gewerblich-technischen Bereich übereinstimmend die Art der kognitiven Auseinandersetzung (Taxonomie), eigenständige Modellierungsnotwendigkeiten sowie der Vertrautheitsgrad der Aufgaben als Prädiktoren der Itemschwierigkeit identifizieren (WINTHER & ACHTENHAGEN 2009; NICKOLAUS, GSCHWENDTNER & ABELE, 2009; SEEBER, 2008). Im Hinblick auf diese Facetten der Aufgabenschwierigkeit gilt, dass für die Aufgaben des Prüfungsbereichs „Geschäftsprozesse“ keine sinnvolle Skalierung gefunden werden kann, da das einzelne Komplexitätsmerkmal auf eine bzw. auf zwei Stufen begrenzt bleibt. Für den Prüfungsbereich „Geschäftsprozesse“ lässt sich damit zunächst kein angemessenes Modell entsprechend der Intention und Ausgestaltung der Testaufgaben entwickeln. Unter Rückgriff auf die von NUNNALLY (1978) definierten Beurteilungsmaßstäbe ist damit auf Basis der konzeptuellen Analyse festzuhalten, dass die Fähigkeit, ergebnis- und kundenorientiert in Geschäftsprozessen handeln zu können, durch die Items des Prüfungsbereichs „Geschäftsprozesse“ nicht hinreichend repräsentiert wird. Zum einen sind die essenziellen Aspekte des zu messenden Merkmals über die Items nicht identifizierbar und zum anderen zeigen sich im Hinblick auf die Iteminhalte und die Itemschwierigkeiten konzeptionelle Schwächen in der Testkonstruktion.

3.2 Skalen- und Itemanalyse des Prüfungsbereichs „Geschäftsprozesse“

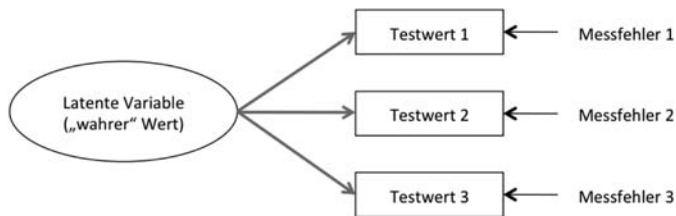
Die erkennbaren inhaltlichen Schwierigkeiten des Prüfungsbereichs „Geschäftsprozesse“ spiegeln sich in den empirischen Befunden. Für die Analyse stehen

Informationen aus sechs IHK-Bezirken zur Verfügung. Insgesamt liegen 1.768 Prüfungshefte einschließlich der Beurteilung durch die Lehrenden vor. Die Analyse der Prüfungsaufgaben erfolgt sowohl auf Basis der Klassischen Testtheorie als auch unter Rückgriff auf die Item Response-Theorie. Mit der Klassischen Testtheorie und der Item Response-Theorie liegen zwei Möglichkeiten vor, wie aus dem Testverhalten auf das zu untersuchende Personenmerkmal (latente Variable) geschlossen werden kann (vgl. *Abbildung 2*):

Abbildung 2: Unterschied zwischen den Testtheorien (in Anlehnung an EMBRETSON, 1999; entnommen WINther, 2010, S. 126)

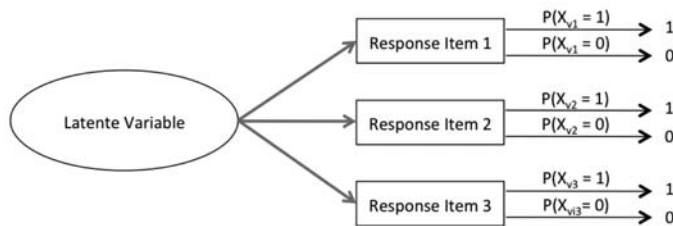
Klassische Testtheorie:

Beobachteter Wert = „wahrer“ Wert + Messfehler



Probabilistische Testtheorie (IRT):

Wahrscheinlichkeit $P(X_{vi} = 1) = \text{Funktion}[(\text{Personenfähigkeit}) - (\text{Itemschwierigkeit})]$



In der Klassischen Testtheorie ergibt sich der beobachtbare Testwert als Summe des „wahren“ Testwerts und eines Messfehlers. Dabei gilt, dass Messfehler eine uniforme Normalverteilung aufweisen, einen Erwartungswert von Null haben und unkorreliert mit allen anderen Variablen sind. Nach EMBRETSON (1999, pp. 5/6) weist die Klassische Testtheorie trotz der Messfehlerannahmen einige Beschränkungen auf, die sich insbesondere auf die Forderung nach test- und stichprobenunabhängigen Schätzern beziehen lassen:

- Der Messfehler lässt sich nur auf Items eines spezifischen Tests beziehen, da keine Annahmen über mögliche Variationen der Itemparameter im Rahmen der Klassischen Testtheorie getroffen werden können.
- Die Messfehlertheorie basiert auf zwei unabhängigen Variablen (dem „wahren“ Wert und dem Messfehler). Diese lassen sich jedoch nicht wirklich trennscharf separieren.

- Die Itemeigenschaften sind nicht an das Testverhalten gebunden und müssen daher über Zusatzstatistiken abgesichert werden (Varianzen, Reliabilitäten).

Im Rahmen der Item Response-Theorie wird nach einer mathematischen Relation gesucht, die den Zusammenhang zwischen Personenfähigkeit θ_v und Itemschwierigkeit δ_i im Hinblick auf das Antwortverhalten X_{vi} beschreibt (vgl. u. a. WILSON, 2004c). Dabei wird davon ausgegangen, dass θ_v und δ_i Orte auf dem Kontinuum des Personenmerkmals repräsentieren und folglich auf eine gemeinsame Skala transformiert sind. Ihre natürliche Beziehung lässt sich damit als Differenz ($\theta_v - \delta_i$) ausdrücken. Diese Differenz von Personenfähigkeit θ_v und Itemschwierigkeit δ_i lässt sich logisch auf das Antwortverhalten X_{vi} beziehen. Ist die Personenfähigkeit größer als die Itemschwierigkeit, wird angenommen, dass das Item gelöst werden kann; ein negativer Differenzwert deutet umgekehrt darauf hin, dass das Item nicht gelöst wird. Da es in der Bearbeitung von Tests dennoch dazu kommen kann, dass Personen mit einer geringen Ausprägung der Personenfähigkeit ein in Relation dazu schwieriges Item lösen können, greift die Item Response-Theorie nicht auf deterministische Relationen, sondern auf die Annahme von Wahrscheinlichkeiten zurück (probabilistische Testtheorie).

Die Wahrscheinlichkeit, dass eine Person ein Item bewältigen kann oder aber bei der Bearbeitung scheitert, ist damit ausschließlich abhängig von der Fähigkeit der Person und der Itemschwierigkeit. Bei einem positiven Differenzwert besteht demnach eine mehr als 50-prozentige Wahrscheinlichkeit, das Item lösen zu können; ein negativer Differenzwert steht für eine mehr als 50-prozentige Wahrscheinlichkeit, das Item nicht zu lösen. Entspricht die Personenfähigkeit exakt der Itemschwierigkeit, ist die Wahrscheinlichkeit hingegen genau $P = 0,5$ (vgl. ausführlich WILSON, 2005; für den Bereich der kaufmännischen Bildung WINTHER, 2010).

In beiden Testtheorien wird im Rahmen der Test- und Itemanalyse auf Itemschwierigkeiten, Itemtrennschärfen und Varianzen zurückgegriffen, um die Frage zu beantworten, wie ein Item generell über alle Personen hinweg zum gewählten Messmodell passt. Zur Überprüfung der Passung eines Items bezüglich eines gewählten Messmodells (Itemfit) werden in der Item Response-Theorie zusätzlich die gewichteten Abweichungsquadrate eines Items über alle Personen (weighted Mean Squares; wMNSQ) analysiert. Der wMNSQ ist der Standardparameter für den Itemfit; für einen akzeptablen Itemfit sollte der Wert des wMNSQ innerhalb des Intervalls $0,75 \leq \text{wMNSQ} \leq 1,33$ liegen (Adams & Khoo, 1996). Um einschätzen zu können, ob Items eine zu geringe Trennschärfe aufweisen, wird zusätzlich die T-Statistik herangezogen. Die T-Statistik prüft, ob die Abweichungen zwischen erwarteten und beobachteten Häufigkeiten statistisch signifikant sind. Da der T-Wert annähernd normalverteilt ist, lässt sich das Ergebnis des Signifikanztests anhand der Normalverteilungstabelle bestimmen. Ist der empirisch ermittelte z-Wert größer als der für das $\alpha = 5\%$ -Niveau bei zweiseitigem Test erwartete Testwert ($z = \pm 1,96$), deutet dies auf eine signifikante Abweichung in Richtung einer zu niedrigen Trennschärfe hin (vgl. u. a. Bortz, 2005). In der Folge wäre inhaltlich zu prüfen, ob das Item aus dem Test genommen werden sollte.

Die Ergebnisse der Skalen- und Itemanalyse sind in *Tabelle 3* dargestellt. Die Trennschärfekoeffizienten wurden auf Basis der Korrelation eines Items mit dem Gesamtskalenwert ermittelt (Korrelationskoeffizient nach Pearson; corrected item-total-correlation). Der Prüfungsbereich „Geschäftsprozesse“ besteht aus 38 Items, mit denen die Fähigkeit, situative Fallsituationen in den Inhaltsbereichen: Marketing

und Absatz (14 Items), Beschaffung und Bevorratung (7 Items), Personal (7 Items) sowie Leistungserstellung (10 Items), bearbeiten zu können, erfasst werden soll. In einem ersten Zugang wurden diese Inhaltsbereiche als Skalen interpretiert, die jedoch keine zufriedenstellenden internen Konsistenzen aufweisen, so dass zunächst von einer Testskala „Geschäftsprozesse“ ausgegangen wird.³ Im Mittel haben die Auszubildenden 73,38 Punkte (Min = 14, Max = 100) erreicht; die Standardabweichung beträgt 12 Punkte. Der Standardfehler des Mittelwerts ist mit 0,290 für einen Leistungstest vergleichsweise gering. Die Verteilung des Antwortverhaltens weist die Form eines rechtssteilen Normalverteilung (Schiefe = -0,683; Kurtosis = 0,613) auf. Der Test ist damit insgesamt zu leicht. 25,4 % der Items weisen eine Itemschwierigkeit von größer als 0,80 auf, und bei keinem Item liegen die Lösungsquoten unter 35 Prozent. Ein weiteres Defizit stellen die Itemtrennschärfen dar: Bei 42 Prozent der Items liegt die Trennschärfe unter 0,30 und damit außerhalb der für Leistungstests empfohlenen Untergrenze. Insgesamt weisen 18 von 38 Items unter Rückgriff auf die Klassische Testtheorie nicht die Güte auf (zu geringe Trennschärfen bzw. zu hohe Lösungsquoten), die notwendig wäre, um ausgehend von dem in den Items gezeigten Antwortverhalten auf eine durch den Test definierte Personenfähigkeit zu schließen.

Tabelle 3: Skalen- und Itemkennwerte des Prüfungsbereichs „Geschäftsprozesse“

Skala	Item-anzahl	Interne Konistenz (Cronbachs α)	Mittlere Item-trennschärfe	Mittlere Item-schwierigkeit
Geschäftsprozesse	38	.841	.326 (.12/.49)	.623 (.35/.91)
Marketing und Absatz	14	.699	.349 (.22/.49)	.650 (.35/.91)
Beschaffung und Bevorratung	7	.500	.306 (.20/.46)	.620 (.35/.87)
Personal	7	.626	.349 (.22/.45)	.600 (.40/.78)
Leistungserstellung	14	.557	.293 (.12/.46)	.599 (.40/.89)

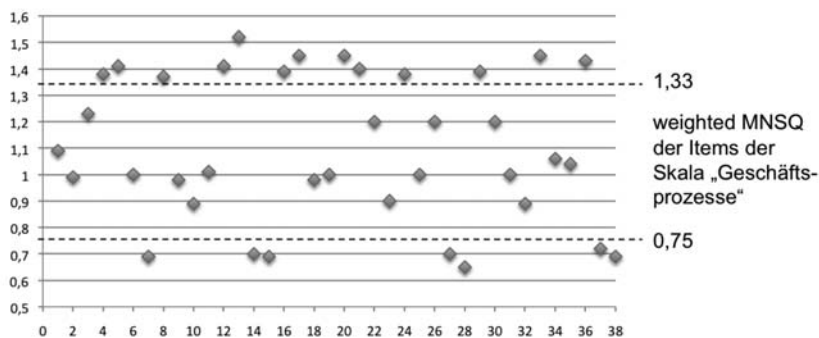
* Anmerkung: In den Klammern sind Minimum sowie Maximum (Min/Max) der Kennwerte für Item-trennschärfe und Itemschwierigkeit angegeben.

Die Analyse auf Basis der Item Response-Theorie zeigt, dass die gewichteten Abweichungsquadrate als Infit-Schätzer für den Test zwischen 0,65 und 1,52 variieren. 20 der 38 Items liegen dabei außerhalb des zulässigen Intervalls von $0,75 \leq wMNSQ \leq 1,33$ (ADAMS & KHOO, 1996); die auf Basis der T-Statistik ermittelten z-Werte liegen außerhalb der entsprechenden Vertrauensintervalle und indizieren damit eine zu niedrige Trennschärfe. Die strengere Prüfung der Items im Rahmen der probabilistischen Testtheorie ist dadurch gegeben, dass der gewählte Kennwert für den Itemfit sehr sensitiv auf Ausreißer in den Daten reagiert. Bereits dann, wenn nur sehr vereinzelt (extrem) unerwartete Antworten auf ein Item gegeben werden, wird ein Missfit des Items zum Modell angezeigt. Durch die Verwendung des weighted

3 Auch die im Prüfungskatalog der Aka (2009) vorgeschlagene Unterscheidung in Teilaspekte einer betrieblichen Handlung – Planen, Durchführen und Kontrollieren – stellt keine Lösung für eine differenzierte Skalierung des Prüfungsbereichs „Geschäftsprozesse“ dar.

MNSQ verlieren extreme Abweichungen zwischen den Parameterschätzern zwar relativ an Bedeutung, aufgehoben wird dieser Effekt dadurch allerdings nicht.

Abbildung 3: Fit-Statistik für die Itemparameter (Item Response-Theorie)

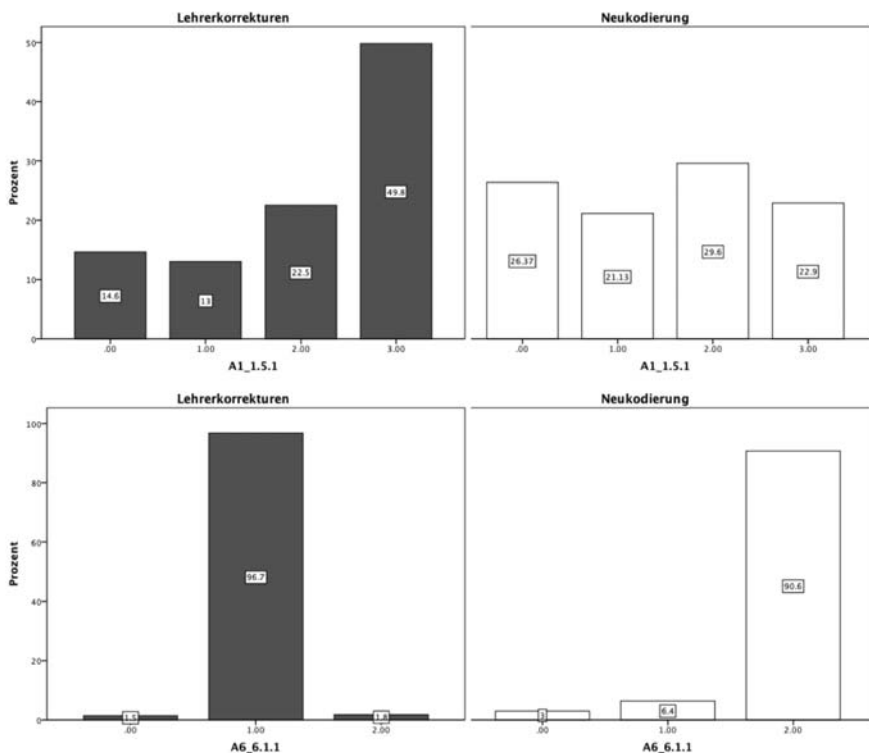


* Anmerkung: Der weighted MNSQ der Items wurde auf Basis eines eindimensionalen Partial Credit-Modells geschätzt.

Die Frage, ob die unzureichenden Passungen einer großen Zahl von Testitems vorrangig der Itemkonstruktion, dem gewählten Sampling der Items oder den Korrekturen, und damit der Codierung der Items, geschuldet sind, ist bislang nicht geklärt. Es liegen allerdings erste Befunde vor, die aufzeigen, dass bei Neukodierungen der Items die Häufigkeitsverteilungen des Antwortverhaltens in einem Item i erheblich von den Verteilungen auf Basis der Korrekturen der Lehrenden abweichen.⁴ Hierbei lassen sich zwei Muster identifizieren (vgl. *Abbildung 4*): Zum einen führt die Neukodierung des Antwortverhaltens dazu, dass die Antwortkategorien eines Items gleichmäßiger besetzt sind (Häufigkeiten der Antwortkategorien der Aufgabe A1_1.5.1 in *Abbildung 4*). Damit steigt in der Terminologie der Item Response-Theorie die Wahrscheinlichkeit, dass eine Person v mit einer spezifischen Personenfähigkeit θ_v in einer der Leistungsfähigkeit angepassten Antwortkategorie antwortet – dies hat positive Konsequenzen für den Itemfit. Hierhinter steht die Prämisse, dass das Antwortverhalten einer Person v ausschließlich durch die Itemschwierigkeit bzw. bei ordinalen Items durch die Schwellenschwierigkeiten eines Items und die Personenfähigkeit bestimmt wird. Unter Rückgriff auf diese Prämisse wird ein komplettes Verhaltensmodell formuliert, bei dem die Aussagen über die Personen, die den Test bearbeitet haben, und über die Items dieses Tests nur von den Merkmalen der Personen und Items abhängen, nicht jedoch von den Methoden, mit denen diese Merkmale gemessen werden (vgl. u. a. EMBRETSON & REISE, 2001, p. 49).

- 4 Hierfür wurden Neukodierungen und damit verbunden Neukorrekturen aller vorliegenden Prüfungshefte notwendig - insbesondere da die von der AKA gegebenen Lösungshinweise als Lösungsskizzen interpretiert wurden. Die Neukodierungen greifen auf das vorliegende Datenmaterial zurück und prüfen insbesondere, welche Lösungsqualität die formulierten Aufgaben überhaupt zulassen. Bislang sind vier der sechs IHK-Bezirke entsprechend ausgewertet ($n=1.153$).

Abbildung 4: Vergleich der Kodierungen des Antwortverhaltens



Zum anderen führen die Neukodierungen des Antwortverhaltens dazu, dass bisherige Antwortkategorien, die auf Basis des vorliegenden Datenmaterials kaum besetzt sind, durch inhaltlich andere Antwortkategorien ersetzt werden. Für die Aufgabe A6_6.1.1 (Abbildung 4) ist auffällig, dass 97 Prozent der Lehrenden die Leistung der Auszubildenden auf der mittleren Antwortkategorie verorteten und die niedrigste sowie die höchste Antwortkategorie faktisch unbesetzt sind. Ein Unterschied zwischen den Lösungsqualitäten der Probanden (1,8 Prozent), die diese Aufgabe auf der höchsten Antwortkategorie beantwortet haben, und der Mehrzahl derjenigen, die der mittleren Kategorie zugeordnet wurden, ließ sich in der Re-Analyse des Tests allerdings nicht feststellen. Als Ergebnis dieses inhaltlichen Abgleichs wird die mittlere Kategorie der Kodierung der Lehrenden als höchste Kategorie im Rahmen der Neukodierung interpretiert – am Beispiel der Aufgabe A6_6.1.1 wird in diesem Zusammenhang deutlich, dass die Aufgabenformulierung und die zur Aufgabe gegebenen Lösungshinweise nicht kongruent sind. Oder anders formuliert: Für das in den Lösungshinweisen eingeforderte Antwortverhalten sind in der Aufgabenformulierung keine hinreichenden Stimuli verankert.⁵

5 Für die Neukodierungen wurde insbesondere darauf geachtet, den durch die Aufgabenstellung möglichen Lösungsraum auf Basis des vorliegenden Datenmaterials zu erfassen – dies umfasst die Aufgabenmerkmale und das resultierende Verhalten.

3.3 Konstruktvalidität des Prüfungsbereichs „Geschäftsprozesse“

Gute Messwerte basieren auf mehreren Beobachtungen. Für kompetenzorientierte Assessments bedeutet dies, dass für jede zu messende Teilkompetenz eine Menge hinreichend ähnlicher realer Situationen zu konstruieren sind, in denen bestimmte, ähnliche Anforderungen bewältigt werden müssen, die die zu erfassende Kompetenz auf einem Kontinuum abbilden (Definition des Kompetenzkonstrukts). Die zentrale Herausforderung ist, gemeinsame ähnliche Anforderungen zu beschreiben. Hierfür ist eine hinreichend genaue Vorstellung darüber zu entwickeln, welche Prozesse beim Bearbeiten und Lösen der Anforderungssituationen ablaufen, d. h. welche betrieblichen Handlungen in spezifischen Anforderungen relevant werden.

Werden Assessments von vornherein als Kompetenzmessungen gestaltet, stehen für die Beschreibung von Anforderungssituationen sowie folgend für die Konstruktion von adäquaten Testitems Modelle des zu erfassenden Kompetenzkonstrukts zur Verfügung (vgl. hierzu insbesondere die Ergebnisse der VET-LSA-Studie für den kaufmännischen Bereich; WINTHER & ACHTENHAGEN 2009; 2010; WINTHER, 2010). Diese Modelle verbinden die inhaltlichen Zielvorstellungen einer spezifischen Domäne (Domänenmodell) mit kognitionspsychologischen Annahmen über die Bewältigung unterschiedlich komplexer Situationen in den gewählten Handlungsfeldern der Domäne (Kompetenzmodell). Für den Prüfungsbereich „Geschäftsprozesse“ sind sowohl die Domäne als auch die zu erfassende Kompetenz über den Prüfungskatalog für die IHK-Abschlussprüfungen definiert. In den Abschlussprüfungen ist die Befähigung zur beruflichen Handlungskompetenz durch das Bearbeiten komplexer praxisbezogener Aufgaben, die sich auf Ausschnitte betrieblicher Handlungen – Planen, Durchführen, Kontrollieren – beziehen, nachzuweisen. Diese Ausschnitte betrieblicher Handlungen werden dabei allgemein als Umschreibung – und damit auch als Operationalisierung – von beruflicher Handlungskompetenz verstanden (Aka, 2009, S. 3). Als Konsequenz müssten die Teilaufgaben des Prüfungsbereichs „Geschäftsprozesse“ betriebliche Tätigkeiten abbilden, die sich auf das selbständige Planen, Durchführen und Kontrollieren von spezifischen Geschäftsvorfällen beziehen. Eine solche Systematisierung lässt sich – insbesondere bei dem explorativen Charakter der hier vorliegenden Re-Analyse kaufmännischer Prüfungen – empirisch auf sehr einfache Weise durch die Analyse der Inter-Item-Korrelation (Pearsons product-moment correlation coefficient) sowie eine faktorenanalytische Zerlegung der entsprechenden Inter-Item-Korrelations-Matrix prüfen. Die bislang im Bereich der beruflichen Kompetenzmessung vorliegenden Befunde ließen einerseits eine Struktur der zu erfassenden Kompetenz erwarten, die die unterschiedlichen Bearbeitungsheuristiken und -zugänge der Probanden in den einzelnen Aufgaben abbildet. Für den Prüfungsbereich „Geschäftsprozesse“ wären dies die drei Teilelemente einer betrieblichen Handlung: Planen, Durchführen, Kontrollieren. Es kann davon ausgegangen werden, dass diese Handlungselemente eine Anforderungssituation typischerweise charakterisieren und dass folglich zur Bewältigung dieser Anforderungssituationen verschiedene Fähigkeitsstrukturen benötigt werden. Andererseits verweisen die empirischen Befunde auf die besondere Relevanz von betrieblichen Inhalten, deren charakteristischer Einfluss auf die Anforderungssituationen – beispielsweise durch vorgefertigte Lösungsalgorithmen oder zu verwendende Hilfsmittel – das Antwortverhalten der Probanden determiniert (vgl. zu beiden Strukturüberlegungen beruflicher Kompetenz insbesondere WINTHER

& ACHTENHAGEN, 2009; SEEGER, 2008; NICKOLAUS, GSCHWENDTNER & ABELE, 2009). Für den Prüfungsbereich „Geschäftsprozesse“ liegen mit der Unterscheidung zwischen Marketing und Absatz, Beschaffung und Bevorratung, Personal sowie Leistungserstellung vier Inhaltsbereiche vor, die sich im Hinblick auf das Antwortverhalten der Probanden voneinander unterscheiden könnten. In den empirischen Re-Analysen des Prüfungsbereichs „Geschäftsprozesse“ lässt sich jedoch eine entsprechende Struktur weder im Hinblick auf Handlungs- noch auf Inhaltsfacetten sichern. Dies gilt auch für eine Kombination beider Strukturüberlegungen. In *Tabelle 4* sind Ausschnitte der Inter-Item-Korrelations-Matrix zum einen für das Handlungselement Planen und zum anderen für den Inhaltsbereich Leistungserstellung dargestellt.

Tabelle 4: Inter-Item-Korrelations-Matrix (systematisierte Ausschnitte)

		Korrelationen			
		A1_1.3.2	A1_1.5.1	A4_4.2.1	A6_6.1.1
Items, die auf das Handlungselement Planen abstellen	A1_1.3.2	...	...	...	...
	A1_1.5.1	.13***	...	...	...
	A4_4.2.1	.12***	.03	...	...
	A6_6.1.1	.07**	.15***	.09**	...
		Korrelationen			
		A6_6.2.2	A6_6.3.1	A6_6.3.2	A6_6.3.3
Items des Inhaltsbereichs Leistungserstellung	A6_6.2.2	...	...	...	...
	A6_6.3.1	.05*	...	...	...
	A6_6.3.2	.05*	.10***	...	...
	A6_6.3.3	.03	.08**	.04	...

* Anmerkung: Produkt-Moment-Korrelation nach Pearson; ***p < 0,01; **p < 0,05; p < 0,10

Die Korrelationskoeffizienten für beide Strukturüberlegungen sind sehr gering bis moderat. Eine explorative faktorenanalytische Prüfung der Inter-Item-Korrelationsmatrix⁶ extrahiert eine Lösung mit 16 Faktoren, von denen nur vier Faktoren überhaupt substantielle Faktorladungen aufweisen und inhaltlich adäquat zu interpretieren wären. Eine sinnhafte Zusammenstellung von Testaufgaben, mit denen ein spezifisches Kontinuum eines Kompetenzkonstrukts abgebildet werden könnte, zeigt sich im Prüfungsbereich „Geschäftsprozesse“ auf Basis der gegebenen Testaufgaben und des durch die Lehrenden kodierten Antwortverhaltens der Auszubildenden nicht. Unter Rückgriff auf die bislang berichteten Befunde muss daher geschlossen werden, dass die Prüfung im Sinne der Verordnung über die Berufsausbildung zum Industriekaufmann/zur Industriekauffrau (Anlage 1 zu § 5 AO sowie § 2 Abs. 2, 2002) sowie unter Rückgriff auf den Prüfungskatalog für die IHK-Abschlussprüfungen nicht konstruktvalid ist. Vertiefende Analysen zeigen darüber hinaus, dass selbst

6 Die extrahierte Faktorenlösung erklärt 65,82 Prozent Gesamtvarianz (Hauptkomponentenmethode, Varimax-Rotation mit Kaiser-Normalisierung; Eigenwerte größer 1; Rotation ist in 34 Iterationen konvergiert, Kaiser-Meyer-Olkin-Kriterium der Stichprobeneignung: 0,853).

bei Items mit moderatem Zusammenhang systematische Verzerrungen im Hinblick auf exogene Variablen (wie Betriebsgröße, Geschlecht der Prüflinge und bisheriger Schulabschluss) auftreten. Dies hat zur Folge, dass bezogen auf spezifische Kriterien selbst Items, die inhaltlich ein ähnliches Konstrukt erfassen, keine konsistente Beziehung aufweisen. In *Tabelle 5* sind einzelne Inter-Item-Korrelationen sowie Korrelationen der Items mit verschiedenen Hintergrundvariablen am Beispiel eines Sets aus vier Items abgebildet. Die Hintergrundvariablen sind insbesondere zur Klärung exogener Effekte auf das Testergebnis sowie im Zusammenhang der Prüfung von Testfairness relevant. Sie können jedoch auch dazu genutzt werden, interpretierbare Muster innerhalb eines Tests zu identifizieren.

Tabelle 5: Inter-Item-Korrelation sowie Korrelation mit Hintergrundvariablen

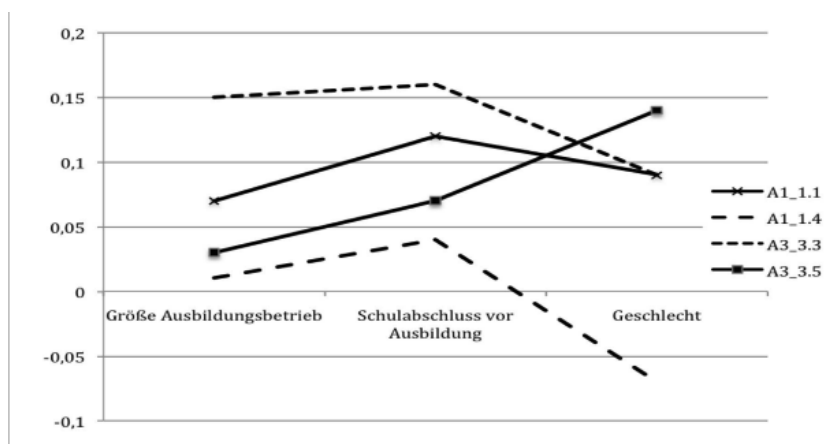
	Inter-Item-Korrelation				Korrelation Hintergrundvariable		
	A1_1.1	A1_1.4	A3_3.3	A3_3.5	Größe Ausbil- dungs- betrieb	Schulab- schluss vor Aus- bildung	Ge- schlecht
A1_1.1	...	...	...	...	.07**	.12***	.09**
A1_1.4	.22***	...	...	...	.01	.04*	-.07**
A3_3.3	.30***	.21***	...	...	.15***	.16***	.09**
A3_3.5	.17***	.17***	.20***	...	.03	.07**	.14***

* Anmerkung: Produkt-Moment-Korrelation nach Pearson; *** $p < 0,01$; ** $p < 0,05$; $p < 0,10$; negative Korrelation beim Geschlecht deutet auf Lösungsvorteile für weibliche Auszubildende hin.

Die Korrelationen der einzelnen Testitems mit den Hintergrundvariablen sind sehr unterschiedlich; die Einflüsse des Geschlechts, der Größe des Ausbildungsbetriebes sowie des vor der Ausbildung erworbenen Schulabschlusses auf das Antwortverhalten variieren zwischen den gewählten Beispielitems deutlich. In *Abbildung 5* sind die einzelnen Korrelationsprofile der Testitems im Vergleich geplottet. Die Abbildung veranschaulicht, dass sich das Profil des Testitems A3_3.5 durch seinen ansteigenden Verlauf und das Profil des Testitems A1_1.4 durch die negative Korrelation mit dem Merkmal Geschlecht von den Profilen der anderen beiden Testitems unterscheiden. Dies lässt den Schluss zu, dass die Beispielitems A3_3.3 und A1_1.1 innerhalb der theoretisch definierten Domäne, die sich zusätzlich durch die gewählten exogenen Variablen spezifizieren lässt, dasselbe Konstrukt erfassen. Diese Items zeigen eine konsistente Beziehung zu den exogenen Variablen, während sich für die zwei anderen Items der Zusammenhang unspezifisch darstellt. Oder anders formuliert: Trotz der signifikanten Korrelationen der vier Items ist es nicht angezeigt, alle Items als Skala bzw. als Teil einer Skala zu kombinieren.

Der Unterschied zwischen den Korrelationsprofilen lässt sich als Proportionalitäts-Ratio (P^2) berechnen, um die Frage zu beantworten, zu welchem Grad sich zwei Korrelationsreihen affin zueinander verhalten. Hierzu wird die Summe der Kreuzprodukte der einzelnen Korrelationen zu den Summen der quadrierten Korrelationen ins Verhältnis gesetzt. Diese Messgröße zeigt an, welche Itempaare im Hinblick auf relevante Hintergrundvariablen proportionale Korrelationen aufweisen und damit

Abbildung 5: Korrelationsprofil am Beispiel von vier Testitems



in kohärenter Beziehung zueinander sowie zum zu erfassenden Konstrukt stehen (PIAZZA, 1980, p. 592):

$$P^2_{x_1x_2} = \frac{\left(\sum_k r_{x_1z_k} r_{x_2z_k} \right)^2}{\left(\sum_k r_{x_1z_k}^2 \sum_k r_{x_2z_k}^2 \right)}$$

mit $k = 1, \dots, N$; die z_k sind die N exogenen Variablen

Die Proportionalitäts-Ratio ist ein deskriptives Instrument zur Klärung der Konstruktvalidität unter Berücksichtigung exogener Variablen, von denen Effekte auf das Test- und Antwortverhalten theoriebasiert erwartet werden können. Die P^2 -Werte der vier Beispielimts sind, ergänzt um weitere P^2 -Werte, in *Tabelle 6* dargestellt. Der numerische Ausdruck festigt die graphische Analyse: Die Items A3_3.3 und A1_1.1 weisen einen hohen Proportionalitätskennwert auf ($P^2 = 0,928$)

Für die Aufgaben A1_1.1 sowie A3_3.5 ergibt sich – anders als in der graphischen Analyse vermutet – ebenfalls ein guter Proportionalitätskennwert ($P^2 = 0,767$).⁷ Hier werden die Schwächen des Messwertes deutlich: hohe Korrelationen werden stärker als kleinere Korrelationen in der Proportionalitäts-Ratio berücksichtigt, so dass schwache Korrelationen der Items mit exogenen Variablen durch starke überlagert werden. Insgesamt ist dieses Verfahren jedoch geeignet, nach theoretisch interpretierbaren Mustern in einem Test – der bislang nicht skaliert ist – zu suchen. Statt die Inter-Item-Korrelationsmatrix explorativ zu analysieren, ist es sinnvoll, auf die P^2 -Matrix zurückzugreifen. Die Verwendung explorativer Modelle wird im Rahmen

7 Für die Interpretation eines guten Proportionalitätskennwertes liegen keine Verteilungen des P^2 -Wertes vor. Die Einschätzungen sind ausschließlich deskriptiv (vgl. hierzu PIAZZA, 1980).

Tabelle 6: Matrix der P²-Werte

	P ² (Proportionalitäts-Ratio)						
	A1_1.1	A1_1.4	A3_3.3	A3_3.5	A4_4.1.1	A4_4.1.2	A6_6.3.3
A1_1.1	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
A1_1.4	.040	...	...	...	...	...	...
...	...	...	...	...	...	...	...
A3_3.3	.928	.268	...	...	...	...	...
...	...	...	...	...	...	...	...
A3_3.5	.767	.070	.561	...	...	...	...
...	...	...	...	...	...	...	...
A4_4.1.1	.803	.688	.122	.789	...	...	...
A4_4.1.2	.378	.933	.075	.445	.912	...	...
...	...	...	...	...	...	...	...
A6_6.3.3	.767	.070	.541	.091	.897	.347	...

*Anmerkung: Ein P²-Wert von 1 indiziert perfekte Proportionalität; ein P²-Wert von -1 indiziert perfekte inverse Proportionalität.

der Analyse kaufmännischer Abschlussprüfungen notwendig, da nach den Strukturen im Datensatz zunächst gesucht werden muss und im Vorfeld der Analyse der Zusammenhang zwischen den Daten nicht bekannt war. Durch die P²-Matrix wird dieser Zusammenhang durch eine Auswahl exogener Variablen definiert, wodurch ein deduktiv-logisches Referenzsystem vorgegeben wird.

4 Schlussfolgerungen und weitere Analyseschritte

Die bisherigen Ausführungen haben deutlich gemacht, dass der Prüfungsbe-
reich „Geschäftsprozesse“ der IHK-Abschlussprüfungen im Ausbildungsberuf
Industriekaufmann/-frau inhaltlich wie empirisch in hohem Maße Limitationen auf-
weist. Dabei sind insbesondere die verschiedenen Validitätskonzepte betroffen. Der
Anspruch diagnostischer Verfahren setzt jedoch insbesondere psychometrische
Validität und Vergleichbarkeit voraus und ist an eine stringente Konstruktdefinition
gebunden. Es muss Klarheit darüber bestehen, was eigentlich wie gemessen werden
soll. Mittels explorativer Verfahren lassen sich Strukturen in den vorliegenden Daten
extrahieren, die einer inhaltlichen Analyse bedürfen. Die besondere Schwierigkeit
hierbei ist, dass die ersten empirischen Befunde systematische Verzerrungen in
den Korrekturen der Aufgaben des Prüfungsbereichs vermuten lassen. Dies macht
eine vollständige Neukodierung aller Prüfungshefte notwendig. Auf Basis dieser
Neukodierungen wird erneut nach interpretativen Zusammenhängen und Mustern
gesucht werden. Hierzu wird unter Rückgriff auf die P²-Matrix ein exploratives mehr-
dimensionales Rasch-Modell (vgl. u. a. McKINLEY & RECKASE, 1982; zit. n. EMBRETSON

& REISE, 2001) mit dem Ziel der Hypothesengenerierung geschätzt. Auf Basis der so gewonnenen Ergebnisse werden Re-Modellierungen der Prüfungsaufgaben vorgenommen, die dann konfirmatorische Analysen – also Analysen der Hypothesenprüfung – zulassen. Das Ziel der Re-Modellierungen ist es, im Vorfeld eines Assessments detaillierte, kognitiv determinierte Überlegungen darüber anzustellen, wie Items zu konstruieren sind und welche kognitiven Fähigkeiten diese Items bei den Probanden ansprechen können und sollen. Diese Überlegungen umfassen folglich auch eine Abbildung möglicher Antwortmuster und die Fragen, wie Probanden auf die Items reagieren werden und wie sich diese Reaktionen in Messwerte der Personenfähigkeit transformieren lassen – Fragen die zentral die Konstruktdefinition betreffen. Die Re-Modellierungen werden daher systematische Variationen des Anspruchsniveaus beinhalten, um ein mögliches Kontinuum beruflicher Fähigkeiten so abzubilden, dass sich ein Kompetenzmodell generieren lässt. Vorbild hierfür können die Experimente von QUILICI und MAYER (1996) sein, in denen Kovariationen zwischen Oberflächenstruktur einer Aufgabe (wie ist eine Aufgabe gestellt) und Lösungsprinzipien (wie wird eine Aufgabe bearbeitet) nachgewiesen werden konnten. Die zentrale Frage hinter systematischen Variationen des Anspruchsniveaus im Rahmen kognitiver Leistungstests ist die nach den Mechanismen des Erwerbs inhalts- bzw. domänenspezifischer Wissensstrukturen (vgl. hier u. a. THUSSBAS & CHOURDAKIS, 2002) als Voraussetzung für die Beschreibung berufsfachlicher Kompetenz.

Literatur

- Achtenhagen, F. (2004). Prüfung von Leistungsindikatoren für die Berufsbildung sowie zur Ausdifferenzierung beruflicher Kompetenzprofile nach Wissensarten. In M. Baethge, K.-P. Buss & C. Lanfer (Hrsg.), *Expertisen zu den konzeptionellen Grundlagen für einen nationalen Bildungsbericht – Berufliche Bildung und Weiterbildung/Lebenslanges Lernen* (S. 11-32). Bonn, Berlin: BMBF.
- Adams, R. J. & Khoo, S. T. (1996). *Quest*. Melbourne, Australia: Australian Council for Educational Research.
- American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME) (Eds.) (1999). *Standards for Educational and Psychological Testing*. Washington, DC: National Academy Press.
- Aufgabenstelle für kaufmännische Abschluss- und Zwischenprüfungen (Aka) (Hrsg.) (2009). *Prüfungskatalog für die IHK-Abschlussprüfungen, 3. Aufl.* Nürnberg: Aka.
- Baethge, M., Achtenhagen, F., Arends, L., Babic, E., Baethge-Kinsky, V. & Weber, S. (2006). *Berufsbildungs-PISA – Machbarkeitsstudie*. Stuttgart: Steiner.
- Baethge, M. & Arends, L. (Eds.) (2009). *Feasibility Study VET-LSA. A Comparative Analysis of Occupational Profiles and VET Programmes in 8 European Countries*. International Report. Göttingen: Soziologisches Forschungsinstitut.
- Baxter, G. P. (1995). Using Computer Simulations to Assess Hands-on Science Learning. *Journal of Science Education and Technology*, 4 (1), 21–27.
- Beck, K. (2001). WBT – Wirtschaftskundlicher Bildungstest. In W. Sarges & H. Wottawa (Hrsg.), *Handbuch wirtschaftspsychologischer Testverfahren* (S. 559–562). Lengerich: Pabst Science Publishers.
- Black, P. & Wiliam, D. (2004). The Formative Purpose: Assessment Must First Promote Learning. In M. Wilson (Ed.), *Towards Coherences Between Classroom Assessment and Accountability. 103rd Yearbook of the National Society for the Study of Education* (pp. 20–50). Chicago, IL: The University of Chicago Press.
- Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler*, 6. Aufl. Heidelberg: Springer.
- Breuer, K., Hillen, S. & Winther, E. (2009). Comparative International Analysis of Occupational Tasks and Qualification Requirements for the Labour Market and Assessment Tasks at

- the End of VET in Participating Countries – Business and Administration. In M. Baethge & L. Arends (Eds.), *Feasibility Study VET-LSA. A Comparative Analysis of Occupational Profiles and VET Programmes in 8 European Countries*. International Report. Göttingen: Soziologisches Forschungsinstitut.
- Cheng, L., Watanabe, Y. & Curtis, A. (Eds.). (2004). *Washback in language testing. Research contexts and methods*. Mahwah: Lawrence Erlbaum.
- Connel, M. W., Sheridan, K. & Gardner, H. (2003). On Abilities and Domains. In R. J. Sternberg & E. L. Grigorenko (Eds.), *The Psychology of Abilities, Competencies, and Expertise* (pp. 126–155). Cambridge: Cambridge University Press.
- Department of Education and Science (Ed.) (1987). *National Curriculum Task Group on Assessment and Testing: A Report*. London: HMSO.
- Embretson, S. E. (1999). Issues in the Measurement of Cognitive Ability. In S. E. Embretson & S. L. Hershberger (Eds.), *The New Rules of Measurement* (pp. 1–15). Mahwah, NJ: Lawrence Erlbaum.
- Embretson, S. E. & Reise, S. P. (2001). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Fuhrman, S. H. & Elmore, R. F. (Eds.) (2004). *Redesigning accountability systems for education*. New York: Teachers College Press.
- Grollmann, P. & Spöttl, G. (2008). Berufsbildungs-PISA ist überstürzt und in der politischen Wirkung schädlich. WAP.
- Gustafsson, J.-E. & Undheim, J. O. (1996). Individual Differences in Cognitive Functions. In D. Berliner & R. Calfee (Eds.), *Handbook of Educational Psychology* (pp. 186–242). New York, NY: Macmillan.
- Hartig, J., Klieme, E. & Leutner, D. (2008). *Assessment of Competencies in Educational Contexts*. Göttingen: Hogrefe & Huber.
- IHK Aachen (Hrsg.) (2008). *Prüfungsbereiche der IHK Abschlussprüfungen für den Ausbildungsberuf Industriekaufmann/Industriekauffrau*. Aachen: IHK.
- Klieme, E. & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. *Zeitschrift für Pädagogische Psychologie*, 20, 137–138.
- KMK (Hrsg.) (2002). *Rahmenlehrplan für den Ausbildungsberuf Industriekaufmann/Industriekauffrau, Beschluss vom 14.06.2002*. Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland: Bonn.
- KMK (Hrsg.) (2003). *Bildungsstandards im Fach Mathematik für den Mittleren Bildungsabschluss, Beschluss vom 04.12.2003*. Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland: Bonn.
- Mislevy, R. J. (1994). Evidence and Inference in Educational Assessment. *Psychometrika*, 59, 439–483.
- Mislevy, R. J. & Haertel, G. D. (2006). Implications of Evidence-Centered Design for Educational Testing. *Educational Measurement: Issues and Practice*, 25 (4), 6–20.
- Moss, P. A., Pullin, D. C., Gee, J. P., Haertel, E. H. & Young, L. J. (Eds.) (2008). *Assessment, Equity, and Opportunity to Learn*. Cambridge, MA: Cambridge University Press.
- Musekamp, F., Spöttl, G. & Becker, M. (2010). Schriftliche Arbeitsaufträge zur Erfassung von Differenzen in der Expertise von Facharbeitern und Auszubildenden. *Zeitschrift für Berufs- und Wirtschaftspädagogik*, 106 (3), 336–360.
- National Council on Economic Education (Ed.) (1964). *Test of economic understanding*. New York, NY: National Council on Economic Education (NCEE).
- Nickolaus, R., Gschwendtner, T. & Abele, S. (2009). *Die Validität von Simulationsaufgaben am Beispiel der Diagnosekompetenz von Kfz-Mechatronikern. Vorstudie zur Validität von Simulationsaufgaben im Rahmen eines VET-LSA*. Abschlussbericht für das Bundesministerium für Bildung und Forschung: Universität Stuttgart.
- Nunnally I. C. (1978). *Psychometric theory*. New York: McGraw-Hill Book Company.
- Pellegrino, J., Chudowsky, N. & Glaser, R. (Eds.) (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. National Research Council (NRC) Committee on the Foundations of Assessment. Washington, DC: National Academy Press.
- Piazza, T. (1980). The analysis of attitude items. *American Journal of Sociology*, 86 (3), 584–603.

- Quilici, J. L. & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology*, 88, 144–161.
- Roth, H. (1971). *Pädagogische Anthropologie. Bd. 2: Entwicklung und Erziehung. Grundlagen einer Entwicklungspädagogik*. Hannover: Schroedel.
- Reetz, L. & Hewlett, C. (2009). *Das Prüferhandbuch. Eine Handreichung zur Prüfungspraxis in der beruflichen Bildung*. Hamburg: Ver.di.
- Scheerens, J., Glas, C. & Thomas, S. M. (2003). *Educational evaluation, assessment, and monitoring. A systematic approach*. Lisse: Swets and Zeitlinger.
- Seeber, S. (2008). Ansätze zur Modellierung beruflicher Fachkompetenz in kaufmännischen Ausbildungsberufen. *Zeitschrift für Berufs- und Wirtschaftspädagogik*, 104, 74–97.
- Shavelson, R. J., Baxter, G. P. & Pine, J. (1991). Performance Assessment in Science. *Applied Measurement in Education*, 4, 347–362.
- Shavelson, R. J., Ruiz-Primo, M. A. & Wiley, E. W. (1999). Note on Sources of Sampling Variability in Science Performance Assessments. *Journal of Educational Measurement*, 36, 61–71.
- Soper, J. C. & Walstad, W. B. (1987). *The test of economics literacy* (2nd Ed.): Examiner's manual. New York, NY: National Council on Economic Education (NCEE).
- Straka, G. A. & Macke, G. (2010). Kompetenz – nur eine "kontextspezifische cognitive Leistungsdisposition"? *Zeitschrift für Berufs- und Wirtschaftspädagogik*, 106 (3), 444–451.
- Thußbas, C. & Chourdakis, D. (2002). Wie unterschiedlich sollten Beispiele sein? *Zeitschrift für Pädagogische Psychologie*, 16 (2), 117–123.
- Walstad, W. B. & Rebeck, K. (2001). Assessing the Economic Understanding of U.S. High School Students. *American Economic Review*, 91 (2), pp. 452–457.
- White, B. & Frederiksen, J. (1998). Inquiry, Modeling, and Metacognition: Making Science Accessible to All Students. *Cognition and Instruction*, 16 (1), 3–118.
- Wilson, M. (Ed.) (2004a). *Towards Coherences Between Classroom Assessment and Accountability. 103rd Yearbook of the National Society for the Study of Education*. Chicago, IL: The University of Chicago Press.
- Wilson, M. (2004b). Assessment, Accountability and the Classroom: A Community of Judgement. In M. Wilson (Ed.), *Towards Coherences Between Classroom Assessment and Accountability. 103rd Yearbook of the National Society for the Study of Education* (pp. 1–19). Chicago, IL: The University of Chicago Press.
- Wilson, M. (2004c). On Choosing a Model for Measuring. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch-Measurement* (pp. 11–42). Maple Grove, MN: JAM Press.
- Wilson, M. (2005). *Constructing Measures: An Item-response Modeling Approach*. Mahwah, NJ: Lawrence Erlbaum.
- Wilson, M. & Sloane, K. (2000). From Principles to Practice: An Embedded Assessment System. *Applied Measurement in Education*, 13 (2), 181–208.
- Winther, E. (2010). *Kompetenzmessung in der beruflichen Bildung*. Bielefeld: wbv.
- Winther, E. & Achtenhagen, F. (2009). Skalen und Stufen kaufmännischer Kompetenz. *Zeitschrift für Berufs- und Wirtschaftspädagogik*, 105 (4), 521–556.
- Winther, E. & Achtenhagen, F. (2010). Berufsfachliche Kompetenz: Messinstrumente und empirische Befunde zur Mehrdimensionalität beruflicher Handlungskompetenz. *Berufsbildung in Wissenschaft und Praxis*, 1/2010, 18–21.
- Wolf, D. P. (1989). Portfolio Assessment: Sampling Student Work. *Educational Leadership*, 46, 35–39.

Anschrift der Autorin: Prof. Dr. Esther Winther, Professur für Wirtschaftspädagogik, Universität Paderborn, Warburgerstr. 100, 33098 Paderborn (esther.winther@wiwi.upb.de)