Susan Seeber

Ansätze zur Modellierung beruflicher Fachkompetenz in kaufmännischen Ausbildungsberufen

Kurzfassung: Im vorliegenden Beitrag werden Möglichkeiten diskutiert, berufliche Fachkompetenzen im Ausbildungsberuf Bürokaufmann/Bürokauffrau zu messen. Im Vergleich mehrerer formaler Modelle wird die zu Grunde liegende psychometrische Struktur des eingesetzten Tests erörtert. Die Ergebnisse der Prüfung auf Modellgeltung und die Dimensionalität beruflicher Fachkompetenz verweisen auf domänenspezifische Verständnisfaktoren. Darüber hinaus werden auf der Grundlage kognitionspsychologischer Überlegungen Ansätze zur Identifikation von Anforderungsmerkmalen der Testaufgaben und empirische Verfahren der Bestimmung von Kompetenzniveaus zur Diskussion gestellt. Über die so ermöglichte kriteriumsorientierte Interpretation der Testleistungen wird das diagnostische Potenzial der Kompetenzmessung aufgezeigt. In diesem Zusammenhang werden auch Befunde, die auf Diskrepanzen zwischen den curricularen Ansprüchen und den erreichten beruflichen Fachleistungen gegen Ende der beruflichen Ausbildung verweisen, vorgestellt.

Abstract: The article presents several approaches to assessing vocational competencies in a training programme for future employees in business administration ("Bürokaufmann"). In a comparative fashion, different models are used to ascertain the psychometric structure of the test employed. Findings from statistical analyses of model fit and dimensionality, applied to the test given, suggest the existence of manifold content-specific cognitive competencies in this particular domain. Moreover, tenets of cognitive psychology are used to define of factors underlying the difficulty of items and evidence-based methods of determining distinct competency levels. This facilitates criterion-referenced interpretations of test results, implying a considerable diagnostic potential. In this context, findings are discussed which demonstrate discrepancies between the intended and the achieved curriculum towards the end of vocational education and training.

1. Einleitung

Der Beitrag hat das Ziel, Möglichkeiten der Überprüfung und damit ggfs. der Optimierung der Anforderungsstruktur berufsbezogener Fachleistungstests vergleichend zu diskutieren. Im Rahmen der Längs- und Querschnitts-Studien „**U**ntersuchung von **L**ernständen, **M**otivation und **E**instellungen Hamburger Schülerinnen und Schülern an Berufsschulen sowie teil- und vollqualifizierenden Berufsfachschulen" (ULME I-III) liegen Erfahrungen zur Erfassung von berufsbezogenen, berufsspezifischen und berufsübergreifenden Kompetenzen auf der Grundlage interner Regulationsbedingungen vor. In ULME III wurden dabei berufliche Fachleistungstests für insgesamt 17 Berufe, davon sieben im Berufsfeld Wirtschaft und Verwaltung, entwickelt und erprobt.

Die auf der Basis der probabilistischen Testtheorie (vgl. Fischer & Molenaar, 1995) skalierten beruflichen Fachleistungstests umfassen jeweils ein breites berufliches Anforderungsspektrum, das sowohl Niveaueinstufungen von Aufgaben auf der Grundlage taxonomischer Überlegungen (vgl. dazu Brand, Hofmeister & Tramm, 2005) als auch Analysen zu den Dimensionen der erfassten berufsspezifischen Fähigkeiten erlaubt. Für die Untersuchung der gegebenen Anforderungs-

strukturen der beruflichen Fachleistungstests stehen darüber hinaus verschiedene fachübergreifende Tests zur Verfügung. Zudem liegen in den meisten Berufen auch Erkenntnisse über die allgemeinen Lernstände zu Beginn der Ausbildung vor.

Während über Kompetenzstrukturmodelle die Zusammenhänge zwischen spezifischen Kompetenzen und Teilkompetenzen abgebildet werden können, befassen sich Kompetenzstufenmodelle mit der kriteriumsorientierten Testinterpretation, indem Abschnitte auf Kompetenzskalen hinsichtlich komplexer Anforderungen, die Personen auf diesen Niveaus bewältigen können, beschrieben werden. Für die Definition entsprechender Kompetenzniveaus („proficiency scaling") finden in der empirischen Bildungsforschung vor allem zwei auf unterschiedlichen methodischen Herangehensweisen beruhende Verfahren Anwendung: die Bestimmung kritischer Schwellen nach Beaton & Allen (1992) und die Prädiktion von Itemschwierigkeiten unter Rückgriff auf Aufgabenmerkmale (Hartig, 2007).

Im vorliegenden Beitrag werden für den beruflichen Fachleistungstest des Bürokaufmanns/der Bürokauffrau Ergebnisse der Dimensionalitätsprüfung und Verfahren der Kompetenzmodellierung exemplarisch diskutiert sowie zentrale Befunde und Determinanten beruflicher Fachleistungen herausgearbeitet.

2. Teststruktur und Datenbasis

Der berufliche Fachleistungstest für den Ausbildungsberuf Bürokaufmann/Bürokauffrau bezieht sich auf die Erfassung kontextspezifischer kognitiver Leistungsdispositionen, wobei unter Kontext die spezifischen Anforderungen und Situationen der Fachausbildung an beruflichen Schulen im jeweiligen Ausbildungsberuf verstanden werden (zur Kontextabhängigkeit von Kompetenzen vgl. Klieme, Maag-Merki & Hartig, 2007, 7). Die 51 Testaufgaben, bestehend aus 112 Einzelitems, wurden im Multiple-Choice-Format, als Einfachwahl- und in Form von Zuordnungsaufgaben sowie als Aufgaben mit offenem Antwortformat konzipiert. Im Interesse einer effizienten Nutzung der verfügbaren Testzeit von 90 Minuten für den beruflichen Fachleistungstest (zum Gesamtkonzept der Untersuchung vgl. Lehmann & Seeber, 2007; zu den Testinstrumenten Lehmann & Hunger, 2007) wurden, um eine differenzierte Erfassung breiter berufsfachlicher Anforderungen zu gewährleisten, überwiegend Aufgaben mit geschlossenem Format konzipiert. Aufgaben mit divergenten Anforderungen, für deren Lösung konstruktive und argumentative Schritte erforderlich sind, die gerade in der kaufmännischen Ausbildung eine zentrale Rolle spielen, konnten angesichts der Rahmenbedingungen der Gesamtstudie hier nicht berücksichtigt werden.

Mit dem Einsatz von Testverfahren verbindet sich, im Unterschied zu Klassenarbeiten, der Anspruch psychometrisch fundierte Aussagen über latente Fähigkeiten zu machen; es geht also nicht allein darum, wie gut oder wie schlecht jemand eine konkrete Aufgabe oder einen einzigen Typ von Aufgaben lösen kann (vgl. Klieme et al. 2000, 108). Die Beurteilung von Kompetenzen innerhalb einer Domäne oder eines Fachgebiets erfordert vielmehr eine ausreichende Breite von Anforderungen, Inhalten und Antwortformaten. Die Aufgaben selbst stellen Indikatoren einer umfassenden Fähigkeit dar, im vorliegenden Falle die ökonomische Fachkompetenz von angehenden Bürokaufleuten. Ziel der Testentwicklung war es somit, möglichst über die gesamte Breite berufsfachlicher Anforderungen des entsprechenden Ausbildungsberufs Aufgabenstellungen zu entwickeln, die sich sowohl auf unter-

schiedliche Inhaltsbereiche beziehen, als auch in verschiedenen situativen Kontexten verankert sind (zum Situationsprinzip vgl. Reetz, 1984). Gleichzeitig sollten die Aufgaben differenzierte Analysen zur Dimensionalität, zum Anforderungsniveau und zur hierarchischen Struktur der Fachkompetenz ermöglichen, womit sich hohe Ansprüche an Testdesign und -struktur verbinden. Neben der beruflichen Relevanz spielte bereits bei der Testkonstruktion vor allem die curriculare Validität der Aufgabensätze eine erhebliche Rolle (vgl. Brand, Hofmeister & Tramm, 2005). Durch die Fachdidaktik-Experten des Instituts für Berufs- und Wirtschaftspädagogik der Universität Hamburg erfolgte vorab eine Sichtung der vorliegenden Rahmenlehrpläne und der Hamburger Bildungspläne für den jeweiligen Ausbildungsberuf, die nach den Vorgaben des entsprechenden Modellversuchs für den hier untersuchten Beruf und Ausbildungsjahrgang in Hamburg bereits lernfeld- und kompetenzorientiert formuliert waren. Weiterhin wurden in diese Dokumentenanalysen auch die in Hamburg eingesetzten Lehrbücher und Aufgaben der Kammerprüfungen der letzten Jahre einbezogen, um den expliziten wie auch impliziten curricularen Anforderungen Rechnung zu tragen (zur Unterscheidung von intendierten und implementierten Curriculum vgl. Baumert, Lehmann et al., 1997). Darüber hinaus waren Lehrende aus Hamburger Berufsschulen maßgeblich in die Itementwicklung und in die Beurteilung der curricularen Validität des Tests einbezogen. Auch im Rahmen einer Pilotierung an Berliner Berufsschulen wurden die Lehrenden gebeten, die einzelnen Testaufgaben hinsichtlich ihrer curricularen und berufsbezogenen Relevanz einzuschätzen. Freilich konnte – auch angesichts der zeitlichen Rahmenbedingungen für die Testentwicklung und -pilotierung – mit diesen Maßnahmen nur ansatzweise die curriculare Validität überprüft werden. Es steht aber außer Frage, dass in Folgestudien dem Aspekt der curricularen und berufsbezogenen Validität auch mit anderen Verfahren, beispielsweise in Anlehnung an die Angoff-Methode (vgl. dazu Whetton, Twist und Sainsbury, 1999), nachgegangen werden muss.

Der von den Fachdidaktik-Experten für die Itementwicklung zugrunde gelegte Klassifikationsrahmen unterscheidet – in Anlehnung an das kognitionspsychologische Modell von Anderson & Krathwohl (2001) – Wissens- und Verhaltensdimensionen. Die dort entwickelte Matrix differenziert bei den kognitiven Strukturen zwischen vier Wissens- und sechs Verhaltensdimensionen. Brand, Hofmeister & Tramm (2005) revidierten diese und gingen von drei Wissenskategorien (Faktenwissen, Konzeptwissen und prozedurales Wissen) mit je zwei Unterkategorien aus. In Bezug auf die kognitiven Dimensionen lehnten sie sich an einem Vorschlag von Metzger et al. (1993) an, der zwischen Reproduzieren, Verstehen/Anwenden und Kritisieren/Reflektieren unterscheidet. Metzger et al. (1993) unterstellen bei diesen drei Verhaltenskategorien eine zunehmende Komplexität, die angesichts der vorliegenden empirischen Befunde nicht ganz unproblematisch erscheint. So sprechen die Daten aus den internationalen Schulleistungsuntersuchungen wie TIMSS und PISA dagegen, Verhaltenserwartungen als klar unterscheidbare Kompetenzdimensionen anzusehen. Daher wurde bei TIMSS der Ansatz hierarchisch geordneter kognitiver Operationen zu Gunsten eines kategorialen Rasters typischer Verhaltenserwartungen bei der Lösung von Testaufgaben aufgegeben (für TIMSS vgl. Klieme et al., 2000, 119). Zu ähnlichen Schlüssen gelangte Witt (2006, 407ff.) in Untersuchungen zur ökonomischen Bildung mit dem Wirtschaftskundlichen Bildungs-Test (vgl. Beck & Krumm, 1998) und schlug vor, die an Bloom angelehnten Taxonomiestufen als kategoriale Niveaus, nicht jedoch als hierarchische Ordnung zu

betrachten. Die Ergebnisse aus den Itemanalysen verschiedener berufsbezogener oder berufsspezifischer Tests im Rahmen der Hamburger Studien an teilqualifizierenden Berufsfachschulen und in der dualen Ausbildung erhärten die Vermutung, dass die der kognitiven Dimension zugrunde gelegten Klassifikationsmerkmale auch für berufliche Fachleistungen als weitgehend schwierigkeitsinvariant anzusehen sind (vgl. Seeber, 2005; Seeber, 2007b, 190). Auf den Beitrag der unterschiedenen Wissensarten zur Erklärung der Aufgabenschwierigkeit und damit zur Modellierung von Kompetenzniveaus wird nachfolgend noch näher eingegangen.

Trotz dieser Einschränkungen stellte der zugrunde gelegte Klassifikationsrahmen einen wichtigen Ausgangspunkt für die Konstruktion der Testaufgaben dar, und zwar vor allem, um unausgewogene Verteilungen der Items, z.B. zugunsten des Faktenwissens und des Reproduzierens von Inhalten zu vermeiden (vgl. Brand, Hofmeister & Tramm, 2005; 17ff). Dieses Ziel konnte bei der Mehrzahl der eingesetzten beruflichen Fachleistungstests auch erreicht werden.

Hinsichtlich der angesprochenen Wissensstrukturen dominierten beim Test für den Bürokaufmann Aufgaben, die konzeptuelles Wissen und algorithmisch-prozedurales Wissen erforderten. In Bezug auf die Verhaltenserwartungen enthielt der Test überwiegend Items, die auf Wissensanwendung und Verstehen zielten. Während positiv hervorzuheben ist, dass nur in geringem Umfang Aufgaben mit reproduktivem Charakter vertreten waren, sollten bei Testüberarbeitungen und Weiterentwicklungen auch Aufgaben einbezogen werden, die eine kritische Auseinandersetzung und Reflexion mit ökonomischen Inhalten und Gegenständen erfordern.

Insgesamt handelte es sich um einen anspruchsvollen, die inhaltliche Breite des Ausbildungsberufs gut repräsentierenden Test, der es gestattete, fachliche Kompetenzprofile der Jugendlichen herauszuarbeiten und so spezifische Stärken und Schwächen darzustellen.

Aufgrund der hohen Testgüte der Pilotversion konnten die nun nachfolgend diskutierten Testanalysen und Ansätze zur Kompetenzmodellierung nicht nur auf Basis der Daten der Hauptuntersuchung vorgenommen werden, sondern auch die Daten aus der Berliner Piloterhebung einbezogen werden. An der Pilotierung des Tests nahmen 128 Jugendliche aus sieben Klassen zwei kaufmännischer Oberstufenzentren teil; im Rahmen der Hamburger Leistungsstudie beteiligten sich am beruflichen Kompetenztest für den Ausbildungsberuf Bürokaufmann/Bürokauffrau 156 Jugendliche aus neun Abschlussklassen der beiden zuständigen Hamburger Berufsschulen. Für 135 Schülerinnen und Schüler dieses Ausbildungsberufs liegen Daten aus dem Längsschnitt vor, d. h. hier ist es auch möglich, die Lernentwicklung abzubilden und den Zusammenhängen zwischen den Eingangsleistungen in den allgemein bildenden Domänen und der beruflichen Fachkompetenz am Ende der Ausbildung nachzugehen. Die Tests wurden im dritten Ausbildungsjahr eingesetzt, für die Berliner Jugendlichen am Ende des ersten und für die Hamburger Auszubildenden am Beginn des zweiten Halbjahres im letzten Ausbildungsjahr. Beide Testversionen, Pilotversion und Test für die Hauptuntersuchung, wiesen einen Überschneidungsbereich von ca. 60 Prozent der Items auf. Diese sog. Anker-Items waren hinreichend repräsentativ auf die verschiedenen curricularen Inhalte und kognitiven Anspruchsniveaus verteilt. Eine Verankerung der beiden Testversionen erschien deshalb nicht nur messtheoretisch möglich, sondern sie konnte auch aus kognitionspsychologischer und fachdidaktischer Perspektive inhaltlich gut vertreten werden.

3. Modellgeltung und Skalierung

Das den Untersuchungen zur beruflichen Fachleistung zugrunde liegende Testkonzept ist auf die Messung einer quantitativen Variablen ausgerichtet. Dementsprechend erfolgte die Skalierung des Tests auf Basis der probabilistischen Testtheorie (vgl. dazu Fischer & Molenaar, 1995). IRT-Modelle gehen davon aus, dass dem beobachteten Testverhalten eine Fähigkeit bzw. Disposition zugrunde liegt, die das Testverhalten beeinflusst. Die Fähigkeit einer Person im gemessenen Merkmalsbereich wird als latentes Konstrukt betrachtet; die beobachtbare Leistung ist eine Manifestation dieser Leistungsdisposition. Die Modellierung des Antwortverhaltens erfolgt auf der Grundlage einer nicht-linearen mathematischen Funktion, der sog. Itemfunktion, die die Wahrscheinlichkeit des manifesten Antwortverhaltens auf ein Item in Abhängigkeit von der Ausprägung der zugrunde liegenden Personenfähigkeit (latentes Trait) beschreibt. Aufgabenschwierigkeiten und Personenfähigkeiten werden auf derselben Metrik abgebildet, so dass sie unmittelbar aufeinander bezogen werden können (vgl. Rost, 2004, 96ff.). Üblicherweise wird als Schwierigkeits- oder Lageparameter eines Items der Wendepunkt der Item-Charakterstikkurve gewählt, bei dem die Lösungswahrscheinlichkeit 50 Prozent beträgt. Für den vorliegenden Test wurde, in Anlehnung an die Konventionen in den Large Scale Assessments, über eine Transformation eine höhere Lösungswahrscheinlichkeit von 65 Prozent zur Parametrisierung der Testaufgaben verwendet. Zudem ermöglicht die Skalierung mit probabilistischen Testmodellen eine kriteriumsbezogene Interpretation der Testleistungen, die eine Voraussetzung für die Bildung von Kompetenzniveaus darstellt (Hartig & Jude, 2007, 24).

Die Schätzung der Parameter eines Modells stellt eine wichtige Voraussetzung für die Modellgeltung dar, gleichwohl ist sie keine hinreichende Bedingung (Rost, 2004, 330). Insofern galt es zu prüfen, ob das Rasch-Modell die empirischen Daten angemessen repräsentiert oder ob diese nicht adäquater auf der Grundlage anderer, z. B. klassifizierender, Modelle abgebildet werden. Um die Modellgeltung zu prüfen, wurde unter Nutzung des Computerprogramms WINMIRA auch ein Mixed-Rasch-Modell mit zwei latenten Klassen geprüft. Das Mixed-Rasch-Modell bildet gleichfalls quantitative Personenunterschiede innerhalb von Klassen ab, wobei allerdings die Voraussetzung homogenen Antwortverhaltens preisgegeben wird und Personen mit jeweils ähnlichem Antwortprofil zu Klassen zusammengefasst werden (Rost, 2004, 172f.). In dieser Kombination aus Rasch-Modell und latenter Klassenanalyse wird also davon ausgegangen, dass die getesteten Personen verschiedenen Teilpopulationen angehören, innerhalb derer dann jeweils das Rasch-Modell gilt.

Für die Prüfung der Modellanpassung stehen als Informationsmaße der AIC-Index (Akaike Information Criterion), der BIC-Index (Bayes Information Criterion) sowie der CAIC-Index (Consistent Akaike Information Criterion) zur Verfügung (vgl. dazu ausführlicher Rost, 2004, 220, 157 und 344), über die der Abgleich der testtheoretischen Annahmen erfolgt, wobei das Modell mit den kleineren Werten besser den empirischen Daten entspricht.

Für den vorliegenden Test zeigte sich, dass ein Mixed-Rasch-Modell mit zwei Klassen relativ besser auf die Daten passt im Vergleich zum Rasch-Modell.

Tabelle 1: Ergebnisse der Modellüberprüfung

Index	Modell	
	1 Klasse	2 Klassen
CAIC-Index	21.583.44	21.405.71

In der getesteten Gruppe liegt demnach Personenheterogenität vor, die sich bei der Aufteilung in zwei latente Klassen als signifikant erweist. Die Analyse der Itemprofile zeigte, dass knapp ein Drittel der getesteten Schüler der leistungsstärkeren Gruppe zugeordnet wurden, während etwas mehr als zwei Drittel der leistungsschwächeren Klasse angehörten. Darüber hinaus war an den Itemparametern für die beiden latenten Klassen und an den Profilverläufen erkennbar, dass einige wenige Items das abweichende Profilmuster erzeugten. Im hier vorliegenden Fall wurde daher – unter Anlegung pragmatischer Maßstäbe (Kriterium der Gültigkeit, der Einfachheit und der Brauchbarkeit, vgl. dazu Rost, 2004, 330, 339ff.) – das Rasch-Modell als hinreichende Approximation an die Daten angesehen. Die Geltung des einfachen Rasch-Modells (im strengen Sinne) bedeutet, dass die in der Berücksichtigung latenter Klassen erzeugten Antwortmuster keine zusätzlichen diagnostischen Informationen über die Personen enthalten (Rost, 2004, 199). Die Modellprüfungen zeigten jedoch, dass das quantitative Messmodell die empirischen Daten nicht vollständig erklärt und somit durchaus auch zusätzliche Aufschlüsse über Personeneigenschaften in weiterführenden qualitativen Analysen zu erwarten sind (vgl. dazu auch die Befunde aus dem Test für den Einzelhandelskaufmann in Seeber, 2007b).

4. Dimensionalität des Fachleistungstests

4.1 Hypothesen zur Dimensionalität

Auch wenn in dem hier betrachteten Ausbildungsjahrgang zumindest an den Hamburger Berufsschulen bereits auf der Grundlage eines lernfeldorientierten Ansatzes ausgebildet wurde, schien es aufgrund der unterschiedlichen fachwissenschaftlichen Bezugssysteme des Gegenstandsbereich gerechtfertigt, spezifische inhaltsbezogene Kompetenzen zu prüfen. Die Testaufgaben bezogen sich auf rechtliche, volkswirtschaftliche und einzelbetriebliche Aspekte, die sich in ihren Prinzipien, Begriffen, Begriffssystemen und Strukturen etc. unterscheiden. Insofern war die Interpretierbarkeit eines Gesamttestwertes für den berufsfachlichen Leistungstest kritisch zu prüfen. Zu fragen war folglich, ob die kognitiven Anforderungsstrukturen des hier zur Diskussion stehenden Tests auf ein relativ komplexes, gleichwohl homogenes Fähigkeitsbündel ökonomischen Verständnisses zurückzuführen sind oder ob sie nicht angemessener in einem mehrdimensionalen Bezugsrahmen rekonstruiert werden müssen (zur Dimensionalität vgl. Achtenhagen, 2004; 24; Sloane & Dilger, 2005; auch die Befunde von Straka & Lenz, 2005, 111 zum Zusammenhang zwischen der Variable „ökonomische Bildung“ und „ausbildungsberufsspezifische Fachkompetenz“). Der Test für die angehenden Bürokaufleute umfasste

– Aufgaben mit Bezug auf gesamtwirtschaftliche Zusammenhänge, wirtschaftspolitische Fragen und Probleme sowie Beziehungen zwischen Wirtschaftssubjekten und Ressourcen (*volkswirtschaftliche Dimension*),

- Aufgaben, die sich auf planerische, organisatorische und wirtschaftsmathematische Entscheidungen in Betrieben beziehen und auf eine Optimierung betrieblicher Prozesse zielen (*betriebswirtschaftliche Organisation und Leistungsprozesse*);
- Aufgaben mit Bezug auf Rechtsnormen wirtschaftlichen Handelns, die Regeln des Güter- und Leistungstausches auf dem Markt zwischen Produzenten, Händlern und Konsumenten beinhalten (*rechtliche Dimension*);
- Aufgaben zur systematischen Erfassung und informatorischen Verdichtung der durch den betrieblichen Leistungsprozess entstehenden Geld- und Leistungsströme (Wertschöpfungsdimension); die Aufgaben aus dieser Gruppe zielen auf Konzepte und Prozeduren aus dem Bereich des Rechnungswesens, der aufgrund seines stark axiomatischen Charakters häufig nur schwach mit betriebswirtschaftlichem Inhaltswissen verknüpft ist (vgl. Preiss & Tramm 1996). (Zu dieser Systematik vgl. Tramm & Seeber, 2006).

Im Grundsatz kann – unter Vernachlässigung von Mischformen – von zwei Leitvorstellungen ausgegangen werden, einem „Generalfaktormodell“, das alle Items des Tests berücksichtigt (vgl. Abbildung 1, linke Seite), und einem „Vier-Faktoren-Modell“, in dem die Items zu Teilkompetenzen klassifiziert werden, die freilich im Gegensatz zur traditionellen explorativen Faktorenanalyse korreliert sein können

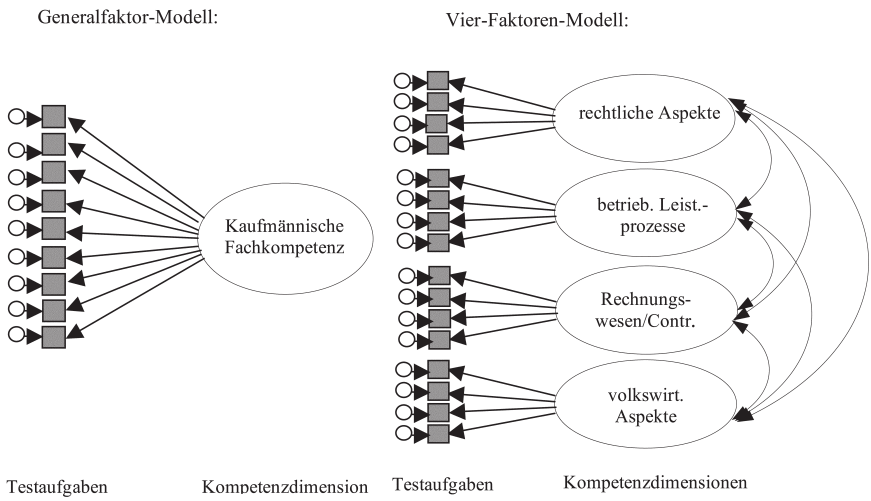

Abbildung 1: Modelle zur Dimensionalität

4.2 Ergebnisse der Dimensionalitätsanalysen

Unter Verwendung des Computerprogramms ConQuest wurden über das einparametrische logistische Modell zunächst zwei Modellvarianten geprüft und über Informationsindizes ihre Anpassung an die empirischen Daten verglichen. Die Modellprüfung erfolgte, wie bereits erwähnt, auf der Grundlage einer gemeinsamen Skalierung der Daten aus der Hamburger und Berliner Stichprobe (N = 284). Die im Vorwege

bereits anhand der Hamburger Stichprobe berechneten Strukturgleichungsmodelle lieferten eindeutige Indizien für domänenspezifische Verständnisfaktoren (vgl. Tramm & Seeber, 2006). Angesichts dessen erschien es sinnvoll, zusätzlich differenzierte Modellprüfungen anhand einer breiteren Stichprobe vorzunehmen.

Die Prüfung der Dimensionalität konnte wegen fehlender Werte im verankerten Datensatz nicht auf der Grundlage von Strukturgleichungsanalysen erfolgen[1], die den Vorteil aussagekräftigerer Modellstatistiken haben, sondern sie musste über die Skalierung ein- und mehrdimensionaler dichotomer Rasch-Modelle und einen Vergleich der Anpassungsindizes erfolgen. Nachfolgende Tabelle gibt Auskunft über die messfehlerbereinigten Korrelationen im vierdimensionalen Modell und enthält darüber hinaus die Subtest-Reliabilitäten sowie die Anzahl der Items in den einzelnen Sub-Dimensionen.

Tabelle 2: Messfehlerbereinigte Interkorrelationen der Sub-Tests im vierdimensionalen Modell für den Ausbildungsberuf „Bürokaufmann/-frau“

Bedeutungskontext	**(1)**	**(2)**	**(3)**	**Anzahl Items**	**Sub-Test WLE- u. EAP/PV Rel.**
BWL: Leist.proz., soz., organis. Dim.(1)	1.00			23	. 58 / .71
Rechtl. Dim. (2)	.65	1.00		15	. 67 / .78
BWL: wertmäßige Dim. (3)	.67	.71	1.00	42	.77 / .80
Volkswirtschaftl. Dim. (4)	.68	.73	.70	15	.55 / .73

Die Interkorrelationen, insbesondere zwischen der Dimension „BWL: Leistungsprozesse, soziale und organisatorische Dimension betrieblichen Handelns“ und den übrigen drei Dimensionen kaufmännischer Kompetenz, fielen verhältnismäßig niedrig aus und deuteten zunächst auf eine relative Eigenständigkeit der berücksichtigten Teilaspekte. Mit dem vorliegenden Test konnte jedoch die Eigenständigkeit zufallskritisch nicht hinreichend abgesichert werden: Beispielsweise zeigte sich bei genauerer Betrachtung der Items der ersten Sub-Dimension, dass mit diesen das obere und untere Leistungsspektrum erfasst wurde, während im mittleren Anforderungsbereich vergleichsweise wenige Items lagen. Dies kann zur relativ geringen Subtest-Reliabilität beigetragen und niedrigere Korrelation zur Folge haben (vgl. Tabelle 2). In ähnlicher Weise war die Subtest-Reliabilität in der volkswirtschaftlichen Dimension für den vorliegenden Test noch nicht zufriedenstellend, was Zweifel an der Legitimität begründete, diese beiden Aspekte als separate Testkomponenten zu behandeln. Trotz viel versprechender Indizien und einer verglichen mit konkurrierenden Konzeptualisierungen besseren Modellanpassung (vgl. Tabelle 3) musste hiernach der vierdimensionale Ansatz einstweilen aufgegeben werden.

Aufgrund des axiomatischen Charakters des Rechnungswesens erschien es folgerichtig, ein weiteres, nunmehr zweidimensionales, Modell zu spezifizieren

1 Multiple Imputationsverfahren mit externen Variablen oder plausible value Technik konnten nicht eingesetzt werden, da für die Daten aus der Berliner Piloterhebung keine externen Variablen, etwa zum Schulabschluss, Geschlecht, Alter, Sozialstatus etc. auf Individualebene zur Verfügung standen.

und zu prüfen. In der ersten Dimension wurden die 56 Items zu den volkswirtschaftlichen Themen, betrieblichen Organisations- und Leistungsprozessen und zu rechtlichen Aspekten kaufmännischen Handelns zusammengefasst; die zweite Dimension umfasste die 39 Items aus dem Bereich des wirtschaftsinstrumentellen Rechnungswesens, also Aufgaben mit wertmäßigem Bezug. Wie im vorgenannten Fall erfolgte auch hier zunächst die Schätzung von Personen- und Itemparametern über das zweidimensionale Rasch-Modell und ein Vergleich der Modellanpassung zwischen der ein- und zweidimensionalen Variante über sog. Informationsindizes (Index aus Likelihood und geschätzter Parameterzahl). Tabelle 3 gibt Auskunft über die Modellanpassung der drei hier geschätzten Modelle.

Tabelle 3: Anpassungsindizes der Dimensionsanalysen des beruflichen Fachtests für den Ausbildungsberuf „Bürokaufmann/-frau“[2]

	Ein-Faktoren-Modell	**Zwei-Faktoren-Modell**	**Vier-Faktoren-Modell**
final deviance/	26681.926	26640.821	26527.038
geschätzte Parameterzahl	96	98	105
Differenz zum eindimensionalen Modell		41,01 bei 2 df; $p < 0.001$	154,89 bei 9 df; $p < 0.001$

Wie die Anpassungsindizes belegen, sind die mehrdimensionalen Ansätze jeweils dem eindimensionalen Modell überlegen; auch die Differenz zwischen der zwei- und vierdimensionalen Variante fällt signifikant zugunsten des differenzierteren Ansatzes aus, allerdings gelten hier die bereits dargestellten Einschränkungen, die eine Annahme des vierdimensionales Modells vorerst als nicht hinreichend begründet erscheinen lässt.

Der *eindimensionale Ansatz* weist bei 95 Items zufriedenstellende Eigenschaften auf, insbesondere eine hohe interne Konsistenz. Jedoch wird eine bessere Passung zwischen empirischen Daten und Testmodell bei einem *zweidimensionalen Ansatz* erreicht. Beide Dimensionen erlangen eine noch akzeptable interne Konsistenz, erfasst anhand der WLE-Reliabilität von 0,81 bzw. von 0,79. Die Information über die Güte der Modellanpassung kann der chi^2-verteilten Differenz aus den beiden Informationsindizes im Verhältnis zu den Freiheitsgraden entnommen werden. Die Differenz von 41,01 bei 2 Freiheitsgraden zeigt eine höchst signifikante Modellverbesserung für den Fall der Annahme der zweidimensionalen Variante (vgl. Tabelle 3). Die messfehlerbereinigte Interkorrelation von 0,78 zwischen den beiden Konstrukten bzw. Subtests deutet darauf hin, dass die Ausdifferenzierung einer eigenständigen Rechnungswesen-Komponente zusätzliche substanzielle Einsichten ermöglichen kann, ungeachtet eines nicht unerheblichen gemeinsamen Varianzanteils von ca. 60 Prozent. Neben differenziellen Analysen rechtfertigt die Höhe der Korrelation

2 Etwaige Abweichungen zu den im ULME-III-Bericht angegebenen Modellparametern des ein- und zweidimensionalen Modells sind auf die Stichprobe zurückzuführen. Erste Dimensionsanalysen im Rahmen der Berichtslegung erfolgten ausschließlich auf Basis der Hamburger Daten, wodurch geringfügig die Kennwerte abweichen. Auch auf breiterer Datenbasis konnten die für Hamburg berichteten Befunde bestätigt werden (vgl. Seeber, 2007, 109).

zwischen den beiden Testkomponenten aber auch Modellierungen auf Basis des eindimensionalen Ansatzes. Es bleibt jedoch Aufgabe nachfolgender Forschungsarbeiten auf der Grundlage einer optimierten Itembasis und breiteren Stichprobe den Fragen der Dimensionalität weiter nachzugehen.

Über inhaltliche Spezifikationen hinaus wurde zusätzlich geprüft, ob die drei Verhaltensklassen (Reproduzieren, Anwenden und vertieftes Verstehen) zwar korrelierte, aber dennoch unterscheidbare Kompetenzdimensionen darstellen. Im Ergebnis der Prüfung zeigten sich latente, messfehlerbereinigte Korrelationen von 0,90 bzw. 0,86 zwischen Aufgaben mit reproduktivem Charakter einerseits und den Aufgaben, die eine Anwendung von Wissen bzw. ein vertieftes Verstehen der Strukturen und Zusammenhänge erfordern. Ähnlich hoch fiel die Korrelation zwischen den Verhaltensklassen Reproduzieren und Verstehen aus ($r = 0,86$). Zwar ergab sich eine signifikant bessere Modellanpassung in der dreidimensionalen Variante, aber die Höhe der Korrelation spricht dagegen, diese als eigenständige Dimensionen zu behandeln.

4.3 Domänenspezifische Befunde zu den Kompetenzstrukturen

Auf der Grundlage der Dimensionsanalysen kristallisierten sich hiernach zwei Inhaltsbereiche (betriebs- und volkswirtschaftliche sowie rechtliche Aspekte als eine und Rechnungswesen als weitere Dimension) heraus, die offenbar durch spezifische Verständnisleistungen jeweils beeinflusst werden.

Auf der Grundlage der zweidimensionalen Rasch-Skalierung wurde es möglich, die beiden Sub-Dimensionen auf zwei miteinander verschränkten Metriken abzubilden und damit die Fachleistungen für die beiden Sub-Tests vergleichend zu diskutieren. Im konkreten Fall wurde die Dimension „Rechnungswesen“ für die Hamburger Gruppe auf einen Mittelwert von 100 und eine Standardabweichung von 25 transformiert; über die Anwendung der gleichen linearen Transformation wurden die Personenlogits der zweiten Dimension „Betriebliche Leistungsprozesse, VWL, Recht“ auf einer damit vergleichbaren Metrik abgebildet, deren Mittelwert für die Hamburger Auszubildenden dann 128 Skalenpunkte bei einer Standardabweichung von 22,9 betrug. Für die Gesamtgruppe der Berliner und Hamburger Stichprobe ergab sich in der wertmäßigen Dimension ein Mittelwert von 108,6 ($SD = 29,6$) und für die Fachkompetenzen auf dem Gebiet volks- und betriebswirtschaftlicher sowie rechtlicher Anforderungen ein Mittelwert von 133,6 ($SD = 30,2$). Freilich steht außer Frage, dass die hier vorgenommenen parallelen linearen Transformationen der Personenlogits und der unmittelbare Vergleich der Leistungen aus den beiden Sub-Tests, die zu relativ günstigeren Kompetenzausprägungen für die eine und weniger erfolgreichen Resultaten für die andere Dimension führen, nur dann gerechtfertigt sind, wenn man von einer Zufallsauswahl der Sub-Testitems aus dem jeweiligen Universum möglicher bzw. curricular sinnvoller Aufgaben ausgeht und wenn Artefakte (z.B. durch Sub-Gruppenunterschiede bedingte Differenzen zwischen den Dimensionen) ausgeschlossen werden können. Soweit als möglich wurde dies sorgfältig geprüft.

Die nachfolgende Verteilung zeigt eine deutlich nach links verschobene Kurve für die Sub-Skala des Rechnungswesens, die dahingehend interpretiert werden kann, dass für diesen Bereich ungünstigere Kompetenzausprägungen vorliegen als für

Anforderungen aus dem Bereich allgemeiner betriebs- und volkswirtschaftlicher Kenntnisse und Fähigkeiten. Man beachte jedoch auch den durchaus beträchtlichen Überschneidungsbereich zwischen den Testbereichen, der aufgrund der substanziellen Korrelationen der beiden Dimensionen auch zu erwarten war.

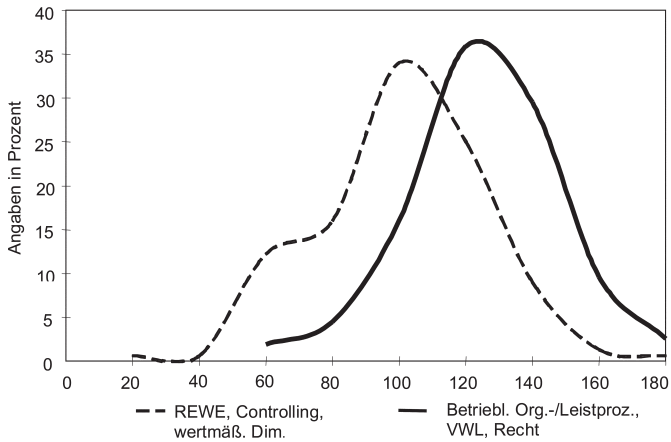

Abbildung 2: Verteilung der Fachleistungen nach Dimensionen

4.4 Determinanten der beruflichen Fachleistung nach Dimensionen

Wie im Abschlussbericht zur Studie herausgearbeitet wurde (vgl. Seeber, 2007, 116), erwiesen sich die mathematischen Kompetenzen zu Beginn der Ausbildung, das Leseverständnis und die Selbsteinschätzungen zu den eingesetzten metakognitiven Strategien im Umgang mit Texten sowie die Fähigkeiten, Informationen aus Grafiken, Tabellen und anderen diskontinuierlichem Texten zu entnehmen einschließlich grundlegender Rechenfertigkeiten (Tests „Texte und Tabellen“) als bedeutsame Prädiktoren für die Erklärung der Varianz in den Schülerleistungen im Gesamttest. In Fortschreibung des hier vorgestellten zweidimensionalen Ansatzes wird nunmehr im Rahmen einer Kommunalitätenanalyse geprüft, welchen spezifischen Beitrag die einzelnen Merkmale, aber auch deren Kombination zur Varianzaufklärung in den beiden untersuchten Sub-Dimensionen leisten. Dieses statistische Verfahren ermöglicht eine Zerlegung der Varianz in spezifische Erklärungsanteile der einzelnen Prädiktoren und in Varianzkomponenten, die sich auf die Kombination von Erklärungsmerkmalen zurückführen lassen. Es wurde mit dem Verfahren der blockweisen Regression gearbeitet, wobei zuerst der „CFT“ (T_2), als nächster Block „Mathematik“ (T_1) und „Texte und Tabellen“ (T_2) und als dritter Block „Leseverständnis“ (T_1) und „Wissen zur Texterschließung“ (T_2) in das Regressionsmodell aufgenommen wurden.[3]

3 T_1 – Messzeitpunkt zu Beginn der Ausbildung, T_2 – Messzeitpunkt gegen Ende der Ausbildung

Insbesondere erwiesen sich dabei die mathematischen Leistungen zu Beginn der Ausbildung und das Verständnis kontinuierlicher und diskontinuierlicher Texte am Ende der Ausbildung, aber auch die Kombination aus „Mathematik/Texte und Tabellen“ einerseits und „Leseverständnis/Metakognitionen im Bereich der Texterschließung“ andererseits als besonders erklärungsmächtig. Rund 8,9 Prozent der Varianz in den Testleistungen der Dimension „Betriebliche Leistungsprozesse, VWL, Recht“ konnten auf die prädiktorspezifische Varianzkomponente „Leseverständnis und Wissen zur Texterschließung (Metakognition)“ zurückgeführt werden. In Bezug auf den Sub-Test *Rechnungswesen* besitzen die beiden zuletzt genannten Merkmale die höchste Erklärungskraft (Leseverständnis und Wissen zur Texterschließung), ebenso tragen die Testleistungen in Mathematik am Ausbildungsbeginn und die kognitiven Grundqualifikationen („Texte und Tabellen“), aber auch deren Kombination zur Variabilität in den Testleistungen bei, freilich mit einem vergleichsweise deutlich niedrigerem Anteil als in der ersten Dimension.

Tabelle 4: Kommunalitätenanalysen für die Determinanten der Fachleistungen in den Subtests „BWL/Leistungsprozesse, Recht und VWL“ und „Rechnungswesen“

Prädiktoren	**1. Dimension: Betriebl. Leistungsprozesse, VWL, Recht**	**2. Dimension: Rechnungswesen/ Controlling**
CFT	0,5	0,3
Mathematik/Texte und Tabellen	15,6	2,5
Leseverständnis/Wissen zur Texterschließung	8,9	5,2
CFT x Mathematik/Texte und Tabellen	0,9	1,5
CFT x Leseverständnis/Wissen zur Texterschließung	0,0	0,7
Mathematik/Texte und Tabellen x Leseverständnis/Wissen zur Texterschließung	11,7	2,6
CFT x Mathematik/ Texte und Tabellen x Leseverständnis/Wissen zur Texterschließung	0,1	0,0
R^2	37,7	12,8

Zusammenfassend ist festzustellen, dass etwa 38 Prozent der Varianz in den Leistungen des Sub-Tests „Betriebliche Leistungsprozesse, VWL und Recht“ durch allgemeine kognitive Grundfähigkeiten, mathematische Fähigkeiten und Lesekompetenz erklärt werden können. Deutlich niedriger fällt mit rund 13 Prozent der Anteil erklärter Varianz durch diese Merkmale für den Teilbereich des Rechnungswesens aus. Hier besitzen offenbar das fachspezifische Vorwissen und die Nutzung fachlicher Konzepte, Begriffssysteme und Prozeduren eine vergleichsweise hohe Bedeutung und können nur bedingt auf allgemeine kognitive Leistungen wie schlussfolgerndes Denken oder die Beherrschung mathematischer Prinzipien und Regeln sowie den Fähigkeiten zur Modellierung von Zusammenhängen zurückgeführt werden. Damit steigt gleichzeitig für diesen Teilbereich die Bedeutsamkeit der Qualität unterrichtlicher Lehr-Lern-Angebote und Entwicklungsräume.

5. Modellierung von Niveaustufen

5.1 Teststruktur und Testgüte des eindimensionalen Modells

Wie im vorangegangenen Abschnitt aufgezeigt, ermöglicht die Teststruktur Auswertungen sowohl auf Basis eines Gesamttestwertes als auch auf der Grundlage zweidimensionaler Kompetenzprofile der Jugendlichen. Die Verbesserung der Modellanpassung durch letztgenanntes Konzept war jedoch vergleichsweise gering. Deshalb soll im Sinne einer theoretisch sparsamen Modellierung von Kompetenzniveaus nachfolgend von der Geltung des eindimensionalen Modells ausgegangen werden.

Der Gesamttest erreicht – basierend auf 95 Items – eine hohe interne Konsistenz mit einer WLE-Reliabilität von 0,87. Die Diskriminanzwerte der Items liegen – bis auf wenige Ausnahmen – zwischen 0,18 und 0,49; die Item-Fit-Werte im Messmodell (Weighted Fit MNSQ) variieren in dem akzeptablen Bereich zwischen 0,86 und 1,15. Die ausgeschlossenen 17 Items blieben aufgrund mangelnder Trennschärfe oder analytischer Abhängigkeit unberücksichtigt.

Insgesamt ist es mit dem vorliegenden Test gelungen, die unterschiedlichen fachbezogenen Inhaltsbereiche jeweils auf breiter Anforderungsbasis abzubilden. Abbildung 3 weist die Schwierigkeitskennwerte der 95 Items aus, einerseits nach Inhaltsbereichen und andererseits nach Verhaltensklassen differenziert. Wie ersichtlich, verteilen sich die Items in jeder Sub-Dimensionen relativ gut über das Schwierigkeitsspektrum.

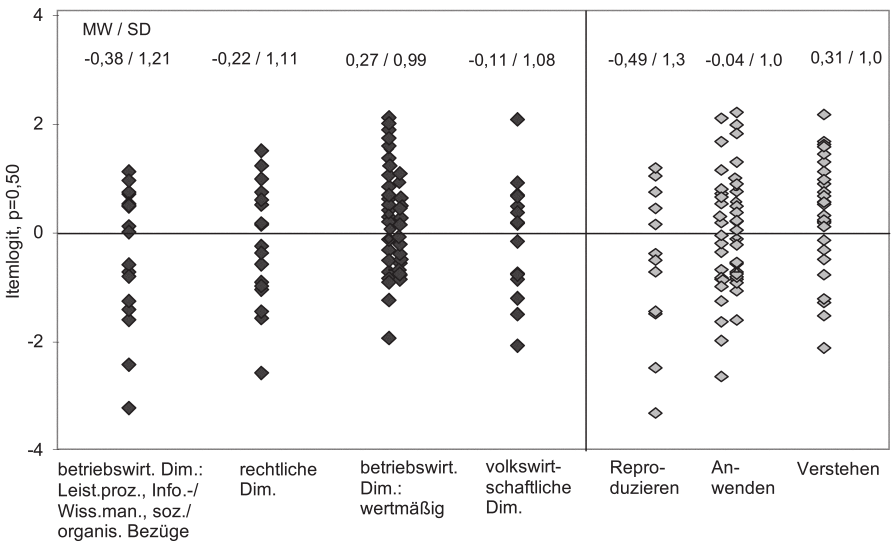

Abbildung 3: Verteilung der Itemschwierigkeiten nach inhaltlicher Dimension und nach Verhaltenserwartungen

5.2 Ansätze und Verfahren der Kompetenzmodellierung

Kompetenzniveaus (oder auch Kompetenzstufen) stellen Abschnitte auf kontinuierlichen Kompetenzskalen zur *kriteriumsorientierten* Beschreibung der erfassten Kompetenzen dar (vgl. Adams & Wu, 2000, 197; Hartig, 2007, 86). Quantitative Differenzen in der Leistungsfähigkeit werden zugleich als qualitative Unterschiede in Bezug auf die kognitive Ausstattung erfasst. Blum et al. (2004, 55) bezeichnen Kompetenzstufen als „heuristisches Hilfsmittel, mit welchem man eine abstrakte Skala „zum Sprechen bringen" kann". Die (kontinuierliche) Fähigkeitsskala wird segmentiert in inhaltlich definierte Kompetenzbereiche, wobei jede Kompetenzstufe „durch kognitive Prozesse und Handlungen von bestimmter Qualität spezifiziert (ist – d. V.), die Schülerinnen und Schüler auf dieser Stufe bewältigen können, nicht aber auf niedrigeren Stufen" (Klieme et al., 2000, 116). Rost (2004, 100) betrachtet dies als einen Versuch, „qualitative Interpretationskategorien in das eigentlich quantitative Messmodell von Rasch zu integrieren". Neben dem Vorteil, die als kontinuierliches Merkmal definierten Fähigkeiten der Schülerinnen und Schüler auf einige wenige anhand von konkreten Anforderungen und kognitiven Operationen beschreibbare Kompetenzniveaus zu reduzieren und damit die Anschaulichkeit und bildungspolitische sowie öffentliche Kommunizierbarkeit von quantitativen Leistungsdaten zu erhöhen (Hartig, 2004, 2), geht die Bildung von Kompetenzniveaus mit dem Nachteil einer Reduzierung von Informationen einher. Zum einen erfolgt keine weitere inhaltliche Ausdifferenzierung der Kompetenzen innerhalb einer Stufe, die sich jedoch zwischen oberen und unteren Rand durchaus unterscheiden können, zum anderen ist davon auszugehen, dass Schülerinnen und Schüler, deren Leistungen an der Grenze zwischen zwei Kompetenzniveaus liegen, sich weitaus ähnlicher sind als Schüler, deren Kompetenz innerhalb eines Niveaus jeweils am unteren und oberen Rand verortet wird.

Im Rahmen großflächiger internationaler wie auch nationaler Schulleistungsuntersuchungen finden verschiedene Verfahren der Kompetenzmodellierung Anwendung, die – in Abhängigkeit vom Ansatz – mit einem unterschiedlichen Grad an kognitionstheoretischer und fachdidaktischer Fundierung sowie empirischer Validierung verbunden sind. Zu nennen sind in diesem Zusammenhang der von Beaton & Allen (1992) entwickelte Ansatz zur Bestimmung von Kompetenzniveaus, der sich auf *Post-hoc-Analysen* der Testaufgaben gründet, und der von Hartig (2007) im Kontext der DESI-Studie entwickelte Ansatz einer empirisch begründeten Gruppierung der Testaufgaben durch *a priori* gewählte schwierigkeitsbestimmende Merkmale. Ein zentrales Merkmal beider Ansätze ist die Unterteilung der kontinuierlichen Kompetenzskala in Kompetenzabschnitte über eine ermöglichte Bestimmung von Schwellen zwischen den Stufen (Hartig, 2004, 5). Konstitutiv ist jeweils die Verortung von Personenfähigkeit und Aufgabenschwierigkeit auf einer einzigen Metrik. Unterschiede zwischen diesen beiden Verfahren sind hinsichtlich der einbezogenen Informationen über die Testaufgaben und des Grads an Generalisierbarkeit von Aufgabenanforderungen festzustellen.

5.3 Gruppierung der Testaufgaben über schwierigkeitsbestimmende Anforderungsmerkmale

Für die Bestimmung von Kompetenzniveaus auf Basis einer a priori-Spezifikation von Aufgabenanforderungen werden im Prinzip bereits vor der Erhebung der Leistungsdaten entsprechende inhaltliche Hypothesen formuliert. Bei diesem Verfahren werden relevante Merkmale (z.B. Merkmale des Lösungsprozesses, der Aufgabenformate, der Situierung der Aufgaben und der Anforderungen an kognitive Prozesse), von denen angenommen werden, dass sie die Schwierigkeiten von Testaufgaben beeinflussen, vorab erarbeitet und jede Testaufgabe wird dementsprechend multipel klassifiziert. An einem solchen Verfahren sind in der Regel Experten aus verschiedenen Disziplinen wie der Fachdidaktik, der Lehr-Lern-Forschung, der Psychometrie sowie ggfs. der Fachwissenschaft beteiligt. Nach der Zuschreibung der Aufgabenmerkmale erfolgt die regressionsanalytische Bestimmung von Schwellen anhand von Aufgaben-Merkmalskombinationen nach theoretischen und empirischen Kriterien (vgl. Hartig, 2007). Dieses anspruchsvolle und stärker modellgeleitete Vorgehen konnte bislang nur selten in internationalen und nationalen Leistungsstudien umgesetzt werden (vgl. die Ergebnisse aus TIMSS sowie die Aufarbeitung der Befunde einschlägiger Forschung für Mathematik in Klieme et al., 2000, 119ff.). Eine Ausnahme stellt hier – wie erwähnt – die PISA-Ergänzungsstudie DESI dar, in deren Rahmen die sprachlichen Kompetenzen der 15jährigen in Deutsch und Englisch untersucht wurden (vgl. die Beiträge in Beck & Klieme, 2007).

Wie bereits im Abschnitt 2 kurz skizziert, erfolgte bei ULME im Rahmen der Itemkonstruktion eine erste Klassifikation der Aufgabenanforderungen durch Fachdidaktik- und Unterrichtsexperten. Über die bereits dargestellten Zuordnungen nach Wissensart und kognitiver Leistung hinaus wurden jedoch der Testentwicklung weitere Klassifikationsmerkmale zugrunde gelegt, zu denen u. a. Merkmale der Informationsstruktur der Aufgaben (Text, Tabellen, Grafiken etc.), die Anzahl der in den Items repräsentierten Inhaltsbereiche, Aspekte der Situierung sowie Aufgabenformate gehörten. Angesichts der Breite der Anforderungen und der Vielfalt der für die Aufgabenlösungen möglichen und notwendigen kognitiven Operationen erwies sich eine a priori Klassifikation der Testaufgaben über die hier genannten Kategorien hinaus als ausgesprochen anspruchsvoll, will man nicht nur Globaleinschätzungen im Sinne eines ‘Mehr‘ oder ‘Weniger’ an Komplexität, an Abstraktionsleistungen etc. zugrunde legen. Vor allem deshalb wurde geprüft, inwiefern die genannten a priori klassifizierten Merkmale einen spezifischen Beitrag zur Erklärung der Aufgabenschwierigkeit liefern. Nach Vorliegen der Ergebnisse erfolgte dann nochmals eine intensive Auseinandersetzung in der Expertengruppe, da einige Zuordnungen schwer mit den empirisch ermittelten Itemschwierigkeiten zu vereinbaren waren. Beispielsweise wurden in den didaktischen Vorgaben alle Aufgaben, die die Anwendung von Prozeduren, also die Abarbeitung eines Algorithmus erfordern, unter prozeduralem Wissen klassifiziert, und zwar ausschließlich hier. Bei einigen Aufgaben zeigte sich jedoch, dass zwar die Grundstruktur der Aufgabe durch die Arbeitung eines Algorithmus gut beschrieben werden kann, dass aber die Zwischenschritte mit vielfältigen konzeptuellen Überlegungen, Über- und Unterordnungen von ökonomischen Kategorien verbunden waren und somit Abstraktionsleistungen und anspruchsvolle Prozesse der Dekontextualisierung erforderten. Ein Beispiel für eine solche Aufgabe stellt die Ermittlung des günstigsten Angebots aus zwei Alternativen

mit jeweils fünf bzw. sechs ökonomisch relevanten Parametern dar, die in der Entscheidungsfindung zu berücksichtigen waren. Dabei war nicht nur die Reihenfolge der Schritte zur Ermittlung der Einstandspreise festzulegen, sondern erforderliches prozedurales Wissen musste mit entsprechendem Wissen über relevante Konzepte und Prinzipien verbunden werden (z. B. Bedeutung und konzeptionelle Basis der Liefer- und Einkaufskonditionen), um so die korrekten Parameter auszuwählen und die Preise über mathematische Operationen zu ermitteln. Im Zuge der Arbeiten zur Kompetenzmodellierung wurden deshalb alle Testaufgaben nochmals inspiziert. Hiernach wurden Items mit prozeduralem Charakter, die neben der Verfügbarkeit einer Routine auch konzeptuelles Wissen verlangten, beiden Kategorien zugeordnet, so dass auch Kombinationseffekte der angesprochenen Wissensarten untersucht werden konnten.

Im Übrigen haben Befunde aus der psychologischen sowie lehr-lern-theoretischen Forschung auf eine kontroverse Diskussion zu den verschiedenen Konzepten und Modellen zur taxonomischen Klassifikation der Wissensarten geführt. Unter dem Oberbegriff der dual components oder dual process theories of cognition wird eine Reihe von unterschiedlichen Ansätzen zur Bildung von kognitionsbezogenen Begriffspaaren unterschieden, die auf Arbeiten aus den Bereichen der kognitiven, pädagogischen und Entwicklungspsychologie zurückgehen, aber auch in der Psycholinguistik und Philosophie verankert sind (vgl. einen Überblick in Schneider, 2006, 48ff.). Zwar ist die Unterscheidung von Wissensarten in der Forschung generell nicht unumstritten (zur ablehnenden Haltung vgl. z. B. die Vertreter der situierten Kognition wie Greeno, 1997), jedoch verweisen Befunde aus der lernpsychologischen und Gedächtnisforschung, insbesondere aus Experimentalstudien, auf die Plausibilität solcher Differenzierungen (vgl. Anderson, 1983, Rittle-Johnson & Alibali, 1999). Wenngleich also ein Einfluss distinkter Wissensarten auf die Schwierigkeit von Testaufgaben empirisch nicht zweifelsfrei belegbar ist, so erschien dennoch für den vorliegenden Test der Versuch angezeigt, derartige Annahmen zu prüfen; zumal erste Befunde zu Kompetenzanalysen in wirtschaftsberuflichen Bildungsgängen auf einen substanziellen Erklärungsbeitrag der Wissensart verwiesen haben (vgl. dazu Seeber, 2007b; Seeber, 2005). Auch in Mikroprozessanalysen zum Einfluss instruktionaler Erklärungen im Rechnungswesen erwiesen sich prinzipienbasierte Überlegungen als bedeutsamer Lernerfolgsprädiktor (vgl. Stark, Hinkofer & Mandl, 2001) und stützen damit gleichfalls diese Erwartungen.

Mit dem Verfahren der Rasch-Skalierung wurde demzufolge für jedes Test-Item j ein Schwierigkeitsparameter σ_j ermittelt, der die Verortung der Aufgabe auf der Kompetenzskala festlegt (vgl. Abbildung 4). Um zu prüfen, inwiefern die vorab den Items zuordneten Aufgabenmerkmale eine Beitrag zur Erklärung der Aufgabenschwierigkeit liefern, wurde das Verfahren der Regressionsanalyse gewählt. Hierfür wurden die 95 Test-Items im Datensatz als 'Fälle' behandelt; die Aufgabenschwierigkeit fungierte als abhängige Variable. Als Prädiktoren gingen die Wissensarten, die repräsentierten Inhaltsbereiche (übergreifend vs. nicht übergreifend) sowie Merkmale der Präsentation der Inhalte (diskontinuierlich in Form von Tabellen und Grafiken vs. nicht diskontinuierlich, nur Textformat) in das Regressionsmodell ein, und zwar jeweils in Form einer Dummy-Codierung. Darüber hinaus wurde angenommen, dass nicht nur die einzelnen Merkmale für sich die Schwierigkeit der Aufgaben beeinflussen, sondern auch die Kombination von Merkmalen substanzielle Erklärungsbeiträge liefert. In der nachfolgenden Tabelle 4 sind die unstandardisierten und

standardisierten Regressionskoeffizienten sowie die Irrtumswahrscheinlichkeit für den jeweiligen α-Fehler ausgewiesen.

Tabelle 4: Determinanten der Aufgabenschwierigkeit

	Nicht standardisierte Koeffizienten B	**Standardisierte Koeffizienten Beta**
(Konstante)	-,651	
Konzeptwissen	1,424	,660***
Prozedurales Wissen	1,396	,363***
Verknüpfung von Konzept- und prozeduralem Wissen	1,150	,403***
lernfeldübergreifende Aufgaben	,886	,310***

Abhängige Variable: Rasch-skalierte Aufgabenschwierigkeit; $R^2 = .41$; *** $p < 0.000$

Wie die Ergebnisse zeigen, ist es vergleichsweise gut gelungen, mit den vorab klassifizierten Aufgabenmerkmalen einen Erklärungsbeitrag für die Schwierigkeit der Testaufgaben zu erlangen (vgl. etwa den für ULME II berichteten Erklärungsanteil R^2=.28 in Seeber, 2005). Als stärkster Prädiktor erwiesen sich Aufgaben, die Struktur- und Zusammenhangswissen erfordern. Auch Aufgaben, deren Bearbeitung Prozeduren wie mathematische Berechnungen, die Anwendung standardisierter Verfahren sowie routinisierte wie auch komplexere Handlungsabläufe verlangten, ließen sich in ihrer Schwierigkeit auf diese Weise gut bestimmen. Insbesondere jene Aufgaben, die eine Kombination von prozeduralem und konzeptuellem Wissen erforderten, für die also Heuristiken unter Rückgriff auf kategoriale und vernetzte Wissenssysteme zu entwickeln und einzusetzen waren, stellten in diesem Test – ausweislich des Regressionskoeffizienten – hohe kognitive Anforderungen. Mit einem etwas geringeren, aber immer noch bedeutsamen Einfluss auf die Aufgabenschwierigkeit waren inhaltsübergreifende Anforderungen verbunden. Der Einfluss anderer Klassifikationsmerkmale wie der der Präsentationsformen der Inhalte, der systemischen Bezüge oder der Textlänge ließ sich im Modell nicht zufallskritisch absichern.

Weitgehend unberücksichtigt oder nur sehr global über die Wissensstrukturen erfasst, blieben die Ansprüche an die kognitiven Operationen, aber auch Aspekte der Motivation wie beispielsweise eine zwischen Aufgaben variierende zusätzliche Stimulanz (zu Befunden aus der Motivationsforschung vgl. z.B. Prenzel, Kramer & Drechsel, 1998 und 2001), Merkmale der Situierung der Aufgaben (z. B. Kürze vs. Weitschweifigkeit; Einfachheit vs. Komplexität) und der Plausibilität der Distraktoren (vgl. hierzu Kirsch, Jungeblut & Mosenthal, 1998, 125). Hier sind vertiefende fachdidaktische und kognitionspsychologische Aufgabenanalysen notwendig, um einer stärker modellgeleiteten Generierung der Kompetenzstufen zu entsprechen. Der Anspruch, ein umfassendes System von Merkmalen zu ermitteln, anhand derer die Schwierigkeit jeder Aufgabe eindeutig prognostiziert werden kann, musste zurückgestellt werden. Angesichts der noch offenen Probleme der Aufgabenanalyse wurde deshalb für den vorliegenden Test zunächst ein Ansatz zur Kompetenzmodellierung auf Basis von post-hoc-Analysen gewählt.

5.4 Post-hoc-Analysen zur Bestimmung von Kompetenzniveaus

Die im Rahmen einer post-hoc-Analyse durchgeführte Definition von Kompetenzniveaus basiert auf einer Inspektion der Testaufgaben (vgl. zu diesem Vorgehen Beaton & Allen, 1992; Klieme et al., 2000, 116ff.). Die Bestimmung von Schwellen erfolgt über sog. Markier-Items, d.h. eine Verankerung von Kompetenzniveaus basiert nicht auf der Gesamtheit der Testaufgaben, sondern auf charakteristischen Aufgaben, die auf entsprechender Stufe mit hinreichender Sicherheit gelöst werden, nicht jedoch von den Personen, die im darunter liegenden Kompetenzniveau verortet sind. Für die Bestimmung der Markier-Items und damit letztlich auch der Lage der Schwellen sind nur jene Aufgaben von Interesse, die an der unteren Grenze eines Kompetenzniveaus liegen und von den Schülerinnen und Schülern der nächst niedrigeren Stufe deutlich weniger erfolgreich bearbeitet werden als auf dieser Stufe erwartet. Nach Beaton & Allen (1992) sollten die charakteristischen Aufgaben eines Kompetenzniveaus mit mindestens 65-prozentiger Wahrscheinlichkeit von den Schülerinnen und Schülern des entsprechenden Niveaus gelöst werden; auf dem niedrigeren Niveau jedoch mit einer Wahrscheinlichkeit unter 50 Prozent. Auf dem ersten Blick erscheint ein solches Vorgehen willkürlich und ist auch nur dann zu rechtfertigen, „wenn es empirisch gelingt, für jede Kompetenzstufe eine ausreichende Zahl charakteristischer Aufgaben zu bestimmen und durch konsistente Expertenurteile hinsichtlich der didaktischen und kognitiven Anforderungen abzusichern“ (Klieme et al., 2000, 118).[4]

Nach Durchsicht und Analyse der Testaufgaben und der Lage der Itemparameter wurde – in Anlehnung an andere Studien – aus Gründen der Konkretisierung der Informationen von einer beschränkten Anzahl von Kompetenzniveaus ausgegangen. Im vorliegend Kontext erschienen vier Niveaus als hinreichend unterscheidbar (vgl. Abbildung 4). Die erste Schwelle, zwischen Niveau I und II, wurde bei 80 Skalenpunkten festgelegt; die zweite Abgrenzung zwischen den Niveaus II und III bei 125 Skalenpunkten und der Übergang von Niveau III auf Niveau IV ist mit gleichem Abstand auf 170 Punkte festgelegt worden. Damit weisen die beiden mittleren Intervalle – entsprechend der methodischen Empfehlungen von Beaton & Allen (1992) – gleich große Abstände auf, während die erste und die höchste Niveaustufe nach unten bzw. nach oben offen sind.

Folgenden Beschreibungen der so festgelegten Kompetenzniveaus sind unter Berücksichtigung der Details begründbar:

> *Kompetenzniveau I:* Jugendliche auf dem unteren Kompetenzniveaus sind in der Lage, ökonomische Aufgabenstellungen aus einem Alltagsverständnis heraus und durch in alltäglicher Kommunikation erworbene Vorstellungen zu bearbeiten. Ein Beispiel für eine solche Aufgabe ist die Zuordnung von verschiedenen Arten des Zahlungsverkehrs

4 Während in TIMSS – und auch für den hier vorliegenden Test – eine relativ starke Anlehnung an das von Beaton & Allen vorgeschlagene Verfahren erfolgte, wurden in PISA für den Bereich der Mathematik nicht nur die Markier-Items, sondern alle 84 Testaufgaben und darüber hinaus die ca. 200 Feld-Testaufgaben zur Beschreibung der Kompetenzniveaus herangezogen worden. Diese große Itemzahl machte es möglich, die Kompetenzstufen nicht nur relativ global zu beschreiben, sondern Stufenmodelle – getrennt nach den vier übergreifenden Ideen (Quantität, Veränderung und Beziehungen, Raum und Zeit, Unsicherheit) zu entwickeln und zu beschreiben, aus denen wiederum eine übergreifende Metabeschreibung für den Gesamttest generiert wurde (Blum et al., 2004, 56).

(Zielkauf, Skonto, Barzahlung, Ratenkauf) zu bestimmten wirtschaftlichen Situationen (Kaufanreiz bei hochpreisigen Produkten, unbekannter Kunde, bekannten Kunden zum Kauf anhalten, eigene Liquidität durch Anreiz für kurzfristigen Rechnungsausgleich sichern). Während die Zuordnung der Barzahlung bei einem Geschäft mit unbekanntem Kunden von den Jugendlichen des untersten Niveaus erfolgreich bearbeitet werden konnte, zeigten sich bei allen übrigen Zuordnungen, die zumindest Kenntnis der ökonomischen Begriffe und basales Wissen über die damit verbundenen wirtschaftlichen Ziele und Absichten voraussetzten, für diese Gruppe deutlich Probleme.

Kompetenzniveau II: Jugendliche auf Niveaustufe II können mit gebotener Sicherheit einfache wirtschaftsmathematische Prozeduren (z. B. Skonto- und Rabattberechnungen) sowie mehrschrittige Berechnungen durchführen (Ermittlung eines Jahreseinkommens anhand eines gegebenen Monatsumsatzes und dem prozentualen Anteil der Provision). Sie verfügen über grundlegendes ökonomisches Faktenwissen und können ökonomische Begriffe und Kategorien in bestimmten Situationen unter Nutzung von mathematischen Prozeduren anwenden (z. B. Ermitteln des Saldos für ein Konto unter Berücksichtigung des Anfangsguthabens und diverser Kontenbewegungen). Die Jugendlichen sind in der Lage, wirtschaftliche Entscheidungen unter Berücksichtigung von zwei, max. drei Randbedingungen zu treffen und einfache betriebliche Zusammenhänge zu modellieren.

Kompetenznivau III: Die Jugendlichen auf Kompetenzniveau III verfügen über konzeptuelles und prozedurales ökonomisches Wissen. Sie sind fähig, Algorithmen und Heuristiken zur Aufgabenlösung zu entwickeln und auch komplexere wirtschaftliche Zusammenhänge zu modellieren, und zwar sowohl auf einzelbetrieblicher als auch auf aggregierter Ebene. Ein sog. Markier-Item des Niveaus III, das von weniger als der Hälfte der Jugendlichen des Niveaus II mit guter Erfolgswahrscheinlichkeit bearbeitet werden konnte, bezieht sich auf die Zusammenhänge zwischen einer Erhöhung der Beitragsbemessungsgrenze für die Krankenversicherung und den möglichen Konsequenzen für Arbeitgeber und Arbeitnehmer. Die Jugendlichen des Niveaus III kennen grundlegende rechtliche Regelungen im Bereich des Güter- und Leistungsaustauschs und können diese auf konkrete Situationen anwenden. Weitere Beispiele für sog. Markier-Items auf Stufe III sind Aufgaben zur Ermittlung der Umsatzsteuerzahllast an das Finanzamt anhand einer eingegangen Lieferantenrechnung und einer entsprechenden Rechnung an den Kunden, nachdem das Produkt im Unternehmen fertig gestellt wurde. Zu den Markier-Items der Stufe III gehört auch die Ermittlung des günstigsten Anschaffungspreises eines PKW, bei dem aus den gegebenen Informationen, jene auszuwählen sind, die für die Preisermittlung tatsächlich relevant sind. Abschließend sind die Preise zu berechnen und zu vergleichen.

Kompetenzniveau IV: Es ist bemerkenswert, dass vor allem Aufgaben aus dem Bereich des wirtschaftsinstrumentellen Rechnungswesen im oberen Kompetenzspektrum liegen, d.h. Aufgaben die ein vertieftes Verständnis ökonomischer Beziehungen erfordern und in der Regel ein anspruchsvolles Abstraktionsniveau der zu modellierenden Zusammenhänge erfordern. Aufgaben dieses Niveaus erfordern einen sicheren Umgang mit Instrumenten, Regeln und Verfahren des Controllings, aber auch konzeptuelles Wissen über die zugrunde liegenden Zusammenhänge. Die Bearbeitung der Aufgaben auf der höchsten Niveaustufe erfordert in der Regel komplexe Informationsverarbeitungsprozesse im Rahmen der Situationsanalyse, einen souveränen Umgang mit ökonomischen Kategorien und die Entwicklung meist vielschrittiger Algorithmen. Ein Beispiel für eine Aufgabe auf Niveau IV ist die Ermittlung des Gewinns resp. des Verlusts anhand der Angaben zur Höhe der Schulden und des Vermögens jeweils am Jahresanfang und Jahresende.

Die nachfolgende Abbildung veranschaulicht die Lage der Itemparameter und die Verteilung der Schülerleistungen auf der Fähigkeits-/Schwierigkeitsskala. Darüber hinaus sind die Kompetenzniveaus abgetragen und mit schraffierten Kästchen die

Markier-Items gekennzeichnet, die von den Schülerinnen und Schülern des darunter liegenden Niveaus mit einer Wahrscheinlichkeit unter 50 Prozent gelöst werden, während Schüler des entsprechenden Niveaus diese mit 65prozentiger Sicherheit bearbeiten. Für Aufgaben oberhalb der Markier-Items nimmt die Wahrscheinlichkeit der erfolgereichen Aufgabenlösung ab.

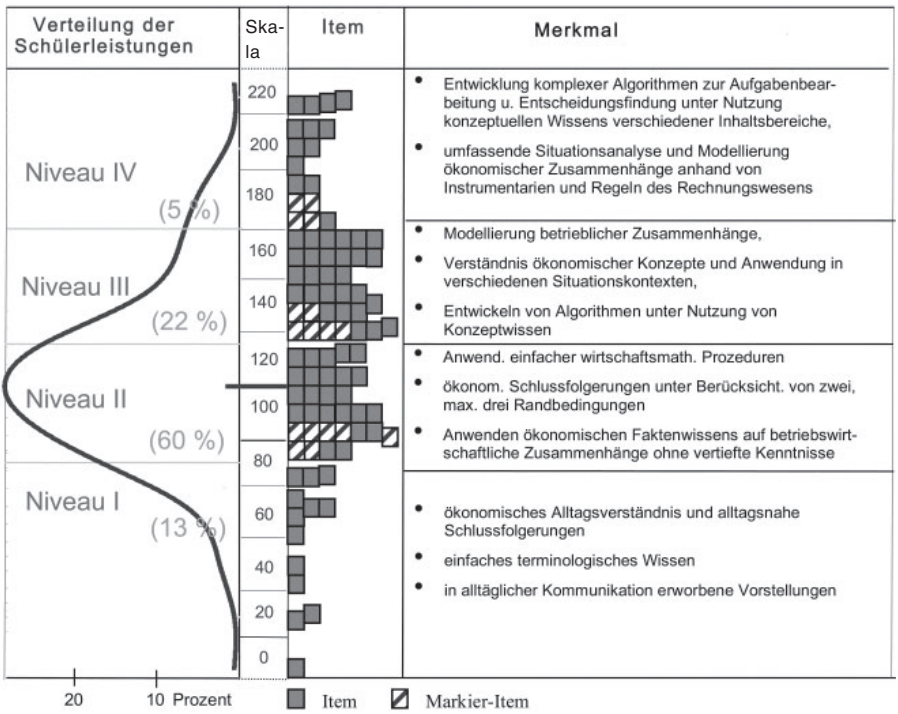

Abbildung 4: Verteilung der Schülerleistungen und der Aufgabenschwierigkeiten im Fachtest „Bürokaufmann/Bürokauffrau“ und Abschnitte auf den Kompetenzskalen

Wie aus der Abbildung hervorgeht, decken die Items ein breites Spektrum von Anforderungen ab, bei denen, wie zuvor ausgeführt, vier Kompetenzniveaus unterschieden werden können. Allerdings zeigt sich, abzulesen an den von der Verteilungskurve jeweils überstrichenen Flächen, dass allenfalls ein Viertel der getesteten Jugendlichen die Aufgaben der beiden oberen Niveaustufen mit der gebotenen Sicherheit zu lösen vermag.

6. Ausblick

In dem vorliegenden Beitrag wurden erste Analysen zur Kompetenzstruktur und zu Kompetenzniveaus für einen ausgewählten Ausbildungsberuf vorgestellt. Nicht wenige Fragen sind damit allerdings eher aufgeworfen als bereits geklärt worden.

Dies gilt beispielsweise für das Problem, wie fach- und situationsspezifische Anforderungsmerkmale systematisch aufeinander zu beziehen sind, wenn generalisierbare und empirische fundierte Niveaubeschreibungen das Ziel sind. Zu untersuchen bleibt auch, wie sich die Ergebnisse etwa zu anderen Ansätzen zur Erklärung von Kompetenzen verhalten. Zu nennen wäre hier beispielsweise der Ansatz Minnameiers (2006), der zur Analyse von Prüfungsaufgaben Varianten der inferentiellen Theorie des Wissenserwerbs anwendet und in diesem Zusammenhang einräumt, dass zwar auf einem solchen Wege die spezifischen (qualitativen – d. V.) Leistungsanforderungen der Aufgaben herausgearbeitet werden konnten, dass aber daraus keine (quantitativen – d. V.) Aussagen zur Aufgabenschwierigkeit ableitbar sind. Aus seiner Sicht könnte die Kopplung von inferentieller Theorie und kognitiven Stufen (im piagetschen Sinne) eventuell zur Diagnostik von Kompetenzen bzw. Kompetenzdefiziten beitragen (Minnameier, 2006, 402ff.). Allerdings ist hier – angesichts der Befunde zur Aufgabenklassifikationen im Rahmen nationaler und internationaler Studien, aber auch im Kontext von Kontroll- und Experimentalstudien, Skepsis angezeigt, über die letztlich nur durch eine empirische Prüfung entschieden werden kann.

Auch aus dem DFG-Schwerpunktprogramm „Lehr-Lern-Prozesse in der kaufmännischen Erstausbildung" (vgl. Beck & Dubs, 1998) und darauf aufbauenden und weiterführenden Forschungsarbeiten liegen eine Vielzahl interessanter Befunde zu Detailfragen der Kompetenzentwicklung und zu Zusammenhängen zwischen Merkmalen der Lernumgebung, der Lernarrangements sowie des Lehrerverhaltens (Prozessmerkmale) und dem Lernerfolg der Schüler (Outputmerkmale) vor. Angesichts der häufig sehr speziellen Fragestellungen dieser Arbeiten ist jedoch nicht sicher, ob die Befunde generalisiert werden können, namentlich in Richtung auf die pragmatisch notwendige Spezifikation von Anforderungsprofilen.

Literatur

Achtenhagen, F. (2004). Prüfung von Leistungsindikatoren für die Berufsbildung sowie zur Ausdifferenzierung beruflicher Kompetenzprofile nach Wissensarten. In Baethge, M., Buss, K.-P. & Lanfer, C. (Hrsg.), Expertisen zu den konzeptionellen Grundlagen für einen Nationalen Bildungsbericht – Berufliche Bildung und Weiterbildung/Lebenslanges Lernen. Bildungsreform Band 8. Bonn: Bundesministerium für Bildung und Forschung (BMBF), 11-32.

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

Anderson, J. R. (1983). The architecture of cognition (10th ed.). Cambridge, MA.: Harvard University Press.

Anderson, L. W. & Krathwohl, D. R. (with Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R. et al.) (Eds.) (2001). A Taxonomy for Learning, Teaching, and Assessing. A Revision of Bloom's Taxonomy of Educational Objectives. New York.

Baethge, M., Buss, K.-P. & Lanfer, C. (2003). Konzeptionelle Grundlagen für einen Nationalen Berufsbildungsbericht: Berufliche Bildung und Weiterbildung/Lebenslanges Lernen. Bildungsreform Band 7. Bundesministerium für Bildung und Forschung. Bonn.

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

Baumert, J. Lehmann, R. & Lehrke, M., Schmitz, B., Clausen, M. Hosenfeld, I., Köller, O. & Neubrand, J. (1997). TIMSS – Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich. Deskriptive Befunde. Opladen: Leske + Budrich.

Baumert, J., Köller, O., Lehrke, M. & Brockmann, J. (2000). Anlage und Durchführung der Dritten Internationalen Mathematik- und Naturwissenschaftsstudie zur Sekundarstufe II (TIMSS/III – Technische Grundlagen. In Baumert, J., Bos, W. & Lehmann, R. (Hrsg.), Dritte

Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Grundbildung am Ende der Schullaufbahn. Band 1: Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit. Opladen: Leske + Budrich, 31-84.

Beaton, A. E. & Allen, N. L. (1992). Interpreting scales through scale anchoring. Journal of Educational Statistics, 17 (2), 191-204.

Beck, B. & Klieme, E. (Hrsg.) (2007). Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International). Weinheim und Basel: Beltz.

Beck, K. & Dubs, R. (Hrsg.). Kompetenzentwicklung in der Berufserziehung: kognitive, motivationale und moralische Dimensionen kaufmännischer Qualifizierungsprozesse. Zeitschrift für Berufs- und Wirtschaftspädagogik, Beihefte, Heft 14. Stuttgart: Franz Steiner.

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

Blum, W., Neubrand, M., Ehmke, T., Senkbeil, M. Jordan, A., Ulfig, F. & Carstensen, C. H. (2004). Mathematische Kompetenz. In PISA-Konsortium Deutschland (Hrsg.), PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs. Münster, New York, München, Berlin: Waxmann, 47-92.

Brand, W., Hofmeister, W. & Tramm, T. (2005). Auf dem Weg zu einem Kompetenzstufenmodell für die berufliche Bildung – Erfahrungen aus dem Projekt ULME. In Brand, W. & Tramm, T. (Hrsg.), Prüfungen und Standards in der beruflichen Bildung. In bwp@ - Berufs- und Wirtschaftspädagogik online, Ausgabe Nr. 8 / Juli 2005.

Fischer, G. H. & Molenaar, I. W. (1995). Rasch models – Foundations, recent developments, and applications. New York: Springer.

Greeno, J. G. (1997). On claims that answer the wrong questions. Educational Researcher, 5-21.

Hartig, J. & Jude, N. (2007). Empirische Erfassung von Kompetenzen und psychometrische Kompetenzmodelle. In Hartig, J. & Klieme, E. (Hrsg.), Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzmodelle. Herausgegeben vom Bundesministerium für Bildung und Forschung (BMBF). Band 20. Bonn, Berlin, 17-36.

Hartig, J. (2004). *Methoden der Skalierung und Definition von Kompetenzniveaus.* Vortrag auf der DESI-Fachtagung „Konzeptualisierung und Messung sprachlicher Kompetenzen", 09. und 10. September 2004, DIPF.

Hartig, J. (2007). Skalierung und Definition von Kompetenzniveaus. In Beck, B. & Klieme, E. (Hrsg.) (2007). Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International). Weinheim und Basel: Beltz, 83-99.

Kirsch, I., Jungeblut, A. & Mosenthal, P. B. (1998). The measurement of adult literacy. In Murray, T. S., Kirsch, I. & Jenkins, L. (Eds.), Adult literacy in ECD countries: Technical report on the First International Adult Literacy Survey. Washington, DC: U.S. Department of Education, National Center for Eudcation Statistics.

Klieme E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.-E. & Vollmer, H. J. (2003). Zur Entwicklung nationaler Bildungsstandards. Eine Expertise. Frankfurt am Main. DIPF.

Klieme, E., Baumert, J., Köller, O. & Bos, W. (2000): Mathematische und naturwissenschaftliche Grundbildung: Konzeptuelle Grundlagen und die Erfassung und Skalierung von Kompetenzen. In Baumert, J., Bos, W. & Lehmann, R. (Hrsg.), TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Grundbildung am Ende der Schullaufbahn. Band 1. Opladen: Leske + Budrich, 85-133.

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

Lehmann, R. & Hunger, S. (2007). ULME III: Anlage und Durchführung der Untersuchung. In Lehmann, R. & Seeber, S. (Hrsg.), ULME III. Untersuchung von Leistungen, Motivation und Einstellungen der Schülerinnen und Schüler in den Abschlussklassen der Berufsschulen. Behörde für Bildung und Sport der Freien und Hansestadt Hamburg. Hamburger Institut für berufliche Bildung HIBB), Hamburg, 21-40.

Lehmann, R. H., Seeber, S. & Hunger, S. (2006). Untersuchung der Leistungen, Motivation und Einstellungen von Schülerinnen und Schülern in den Abschlussklassen der teilqualifizierenden Berufsfachschulen (ULME II). Behörde für Bildung und Sport der Freien und Hansestadt Hamburg (Hrsg.).

Lehmann, R., Seeber, S. & Hunger, S. (2007). ULME III – Ziele der Untersuchung. In Lehmann, R. & Seeber, S. (Hrsg.), ULME III. Untersuchung von Leistungen, Motivation und Einstellungen der Schülerinnen und Schüler in den Abschlussklassen der Berufsschulen. Behörde für Bildung und Sport der Freien und Hansestadt Hamburg. Hamburger Institut für berufliche Bildung HIBB), Hamburg, 15-20.

Metzger, C., Waibel, R., Henning, C., Hodel, M. & Luzi, R. (1993). Anspruchsniveau von Lernzielen und Prüfungen im kognitiven Bereich. IWP St. Gallen.

Minnameier, G. (2006). Aspekte von 'Fachkompetenz' – Kognitive Leistungen im Umgang mit Wissen. In Minnameier, G. & Wutttke, E. (Hrsg.), Berufs- und wirtschaftspädagogische Grundlagenforschung. Lehr-Lern-Prozesse und Kompetenzdiagnostik. Festschrift für Klaus Beck. Frankfurt am Main et al.: Peter Lang, 391-405.

Preiss, P. (1999). Didaktik des wirtschaftsinstrumentellen Rechnungswesens. München u.a.: Oldenbourg.

Preiss, P. & Tramm, T. (1996). Die Göttinger Unterrichtskonzeption des wirtschaftsinstrumentellen Rechnungswesens. In Preiss, P.& Tramm, T. (Hrsg.), Rechnungswesenunterricht und ökonomisches Denken. Wiesbaden: Gabler, 222-323.

Prenzel, M., Drechsel, B. & Kramer, K. (1998). Lernmotivation im kaufmännischen Unterricht: Die Sicht von Auszubildenden und Lehrkräften. Zeitschrift für Berufs- und Wirtschaftspädagogik, Beiheft 14, 169-187.

Prenzel, M., Kramer, K. & Drechsel, B. (2001). Selbstbestimmt motiviertes und interessiertes Lernen in der kaufmännischen Erstausbildung. In Beck, K. & Krumm, V. (Hrsg.), Lehren und Lernen in der beruflichen Erstausbildung. Opladen: Leske + Budrich, 37-61.

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

Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? Journal of Educational Psychology, 91(1), 175-189.

Rost, J. (2004). Lehrbuch Testtheorie – Testkonstruktion, Zweite, vollständig überarbeitete und erweiterte Auflage. Bern, Göttingen: Hans Huber.

Schneider, M. (2006). Konzeptuelles und prozedurales Wissen als latente Variablen: Ihre Interaktion beim Lernen mit Dezimalbrüchen. Dissertation. Technische Universität Berlin. http://deposit.ddb.de/cgibin/dokserv?idn=978818768&dok_var=d1&dok_ext=pdf&filename=978818768.pdf. [aufgerufen am 01.11.2007].

Seeber, S. (2005). Zur Erfassung und Vermittlung berufsbezogener Kompetenzen im teilqualifizierenden Bildungsgang „Wirtschaft und Verwaltung" an Hamburger Berufsfachschulen. In Brand, W. & Tramm, T. (Hrsg.), Prüfungen und Standards in der beruflichen Bildung. In bwp@ - Berufs- und Wirtschaftspädagogik online, Heft 8/2005. http://www.bwpat.de/ausgabe8/seeber_ bwpat8.shtml.

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

Seeber, S. (2007b). Zur Anforderungsstruktur eines Fachleistungstests für Auszubildende des Ausbildungsberufs Einzelhandelskaufmann/Einzelhandelskauffrau. In Münk, D., van Buer, J., Breuer, K. & Deissinger, T. (Hrsg.), Hundert Jahre kaufmännische Ausbildung in Berlin. Schriftenreihe der Sektion Berufs- und Wirtschaftspädagogik der Deutschen Gesellschaft für Erziehungswissenschaft (DGfE). Opladen & Farmington Hills: Barbara Budrich, 184-193.

Sloane, P. F. E. & Dilger, B. (2005). The Competence Clash – Dilemmata bei der Übertragung des 'Konzepts der nationalen Bildungsstandards' auf die berufliche Bildung. In Brand, W. & Tramm, T. (Hrsg.), Prüfungen und Standards in der beruflichen Bildung. In *bwp@*

- *Berufs- und Wirtschaftspädagogik online*, Heft 8/2005. http://www.bwpat.de/ ausgabe8/ sloane_dilger_ bwpat8.shtml.

Stark, R., Hinkofer, L. & Mandl, H. (2001). Beispielbasiertes Lernen im Bereich Buchführung: Einlfuss instruktionaler Erklärungen und multipler Perspektiven auf Lernverhalten und Lernerfolg. Forschungsbericht Nr. 134. Ludwig-Maximilians-Universität München. Institut für Pädagogische Psychologie und Empirische Pädagogik. München.

Straka, G. A. & Lenz, K. (2005). Bestimmungsfaktoren fachkompetenten Handelns kaufmännischer Berufsschülerinnen und Berufsschüler. Ergebnisse einer unterrichtsbegleitenden Pilotstudie. In Frey, A., Jäger, R. S. & Renold, U. (Hrsg.), Kompetenzdiagnostik – Theorien und Methoden zur Erfassung und Bewertung von beruflichen Kompetenzen. Landau: Verlag Empirische Pädagogik, 98-115.

Tramm, T. & Seeber, S. (2006). Überlegungen und Analysen zur Berufsspezifität kaufmännischer Kompetenz. In Minnameier, G. & Wutttke, E. (Hrsg.). Berufs- und wirtschaftspädagogische Grundlagenforschung. Lehr-Lern-Prozesse und Kompetenzdiagnostik. Festschrift für Klaus Beck. Frankfurt am Main et al.: Peter Lang, 273-288.

Whetton, C., Twist, E. & Sainsbury, M. (1999). National Tests and Target Setting: Maintaining Consistent Standards. Paper presented at American Educational Research Association. 1999 Online unter: www.leeds.ac.uk/educol/documents/00001422.htm, Stand: 11/2007.

Witt, R. (2006). Kompetenzstufenmodelle zur Messung ökonomischer Kompetenz. In Minnameier, G. & Wuttke, E. (Hrsg.), Berufs- und wirtschaftspädagogische Grundlagenforschung. Lehr-Lern-Prozesse und Kompetenzdiagnostik. Frankfurt a. M. u. a.: Peter Lang, 407-419.

Autorenanschrift: Dr. Susan Seeber, Deutsches Institut für Internationale Pädagogische Forschung (DIPF), Warschauer Str. 34-38, 10243 Berlin, E-Mail: seeber@bbf.dipf.de