Forum

ESTHER WINTHER

Kompetenzen messen – Zur Notwendigkeit methodologischer und quantitativer Standards im Rahmen beruflicher Kompetenzmessung[1]

Kommentar zum Beitrag von Musekamp, Spöttl & Becker (2010, ZBW, 106 (3)): Schriftliche Arbeitsaufträge zur Erfassung von Differenzen in der Expertise von Facharbeitern und Auszubildenden

Fragen der Kompetenzmessung in der beruflichen Bildung werden seit dem Versuch, einen internationalen Vergleich beruflicher Kompetenzen in verschiedenen Domänen der Berufsbildung durchzuführen (Initiative VET-LSA; BAETHGE, ACHTENHAGEN, ARENDS, BABIC, BAETHGE-KINSKY & WEBER, 2006), verstärkt im wissenschaftlichen Diskurs aufgegriffen. Einer der zentralen Beiträge dieser Initiative ist es, Assessmentverfahren und Methoden der beruflichen Kompetenzmessung sowie psychometrische Modelle zu deren empirischer Validierung entwickelt zu haben, die im internationalen Vergleich anschluss- und konsensfähig sind. Der Beitrag von MUSEKAMP, SPÖTTL & BECKER (2010) befasst sich im Rahmen dieser neueren Debatte mit den Fragen nach der Natur von Kompetenzen und wie sich diese erfassen lässt. Damit werden offene Forschungsdesiderate der beruflichen Kompetenzmessung aufgegriffen. Die Auseinandersetzung mit den Theorien, Messmodellen sowie Konzepten der Kompetenzmessung erfolgt jedoch auf eine Weise, die nicht unkommentiert bleiben kann. Nachfolgend werden fünf Themengebiete diskutiert, in denen gezeigt wird, dass zentrale, international akzeptierte Konventionen zur Erfassung von Leistung und Kompetenz in diesem Beitrag nicht verstanden und/oder nicht zur Kenntnis genommen werden und dass notwendige quantitative Standards für die Veröffentlichung von Forschungsergebnissen nicht eingehalten worden sind; dabei hätte jeder einzelne der beschriebenen Kritikpunkte die Herausgeber dieser Zeitschrift zu einer Ablehnung des Beitrags motivieren müssen. Nachfolgend werden die Ausführungen (1) zu testtheoretischen Annahmen, (2) zu möglichen Messmodellen, (3) zur Konstruktvalidität und zu Konstruktunvollständigkeiten, (4) zu Kompetenzmodellen und letztlich (5) zur gewählten Methodik und Berichtsqualität kommentiert.

1 Ausführungen zur Testtheorie

Im Beitrag von MUSEKAMP, SPÖTTL & BECKER (2010) wird der Anspruch erhoben, auf Basis einer „explorative[n] Untersuchung im Experten-Novizen-Vergleich mithilfe von offenen,

1 Die konzeptionellen Anmerkungen sind 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) sowie des BMBF-Projekts „Konstruktvalidität von Simulationsaufgaben: Computergestützte Messung berufsfachlicher Kompetenz“ entstanden.

schriftlichen Arbeitsaufträgen“ (ebd., S. 338), Ergebnisse generieren zu können, die mit Verfahren der Item Response-Theorie (IRT) so nicht gewonnen werden können. Hierfür wird ausgeführt, dass Verfahren der IRT konfirmatorischer und damit hypothesenprüfender Natur seien und für offene Forschungsdesiderate – wie etwa dem nach der Natur beruflicher Kompetenz – zunächst Studien explorativer und damit hypothesengenerierender Natur auf Basis der Klassischen Testtheorie sinngebend zu seien hätten. Folgende Richtigstellung scheint an dieser Stelle notwenig: Die Entscheidung für oder gegen eine Testtheorie ist **nicht** abhängig von explorativen oder konfirmatorischen Forschungszugängen.

Mit der Klassischen Testtheorie und der Item Response-Theorie liegen zwei Möglichkeiten vor, wie aus dem Testverhalten auf ein zu untersuchendes Personenmerkmal (latente Variable) geschlossen werden kann. 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 – Bedingungen, die hohe Ansprüche an das verwendete Testinstrumentarium stellen. In der Item Response-Theorie werden anstelle von deterministischen Relationen Wahrscheinlichkeitsannahmen formuliert. Es sind insbesondere zwei Eigenschaften, die die Verwendung von IRT-Modellen sinnvoll erscheinen lassen und die dazu führen, dass es diese Modelle sind, die im Rahmen (inter)-nationaler Vergleichsstudien vorrangig zum Einsatz kommen: (1) IRT-Modelle erlauben es, Personen- und Itemparameter parallel zu skalieren. Das Antwortverhalten auf ein Testitem wird damit ausschließlich durch die Itemschwierigkeit und die Personenfähigkeit bestimmt. Diese Separabilität von Personen- und Itemmerkmalen hat zur Folge, dass die Aussagen über die Personen, die einen 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 worden sind. (2) Personen- und Itemparameter sind unabhängige Variablen, die separat geschätzt werden können, d. h., dass die Personenparameter von der Stichprobe der Items und die Itemparameter von der Stichprobe der Personen unabhängig sind. Nur so sind Vergleiche zwischen Personen sowie Vergleiche zwischen Items zulässig (Stichprobenunabhängigkeit). Für die Praxis von Tests bedeutet dies, dass Unterschiede zwischen den Leistungsfähigkeiten von Personen ohne Störfaktoren geschätzt werden können und dass für eine reliable Festlegung der Personenfähigkeit nur Ausschnitte der latenten Variablen abgebildet werden müssen. Separabilität und Stichprobenunabhängigkeit setzen jedoch vor allem (eine hohe Anzahl) Testitems voraus, die unabhängig von einander sind – auch dies stellt eine hohe Anforderung an das zu verwendende Testinstrumentarium dar.

Unabhängig von diesem Plädoyer für die IRT gibt es aber mindestens zwei gute Gründe, sich im Rahmen beruflicher Kompetenzmessungen auch für Modelle der Klassischen Testtheorie zu entscheiden: Berufliche Kompetenzmessung sollte nicht nur auf curriculare Inhalte und Wissenskategorien begrenzt sein, sondern hat ebenso funktionale Aspekte von Bildung einschließlich verschiedener Verhaltenskomponenten und differenzierter Einstellungen in realen beruflichen Lebenswelten zu erfassen (Straka & Macke, 2010; Winther & Achtenhagen, 2009; Sloane & Dilger, 2005). Diese unstrittige Ausweitung des beruflichen Kompetenzkonstrukts hat für entsprechende Erhebungsverfahren zur Folge, dass Testsituationen komplex und dabei fokussiert auf berufliche Handlungen zu gestalten sind. Konsequenterweise und vorrangig einem gewissen Testpragmatismus geschuldet, bedeutet dies, (1) dass die Anzahl der Testitems sehr oft relativ klein ist. Eine geringe Itemanzahl birgt im Rahmen der IRT die Gefahr, dass keine hinreichend gute Verteilung der Items auf der Logit-Skala erzeugt werden kann oder aber dass sich Bündelungen der Items hinsichtlich ihrer Schwierigkeitsindizes zeigen – beides Effekte, die einer angemessenen Skalierung entgegenstehen. Komplexe, arbeitsauthentische Testsituationen können (2) dazu führen, dass die Testitems nicht unabhängig voneinander sind. Damit wäre die Prämisse der lokalen stochastischen Unabhängigkeit verletzt, die eine wesentliche Voraussetzung für

die Anwendung von IRT-Modellen darstellt. Eine Differenzierung zwischen explorativ und konfirmatorisch angelegten Untersuchungen zählt hingegen nicht unter die Gründe, sich für oder gegen eine Testtheorie zu entscheiden, da beide Testtheorien Messmodelle für konfirmatorische und explorative Untersuchungen zur Verfügung stellen. So kommen konfirmatorische und explorative Modelle der Rasch-Familie im Rahmen der IRT zum Einsatz, genau wie explorative und konfirmatorische Faktorenmodelle und Higher-Order-Modelle vor dem Hindergrund klassischer testtheoretischer Überlegungen.

2 Ausführungen zu Messmodellen

Der Beitrag von MUSEKAMP, SPÖTTL & BECKER (2010) stellt zwei Messmodelle vor: Das eindimensionale Rasch-Modell als „das bisher in der Berufsbildungsforschung am häufigsten eingesetzte probabilistische Testmodell" (ebd., S. 341) sowie ein exploratives Faktorenmodell. Die Ausführungen zum Rasch-Modell sollen – so bleibt es der Vermutung des Lesers überlassen – dazu dienen, die testtheoretische Logik und Konzeption gängiger Kompetenzstruktur- und -niveaumodelle zu erläutern. Als Fazit der Autoren bleibt, dass in frühen Stadien des Erkenntnisprozesses explorative Verfahren – am konkreten Beispiel einer explorativen Faktorenanalyse – einzusetzen seien, solange grundlegende Fragen zur Natur von Kompetenzen in beruflichen Domänen nicht geklärt wären (ebd., S. 347).

Zu diesen Ausführungen scheinen die folgenden Korrekturen die dringlichsten zu sein, wenngleich an dieser Stelle umfangreichere Ausführungen insbesondere hinsichtlich messtheoretischer Inhalte notwendig wären (hierzu sei jedoch auf die einschlägige Literatur verwiesen): In den Projekten, die sich quantitativ mit Fragen der Kompetenzmessung befassen, wird dann und nur dann mit konfirmatorischen Modellen (hier erfolgt bewusst keine Eingrenzung auf das 1-parametrische logistische Modell) gearbeitet, wenn im Vorfeld explorativ-quantitative **und** theoretisch-konzeptionelle Erkenntnisse zu Modellen und damit zur Natur des zu erfassenden Konstrukts vorliegen. National verwiesen sei an dieser Stelle auf die Arbeiten des Deutschen PISA-Konsortiums, auf das aktuelle Schwerpunktprogramm der DFG sowie auf die Validierungsstudien im gewerblich-technischen sowie im kaufmännischen Bereich der Berufsbildung im Zuge der VET-LSA-Initiative. Die Wahl des konfirmatorischen Messmodells ist dabei abhängig (a) von der Wahl der Testtheorie, (b) von der angenommenen Relation zwischen dem Anforderungscharakter der Testitems einerseits und der Fähigkeitsstruktur des Probanden andererseits sowie (c) vom Skalenniveau. Dies hat zur Konsequenz, dass in komplexen beruflichen Handlungskontexten im Rahmen der IRT statt 1-parametrischer logistischer Modelle (Raschmodell, das auf dichotomen Antwortkategorien basiert) 2- bzw. 3-parametrische logistische Modelle zum Einsatz kommen, die zur Prüfung der „Natur von Kompetenz" mit multidimensionalen IRT-Modellen (MIRT) der Rasch-Familie verglichen werden (u. a. das Multidimensional Random Coefficients Multinomial Logit Model (MRCMLM; ADAMS, WILSON & WANG, 1997; BRIGGS & WILSON, 2004) für die Extraktion mit dem Anwendungsprogramm Conquest oder das Multicomponent Latent Trait Model (MLTM; EMBRETSON, 1985) für die Extraktion mit dem Anwendungsprogramm MPlus).

Tabelle 1: Überblick über konfirmatorische Messmodelle im Rahmen der Klassischen Testtheorie und der Item Response-Theorie (vgl. HARTIG, 2010)

Latente Variablen	Beobachtete Variablen (Indikatoren der Messung)	
	kontinuierlich	kategorial
kontinuierlich	Strukturgleichungsmodelle, konfirmatorische Faktorenanalysen	Latent trait-Modelle (Raschmodell, 2- und 3-parametrische logistische Modelle)
kategorial	Mischverteilungsmodelle, Clusteranalysen	Latente Klassen-Analysen
	Klassische Testtheorie	Item Response-Theorie

Eine Behauptung aufzustellen, dass im Bereich der Berufsbildungsforschung keine Vorstudien mit explorativ-quantitativen Befunden zu den Fragen vorlägen, was Auszubildende und Arbeitnehmer auf Arbeitsmärkten und in spezifischen beruflichen Anwendungssituationen können und leisten müssen, ist fast als polemisch zu klassifizieren (vgl. u. a. MERTENS, 1974; REETZ, 1989; 1999; BECK, 1989; ACHTENHAGEN, 2004; SLOANE & DILGER, 2005; LEHMANN & SEEBER, 2007; WINTHER, 2010A; WINTHER, 2010b). Auch unter Rückgriff auf Messmodelle der IRT kann man Strukturen innerhalb der Daten suchen und Zusammenhänge zwischen den Daten beschreiben. Als ein illustrierendes Beispiel sei hier auf das explorative mehrdimensionale Rasch-Modell von MCKINLEY & RECKASE (1982; zit. n. EMBRETSON & REISE, 2001) verwiesen:

$$P(X_{vi}=1\,|\,\theta_{vm},\delta_i)=\frac{\exp(\sum_m \theta_{vm}+\delta_i)}{1+\exp(\sum_m \theta_{vm}+\delta_i)}$$

Die Gleichung weist im Vergleich zum eindimensionalen Rasch-Modell folgende Veränderungen auf: (1) Im Hinblick auf den Itemparameter wird im mehrdimensionalen Raschmodell nicht die Schwierigkeit, sondern die Leichtigkeit des Items abgebildet. Dies drückt sich dadurch aus, dass die Relation zwischen den Personen- und Itemparametern nicht durch eine Differenz, sondern durch eine Summe beschrieben wird. (2) Anstatt des Personenparameters θ_v werden im mehrdimensionalen Modell mehrere gleich gewichtete Fähigkeitsparameter für m verschiedene Dimensionen zusammengestellt. θ_{vm} beschreibt dann folglich den Fähigkeitsparameter der Person v in der Dimension m. Diese Aufgliederung stellt sicher, dass nur dann, wenn die multiplen Dimensionen nicht in jedem Item gleich gewichtet sind, eine mehrdimensionale Struktur aufgezeigt werden kann. Bei einem gleichen Gewicht ließen sich für die verschiedenen Dimensionen keine Unterschiede im Antwortverhalten der Probanden abbilden. Dies führte dazu, dass die verschiedenen Fähigkeitsparameter der m Dimensionen nicht separat geschätzt werden können und damit das Modell nicht passt. Als Ergebnis erfolgt vergleichbar zur explorativen Faktorenanalyse eine Dimensionsreduktion, die einer inhaltlich-interpretativen Beschreibung bedarf (für weitere explorative Beispiele im Rahmen der IRT vgl. u. a. WILSON, DE BOECK & CARSTENSEN, 2008).

3 Ausführungen zu Konstruktunvollständigkeiten

MUSEKAMP, SPÖTTL & BECKER (2010) schließen ihre Ausführungen zum eindimensionalen Rasch-Modell damit, dass (1) die Wahl des Messmodells zu „Konstruktunvollständigkeiten" führt und dass (2) diese „Konstruktunvollständigkeiten" auf der Grundlage des Rasch-Modells

systematisch ausgeblendet würden (ebd., S. 344). Hierzu in Kürze nachfolgender Kommentar: Insgesamt kann man sicherlich feststellen, dass explorative Modelle für kognitive Tests im Rahmen der IRT kaum zum Einsatz kommen. Dies liegt aber insbesondere daran, dass im Vorfeld von Assessments detaillierte, kognitiv determinierte Überlegungen darüber angestellt werden, wie ein Test und die entsprechenden Items zu konstruieren seien, um ein spezifisches Konstrukt zu erfassen. D. h. im Rahmen moderner Assessmentmodelle wird zunächst stets die Frage nach der Konstruktvalidität geklärt. Den Annahmen im Beitrag von MUSEKAMP, SPÖTTL & BECKER (2010), dass die vorliegenden Kompetenzmessverfahren, da sie bislang nur sehr einheitlichen Stichproben vorgelegt wurden, Konstrukte unvollständig abbildeten oder überhaupt nicht angemessen zu erfassen in der Lage wären (ebd., S. 344), fehlt jegliche empirische Evidenz; die gezogene Schlussfolgerung, dass berufliche Situationen – am konkreten Beispiel schriftlicher Arbeitsaufträge – nicht nur den Auszubildenden (Novizen: für die betreffende Situation als weniger kompetent angenommen), sondern auch Facharbeitern (Experten: für die betreffende Situation als kompetenter eingeschätzt) vorzulegen wären, um so zu einer Ausweitung des Konstrukts zu kommen, wird empirisch nicht geprüft. Festzuhalten bleibt in diesem Zusammenhang jedoch, dass es definitiv nicht das gewählte Messmodell sein kann, das fehlende Konstruktvalidität bedingt, sondern die fehlende theoriebasierte Auseinandersetzung bezüglich der situationsangemessenen Passung zwischen Test, Proband und zu erfassendem Personenmerkmal. Hier wird von MUSEKAMP, SPÖTTL & BECKER (2010) als „alternative Strategie" (ebd., S. 346) vorgeschlagen, sich dem „Kompetenzkonstrukt von einem möglichst weit fortgeschrittenen *empirischen* Niveau aus zu nähern" (ebd., S. 346), indem (1) nicht normativ formulierte Lernziele oder Bildungsstandards das Konstrukt determinieren und indem (2) die Testsituationen auch beruflichen Experten vorzulegen sind, um so mögliche Entwicklungsstufen zu erfassen. Bei der unter (1) und (2) beschriebenen „alternativen Strategie" ist den Autoren entgangen, dass beispielsweise im Rahmen der Validierungsstudien der VET-LSA-Initiative sowohl im gewerblich-technischen Bereich (NICKOLAUS, GSCHWENDTNER & ABELE, 2009) als auch im kaufmännischen Bereich (ACHTENHAGEN & WINTHER, 2009; WINTHER, 2010a) keine curriculare Validität angestrebt, sondern ein auf berufliche Tätigkeiten abzielendes, sich an der betrieblichen Authentizität messen lassendes Validitätskonzept gewählt wurde, dass zudem (a) national von betrieblichen Experten bewertet und (b) international im Rahmen einer ergänzenden Feasibility-Study (BAETHGE & ARENDS, 2009) auf Konsensfähigkeit geprüft wurde. Das Kontinuum des theoretisch ausgearbeiteten und empirisch validierten Kompetenzkonstrukts beispielsweise im kaufmännischen Bereich variiert von kaufmännischem Grund- und Regelwissen einschließlich Routinetätigkeiten bis zu kaufmännischem Entscheidungshandeln (u. a. WINTHER & ACHTENHAGEN, 2009; 2010). Das Kompetenzkonstrukt erfasst damit Anforderungen und Fähigkeiten, die weit über die Ansprüche der Rahmenlehrpläne und Ausbildungsordnungen hinausgehen, und bildet stattdessen die Beanspruchungsmuster kaufmännischer Angestellter in den verschiedenen Unternehmensbereichen und -positionen ab.

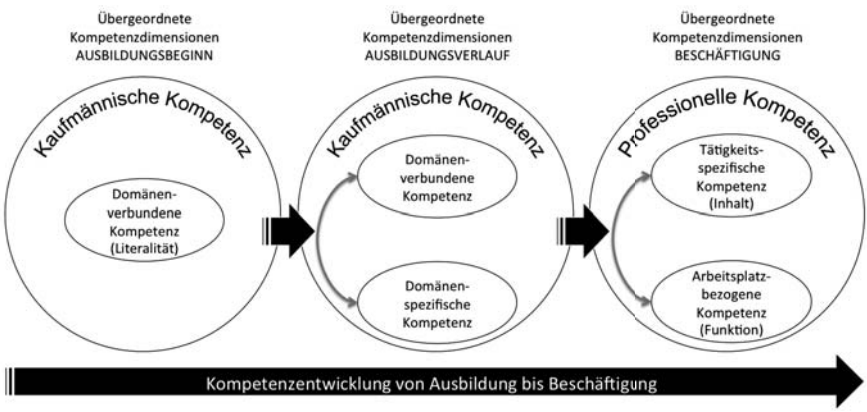
International konsensfähig ist darüber hinaus die Vorstellung, dass Assessments – in welchem Bereich sie auch immer eingesetzt werden – in kongruenter Beziehung zu den Zielen, spezifischen Inhalten und Arbeitshandlungen sowie -techniken eines Faches oder einer Domäne stehen müssen. Oder anders formuliert: Assessments sind an den Lern- und Arbeitsprozessen zu orientieren, die typisch für eine bestimmte Domäne sind (u. a. PELLEGRINO, CHUDOWSKY & GLASER, 2001). In diesem Zusammenhang lässt die gewählte Testform in der von MUSEKAMP, SPÖTTL & BECKER (2010) berichteten Untersuchung zur Kompetenz von Kfz-Servicemechanikern Zweifel an der Gültigkeit des gewählten Verfahrens zu, wenn die „Kompetenzmessung" darin bestand, dass „die Probanden ... gebeten [wurden], diese Aufgaben in drei Schulstunden schriftlich in Aufsatzform zu lösen" (ebd., S. 348).

4 Ausführungen zu Kompetenzmodellen

Im Beitrag von MUSEKAMP, SPÖTTL & BECKER (2010) werden Ontologie und Epistemologie von Kompetenzmodellen skizziert. Es wird festgehalten, dass sich auf Grundlage epistemologischer Annahmen mithilfe des Rasch-Modells statistisch Kompetenzstruktur und -niveaumodelle konstruieren lassen (ebd., S. 342). Hierzu sind die folgenden Korrekturen notwendig: Kompetenzmodelle integrieren qualitative und quantitative Beschreibungen von inter-individuellen sowie intra-individuellen Unterschieden auf verschiedenen Abstraktionsebenen. Die Konstruktion von Kompetenzstrukturen bzw. Dimensionen von Kompetenz erfolgt dabei **qualitativ** im Hinblick auf inhaltlich-semantisch interpretierbare Variablen, mit deren Hilfe sich Kompetenzunterschiede zwischen Personen als Merkmalsträger von Kompetenz beschreiben lassen. Für die Beschreibung inter-individueller Unterschiede stehen **quantitative** Kontinua der einzelnen Teildimensionen als empirische Lösung zur Verfügung, die auf Basis des gewählten Messmodells gewonnen wurden. Das Messmodell dient damit nicht der Konstruktion, sondern stellt die mathematische Relation zur Verfügung, mit der die Passung einer angenommenen Modellstruktur auf die vorliegenden empirischen Daten statistisch getestet werden kann. Dabei ist der Test der Modellgüte auf den Vergleich von Modellen gleicher Familien beschränkt (z. B. g-Faktor-Modell vs. Nested-Faktor-Modell oder eindimensionales vs. mehrdimensionales IRT-Modell).

Unter Rückgriff auf Ontologie und Epistemologie von Kompetenzmodellen argumentieren die Autoren weiter, dass im Hinblick auf die derzeitigen Modelle keine Aussagen bezüglich der zeitlichen Stabilität der Modelle getroffen werden könnten und dass für die vorliegenden Modelle keine Entwicklungsperspektive gälte. Hier böte – so der Fortgang der Argumentation – der Ansatz einer Experten-Novizen-Forschung Lösungsmöglichkeiten. Folgende Korrekturen und Ergänzungen sind in diesem Zusammenhang von Bedeutung: Kompetenzentwicklung kann sich als Veränderung der Kompetenzstruktur sowohl im Sinne einer Differenzierung nach Teilkompetenzen als auch im Sinne einer Integration verschiedener Teilkompetenzen zeigen – hierzu existieren qualitative wie quantitative Befunde, die sich wiederum in Kompetenzmodelle integrieren lassen (vgl. SCHECKER & PARCHMANN, 2006). Ein Beispiel für diese Sichtweise ist WINTHER (2010a) entnommen (Abbildung 1):

Abbildung 1: Veränderung der Kompetenzstruktur im Verlauf professioneller Entwicklung (WINTHER, 2010a, S. 259)



Für die unterschiedlichen Stationen der professionellen Entwicklung von kaufmännischen Auszubildenden und Angestellten lassen sich drei verschiedene theoriebasierte, qualitativ begründete Kompetenz(struktur)modelle unterscheiden, die im Sinne einer Entwicklungsperspektive im Rahmen beruflicher Karrieren aufeinander bezogen werden können. Die Veränderung der Kompetenzstrukturen in den verschiedenen Entwicklungsstadien berücksichtigt, dass sich im Zuge der beruflichen Kompetenzentwicklung Strategien der Anforderungsbewältigung herausbilden, die von dem einzelnen Auszubildenden/Arbeitnehmer als effektiv bewertet werden und die in der Folge zu anderen Kompetenzkonstrukten führen. Während im Verlauf der Ausbildung eine Unterscheidung in domänenspezifische und domänenverbundene Fähigkeitsstrukturen sinnvoll ist, kann diese Trennung im Verlauf der beruflichen Kompetenzentwicklung im Rahmen eines Beschäftigungsverhältnisses durch eine Kategorie abgelöst werden, die als professionelle Kompetenz überwiegend über tätigkeitsspezifische und arbeitsplatzbezogene Faktoren charakterisiert werden kann. Im Zentrum der von MUSEKAMP, SPÖTTL & BECKER (2010) zitierten Studien aus dem allgemeinen (TIMSS, PISA) und beruflichen Bildungsbereich (NICKOLAUS, GSCHWENDTNER & GEISSEL, 2008; LEHMANN & SEEBER, 2007) stand jedoch die konzeptionelle sowie empirische Feststellung von Leistungsständen in spezifischen Domänen und nicht die Abbildung von Entwicklungsprozessen. Die Darstellung der Autoren ist darüber hinaus als unvollständig und verkürzt zu bewerten, wenn die Kompetenzentwicklung an die Veränderung sowie Interpretation von Kompetenzniveaus bzw. Kompetenzstrukturen gebunden wird. MUSEKAMP, SPÖTTL & BECKER (2010) stellen in diesem Zusammenhang fest, dass mit den derzeitigen Verfahren der Kompetenzfeststellung verborgen bliebe, "ob das Erreichen eines bestimmten Kompetenzniveaus in den früheren Phasen eines Kompetenzentwicklungsverlaufes auch funktional für das Erreichen von Kompetenzqualitäten in späteren Phasen der Kompetenzentwicklung ist. Der Blick auf die Kompetenzstrukturen allein über kurze Etappen der Entwicklungszeit kann nicht verlässlich Auskunft darüber geben, was Kompetenz in der höchsten Ausprägung und in ihrem Wesen ausmacht" (ebd., S. 344). Nicht nur, dass der Leser hier eine begriffliche Unschärfe zwischen Kompetenzstruktur und -niveau zu verarbeiten hat; hier werden zusätzlich Annahmen der Kompetenzentwicklung als nomologische Erweiterungen eingeführt. Ziel von Kompetenzentwicklungsmodellen ist es jedoch, Kompetenzniveaus auf den Kontinua der einzelnen Teilkompetenzen nicht allein aus der Analyse von Anforderungssituationen abzuleiten (Kompetenzniveaumodell), sondern die Niveaus als entwicklungspsychologisch begründete qualitative Unterschiede zu interpretieren (u. a. IPN, 2003). Die theoretische und empirische Fundierung von Kompetenzentwicklungsmodellen umfasst damit weit mehr, als die Herausbildung und die Entwicklung von kognitiven Strukturen im Experten-Novizen-Vergleich zu beschreiben. Der Anmerkung der Autoren, die Expertiseforschung sei „im Rahmen der modernen Kompetenzforschung sehr stark in den Hintergrund getreten" (MUSEKAMP, SPÖTTL & BECKER, 2010, S. 347), ist entgegenzuhalten: Aus gutem Grund! Zentral für die Wahl anderer Forschungstheorien als Referenzmodelle für die Kompetenzforschung ist der Fakt, dass sich Unterschiede zwischen Experten und Novizen nur in sehr ausdifferenzierten Domänen und im Hinblick auf klar definierte Anforderungssituationen zeigen. Wenn kognitive Leistungen jedoch in wenig strukturierten Domänen oder unzureichend definierten Anforderungssituationen (Normalfall für komplexe berufliche Situationen) erbracht werden müssen, zeigen sich keine klar interpretierbaren Ergebnisse der Experten-Novizen-Forschung (u. a. ALEXANDER, 1992). Die Unterschiede zwischen Experten und Novizen sind insbesondere dann nicht mehr stringent, wenn die zu bearbeitenden Aufgaben keine eindeutigen Lösungen verlangen oder wenn die zu zeigende Leistung Elemente eines spezifischen, gesellschaftlich geteilten Allgemeinwissens enthält (u. a. ALEXANDER, SCHALLERT & HARE, 1991) – beides sind Annahmen, die im Hinblick auf die Literacy-Konzeption der internationalen Schulleistungsvergleiche, aber auch hinsichtlich beruflicher Kompetenzerfassung relevant werden.

5 Ausführungen zur Methodik und zum Ergebnisbericht

Mittels explorativer Faktorenanalysen versuchen MUSEKAMP, SPÖTTL & BECKER (2010) die Frage zu beantworten, welche Unterschiede zwischen den Aufgabenlösungen von Personen verschiedener Kompetenzniveaus bestehen – mit Kompetenzniveaus sind hier Unterschiede zwischen Personen verschiedener beruflicher Entwicklungsstadien (Auszubildende vs. Facharbeiter) und **nicht** empirisch ermittelte Graduierungen auf den Kontinua der Teilkompetenzen gemeint. Die Analyse dieser Unterschiede soll Hinweise darauf geben, „inwieweit sich das Kompetenzniveau bzw. worin sich die Kompetenzstruktur von Novizen und Experten im Kfz-Service unterscheiden" (ebd., 348; im Original kursiv). Als Testinstrument werden schriftliche Arbeitsaufträge eingesetzt. Die Lösung der Arbeitsaufträge wird auf Basis von 31 Indikatoren (exogene Kriterien) durch Berufspädagogen der Fachrichtung Fahrzeugtechnik sowie durch erfahrene Facharbeiter (Kfz-Meister) bewertet.

Die Darstellung der Untersuchung beginnt mit der Feststellung, dass das eingesetzte Testinstrument „keiner umfassenden methodischen Prüfung unterzogen werden konnte" (ebd., 348). Derartige Aussagen disqualifizieren. Es ist für quantitative Forschungsarbeiten unumgänglich, Hinweise auf die Methodik der Instrumentenentwicklung sowie auf die Güte des Instruments zu geben. Hierzu gehören beispielsweise Angaben zu den Skalierungen der als Bewertungsmaßstab eingesetzten Indikatoren sowie die Anzahl der Experten, die die schriftlichen Arbeitsaufträge entsprechend der gegebenen Indikatoren bewertet haben. Dies ist für die von MUSEKAMP, SPÖTTL & BECKER (2010) berichtete Untersuchung um so dringlicher, da nicht die Lösungen der Probanden, sondern die Einschätzungen der Experten als empirische Daten in die Analyse einflossen. Wird diese Gegebenheit mit der Aussage verbunden, dass die Inter-Rater-Reliabilitäten nur in wenigen Fällen überprüft werden konnten, und für diese Fälle völlig unakzeptable Werte zwischen 0,41 und 0,52 (gewichtetes Kappa) annahmen (ebd., S. 351), muss zwangsläufig die Frage gestellt werden, wie reliabel und valide das Instrument eigentlich ist. Hinzukommt, dass ungeklärt ist, welche Daten dann letztlich explorativ extrahiert worden sind. Die Datenerzeugung auf Basis der Dokumentation der Aufgabenbearbeitung bleibt somit unklar. Es ist zu fragen, was inhaltlich wirklich gemessen wurde (Konstruktvalidität): Sind es die Qualitäten, mit deren Hilfe eine nicht genau bezeichnete Menge von Experten respektive Facharbeitern Handlungsunterschiede dokumentieren können, oder Handlungsqualitäten, die zwischen Novizen und Experten respektive zwischen Auszubildenden und Facharbeitern differenzieren. Für letzteres wäre es notwendig, die schriftlichen Arbeitsaufträge auch beiden Gruppen vorzulegen. Dies ist jedoch nur am Beispiel eines Falls geschehen. Oder anders formuliert: Es wird versucht, auf Basis einer einzigen offenen Aufgabe, Differenzen in den Kompetenzstrukturen von Experten und Novizen zu identifizieren, wobei diese Differenzen aus exogenen Beurteilungen zu extrahieren sind.

MUSEKAMP, SPÖTTL & BECKER (2010) stellen zwei faktorenanalytische Prüfungen vor. Die dem Design und der Fragestellung der Autoren am ehesten angemessene Analyse, ist diejenige des ersten Arbeitsauftrages, da hier das Instrument von allen zu untersuchenden Teilgruppen bearbeitet wurde. Für diesen Fall (Vorgehen bei einer Inspektion; ebd., S. 354) wurden aus den 31 Bewertungsindikatoren des Analyserasters 13 Items extrahiert, die auf zwei inhaltlich interpretierbaren Faktoren laden. Für die restlichen 58 Prozent des eingesetzten Analyserasters lässt sich folglich keine inhaltlich-semantische Struktur in den Daten finden. Damit liegen quantifizierbare Zweifel an der Inhaltsvalidität des Instruments vor: Der inhaltlich-semantische Bereich des zu erfassenden Konstrukts lässt sich mit Hilfe des gewählten Verfahrens einschließlich der zugehörigen Indikatoren für die gegebene Stichprobe nicht abbilden.

Insgesamt zeigt der Ergebnisbericht der von MUSEKAMP, SPÖTTL & BECKER (2010) durchgeführten Untersuchung eklatante Mängel: So fehlen Ausführungen zur Güte der Faktorenanalyse des ersten Arbeitsauftrages, Reliabilitätskennwerte der erzeugten Skalen

und entsprechende Interpretationen, eine Darstellung der Reliabilitäten im Vergleich zu anderen Studien, in denen das gewählte Analyseraster eingesetzt wurde etc. Hinzukommt, dass die Analyse der differenziellen Unterschiede zwischen Experten und Novizen nicht auf statistische Verfahren zurückgreift und damit nicht geklärt werden kann, ob die deskriptiven Unterschiede zwischen den z-transformierten Faktorenwerten zufällig oder statistisch signifikant sind. Die Untersuchung zeigt jedoch deutlich, dass sich weder das eingesetzte Instrument noch das gewählte methodische Verfahren zur Abbildung oder gar Prüfung von Kompetenzen eignen oder zu einer explorativen Klärung der Natur von Kompetenz beitragen können. Es wird im Gegenteil die Gefahr verstärkt, das zu messende Konstrukt durch die Auswahl von spezifischen Beurteilungskriterien auf Basis faktorenanalytischer Extraktion unangemessen zu reduzieren bzw. zu verändern. In diesen Zusammenhängen dann von Kompetenzstrukturen, Graduierungen von Kompetenzen oder sogar von Kompetenzentwicklungen zu sprechen, ist in hohem Maße unangemessen und wird dem aktuellen Forschungsstand im Bereich der Kompetenzmessung und Kompetenzmodellierung in keiner Weise gerecht.

Literatur

Achtenhagen, F. (2004). Prüfung von Leistungsindikatoren für die Berufsbildung sowie zur Ausdifferenzierung beruflcher 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.

Achtenhagen, F. & Winther, E (2009). *Konstruktvalidität von Simulationsaufgaben: Computergestützte Messung berufsfachlicher Kompetenz – am Beispiel der Ausbildung von Industriekaufleuten*. Abschlussbericht zur Einreichung beim BMBF. Professur für Wirtschaftspädagogik der Georg-August Universität Göttingen.

Adams, R. J., Wilson, M. & Wang, W.-C. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement, 21* (1), 1–23.

Alexander, P. A. (1992). Domain Knowledge: Envolving Themes and Emerging Concerns. *Educational Psychologist, 27* (1), 33–51.

Alexander, P. A., Schallert, D. L. & Hare, V. C. (1991). Coming to Terms: How Researchers in Learning and Literacy Talk about Knowledge. *Review of Educational Research, 61*, 315–343.

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.

Briggs, D. & Wilson, M. (2004). An Introduction to Multidimensional Measurement using Rasch-Models. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch-Measurement* (pp. 322–342). Maple Grove, MN: JAM Press.

Beck, K. (1989). "Ökonomische Bildung" – Zur Analyse eines wirtschaftspädagogischen Begriffs. *Zeitschrift für Berufs- und Wirtschaftspädagogik, 85,* 244-265.

Embretson, S. E. (1985). Multicomponent Latent Trait Models for Test Design. In S. E. Embretson (Ed.), *Test Design. Developments in Psychology and Psychometrics* (pp. 195–218). Orlando, FL: Academic Press.

Embretson, S. E. & Reise, S. P. (2001). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Hartig, J. (2010). *Mehrdimensionale IRT-Modelle.* Methodenworkshop im Rahmen des DFG-Schwerpunktprogramms "Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen". Berlin,24.Februar.

IPN (Leibniz-Institut für die Pädagogik der Naturwissenschaften) (Hrsg.) (2003). *Ziele und Perspektiven naturwissenschaftlicher Bildung*. Kiel: IPN.

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

Mertens, D. (1974). Schlüsselqualifikationen: Thesen zur Schulung für eine moderne Gesellschaft. *Mitteilungen aus Arbeitsmarkt- und Berufsforschung, 7*, 36–43.

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.

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.

Nickolaus, R., Gschwendtner, T. & Geißel, B. (2008): Modellierung und Entwicklung beruflicher Fachkompetenz in der gewerblich-technischen Erstausbildung. *Zeitschrift für Berufs- und Wirtschaftspädagogik, 104*, 48–73.

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.

Reetz, L. (1989b). Zum Konzept der Schlüsselqualifikationen in der Berufsbildung. *Berufsbildung in Wissenschaft und Praxis, 6*, 24–30.

Reetz, L. (1999). Zum Zusammenhang von Schlüsselqualifikationen – Kompetenzen – Bildung. In T. Tramm, D. Sembill, F. Klauser & E. G. John (Hrsg.), *Professionalisierung kaufmännischer Berufsbildung* (S. 32–51). Frankfurt: Peter Lang.

Schecker, H. & Parchmann, I. (2006). Modellierung naturwissenschaftlicher Kompetenz. *Zeitschrift für Didaktik der Naturwissenschaften, 12*, 45-66.

Sloane, P. F. E. & Dilger, B. (2005). The Competence Clash – Dilemmata bei der Übertragung des 'Konzepts der nationalen Bildungsstandards auf die berufliche Bildung. *bwp@ – Berufs- und Wirtschaftspädagogik online, 8.*

Straka, G. A. & Macke, G. (2010). Kompetenz – nur eine "kontextspezifische cognitive Leistungsdisposition"? *Zeitschrift für Berufs- und Wirtschaftspädagogik, 106* (3), 444–451.

Wilson, M., De Boeck & Carstensen, C. (2008). Explanatory item response models: A brief introduction. In J. Hartig, E. Klieme, & D. Leutner, D. (Eds.), *Assessment of competencies in educational contexts* (pp. 91–120). Göttingen: Hogrefe & Huber.

Winther, E. (2010a). *Kompetenzmessung in der beruflichen Bildung*. Bielefeld: wbv.

Winther, E. (2010b). Systemorientierte Evaluation des Lernfeldkonzepts: Strategien für das berufliche Übergangssystem. *Zeitschrift für Berufs- und Wirtschaftspädagogik, 106* (3), 427–442.

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 Mehrdimensionalitat beruflicher Handlungskompetenz. *Berufsbildung in Wissenschaft und Praxis, 1/2010*, 18–21.

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