Musekamp, Frank

Validierung eines Multiple-Choice-Instruments zur Erfassung von Kompetenzen in der Domäne Kfz-Service & Reparatur

Faßhauer, Uwe [Hrsg.]; Fürstenau, Bärbel [Hrsg.]; Wuttke, Eveline [Hrsg.]: Grundlagenforschung zum Dualen System und Kompetenzentwicklung in der Lehrerbildung. Opladen ; Berlin ; Farmington Hills, Mich. : Verlag Barbara Budrich 2011, S. 103-115. - (Schriftenreihe der Sektion Berufs- und Wirtschaftspädagogik der Deutschen Gesellschaft für Erziehungswissenschaft (DGfE))

Grundlagenforschung zum Dualen System und Kompetenzentwicklung in der Lehrerbildung

Uwe Faßhauer
Bärbel Fürstenau
Eveline Wuttke (Hrsg.)

Grundlagenforschung zum Dualen System und Kompetenzentwicklung in der Lehrerbildung

Verlag Barbara Budrich
Opladen • Berlin • Farmington Hills, MI 2011

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Dieses Buch steht im OpenAccess Bereich der Verlagsseite zum kostenlosen Download bereit (http://dx.doi.org/10.3224/86649461)
Eine kostenpflichtige Druckversion (Printing on Demand) kann über den Verlag bezogen werden. Die Seitenzahlen in der Druck- und Onlineversion sind identisch.

ISBN 978-3-86649-461-9
DOI 10.3224/86649461

Umschlaggestaltung: Umschlaggestaltung: bettina lehfeldt graphic design, Kleinmachnow
Verlag Barbara Budrich, http://www.budrich-verlag.de

Inhaltsverzeichnis

Teil II: Grundlagenforschung zum Dualen System

Validierung eines Multiple-Choice-Instruments zur Erfassung von Kompetenzen in der Domäne Kfz-Service & Reparatur

Frank Musekamp

Einleitung und Problemstellung

Seit Abschluss der Machbarkeitsstudie zu einem Berufsbildungspisa hat sich die Zahl der entwickelten Instrumente zur Erfassung beruflicher Kompetenzen rasant vergrößert. Für den Kfz-Bereich haben bisher Gschwendtner (2008) ein Instrument zur Erfassung des berufsfachlichen Wissens und Nickolaus/Gschwendtner/Abele (2009) ein simulationsbasiertes Verfahren zur fachspezifischen Problemlösefähigkeit vorgelegt und validiert. Im vorliegenden Beitrag wird in Abgrenzung zu Gschwendtner (2008) ein Instrument auf Validität geprüft, welches weniger auf curricular verankertes berufsfachliches Wissen abzielt, sondern auf handlungsnahes Arbeitsprozesswissen (vgl. Spöttl 2009). Da das zugrunde liegende Kompetenzmodell und die zur Anwendung gelangten Kriterien der Testkonstruktion bereits ausführlich beschrieben wurden (Becker 2009; Spöttl/Becker/Musekamp 2011; Spöttl 2011), werden an dieser Stelle die Strategie und die Ergebnisse zur Validierung des Instruments ins Zentrum gerückt.

Die Validierung von Instrumenten zur Kompetenzmessung ist in der beruflichen Bildung mit zwei Herausforderungen konfrontiert:

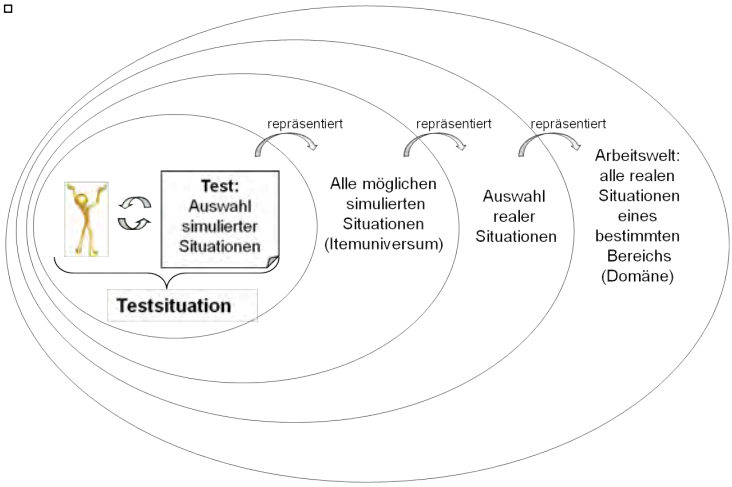
Erstens: Weil Kompetenz durch Lernen in hohem Maße veränderbar ist, muss sichergestellt werden, dass das mit dem Test erfasste Konstrukt entweder die Testpersonen bereits zum Testzeitpunkt in die Lage versetzt, berufliche Aufgaben zu bearbeiten, oder dass das zum Testzeitpunkt erfasste Konstrukt notwendig ist, um diejenigen späteren Stadien der Kompetenzentwicklung zu erreichen, die das berufliche Leistungsverhalten ermöglichen. Insbesondere wenn Tests zu frühen Etappen der Kompetenzentwicklung eingesetzt werden, ist dies nicht unbedingt gegeben. Die Untersuchungen der Gruppe um Mandl verdeutlichen die Problematik der mangelnden Transferierbarkeit von trägem Wissen (im Überblick siehe Gruber/Mandl/Renkl 2000). Musekamp/Spöttl/Becker sprechen bei diesem Aspekt von Konstruktunvollständigkeit (Musekamp/Spöttl/Becker 2010).

Zweitens: Weil sich das Testverhalten in der beruflichen Kompetenzmessung häufig sehr stark von jenem Leistungsverhalten unterscheidet, welches der Test vorherzusagen versucht, unterliegt der Schluss vom Testverhalten auf das berufliche Leistungsverhalten einiger Unsicherheit. Lesen und An-

kreuzen beispielsweise ist so grundsätzlich verschieden vom Einsatz Kfz-spezifischer Werkzeuge in einem Kfz-Betrieb, dass es nicht leicht ist, sicher zu stellen, dass das Verhalten der Probanden bei der Konfrontation mit einer Auswahl simulierter Situationen (Test) repräsentativ ist für das Verhalten in allen möglichen realen Situationen eines Bereichs (siehe Abb. 1).

Beide Aspekte gelten zwar grundsätzlich auch für nicht-berufliche Domänen, z. B. in der Allgemeinbildung oder für psychologische Tests per se. In der beruflichen Bildung ist die Problematik aber besonders ausgeprägt (vgl. Musekamp 2009).

Abbildung 1: Nötige Schlüsse von der Testsituation zur Menge der realen Situationen, auf die sich das Kompetenzkonstrukt bezieht



Zur Validierung von Tests lassen sich nun Strategien zur *internen* und *externen* Validierung als zwei grundsätzlich verschiedene Ansätze heranziehen. „Ein Test heißt intern valide, wenn sich die Annahmen über das Antwortverhalten anhand der Datenmatrix bestätigen lassen“ (Rost 2004, S. 35). Ein Test ist extern valide, wenn das Testverhalten eine Vorhersage auf das interessierende Verhalten außerhalb der Testsituation erlaubt (vgl. ebd.). Während sich die Definition externer Validität demnach auf die zahlreichen Schlüsse von der Testsituation zur Realsituation bezieht, ist interne Validierung auf das Geschehen zwischen Person und Test konzentriert (vgl. Abb. 1). Borsboom (2005) beschränkt seine Definition von Validität ausschließlich auf den Aspekt, der sich mit Strategien interner Validierung beziffern lässt. Für ihn ist ein Test valide, wenn „the attribute to be measured produces variations in the measurement outcomes” (S. 167). Um die Validität eines Tests zu untermauern, ist dann zu belegen, dass eine Eigenschaft (‚attribute‘) existiert

und dass diese Eigenschaft die Unterschiede in den Testscores hervorruft (vgl. Borsboom/Mellenbergh/van Heerden 2004). Um sicher zu stellen, dass das zu erfassende Konstrukt existiert, ist es notwendig, die Testscores außerhalb des Testgeschehens zu verankern. Die häufig verwendete Strategie, ein Kriterium zu definieren und anschließend den Zusammenhang zwischen Testscore und Kriterium als Validität für den Test heranzuziehen, ist dafür nur bedingt brauchbar, weil der Zusammenhang zwischen beiden Werten nicht als kausal angenommen werden kann (Borsboom 2005).

Im Rahmen der Item-Reponse-Theorie – insb. mithilfe des Raschmodells – besteht jedoch die Möglichkeit, Validität entweder durch den Einbezug von realen Aufgaben oder von kompetenten Personen in die Modellgeltungstests elegant zu untermauern. Dieser Grundidee folgend haben beispielweise Nickolaus/Gschwendtner/Abele (2009) simulierte und reale Aufgaben an einem Kraftfahrzeug gemeinsam skaliert. Da sich nur wenige signifikante Unterschiede in den Schwierigkeiten der realen und simulierten Aufgaben zeigten, konnte geschlossen werden, dass zur Lösung der Simulation annähernd dasselbe Konstrukt notwendig war, wie zur Lösung der realen Aufgabe. Da sich im Raschmodell, Personenfähigkeiten und Itemschwierigkeiten mathematisch analog verhalten, lässt sich das gleiche Prinzip auch mit als kompetent erachteten Personen verwirklichen. Dazu wurden im Rahmen der hier beschriebenen Studie[1] neben Auszubildenden am Ende der Ausbildung auch Facharbeiter mit mindestens zweijähriger Berufserfahrung in die Stichprobe einbezogen.

Zur Abschätzung der internen Validität werden die folgenden Hypothesen geprüft:

I. Die Leistung Serviceaufgaben zu lösen, lässt sich auf ein eindimensionales Konstrukt „Servicekompetenz“ zurückführen.

II. Die Leistung Diagnoseaufgaben zu lösen, lässt sich auf ein eindimensionales Konstrukt „Diagnosekompetenz“ zurückführen.

III. Service- und Diagnosekompetenz sind nicht identisch und bilden jeweils eigenständige Dimensionen.

IV. Die Items verteilen sich je nach Arbeitsprozess, den sie betreffen, auf die vier hypothetischen Kompetenzniveaus

Zur Abschätzung der externen Validität werden die folgenden Hypothesen geprüft:

V. Die Testitems funktionieren für Fachkräfte und Auszubildende in gleicher Weise (kein Differential Item Functioning, siehe Osterlind/Everson 2009)

1 Die Erhebungen fanden im Rahmen der Evaluation des zweijährigen Ausbildungsberufes Kfz-Servicemechaniker statt und wurden vom Ministerium für Arbeit, Integration und Soziales des Landes Nordrhein-Westfalen sowie dem Europäischen Sozialfonds gefördert. Die Studie wurde von Prof. Dr. Spöttl geleitet und in Kooperation mit dem Berufsbildungsinstitut Arbeit und Technik (biat) der Universität Flensburg realisiert.

VI. Erfahrene Facharbeiter weisen höhere Kompetenzwerte auf als Auszubildende am Ende der Ausbildung.

Im Folgenden wird in aller Kürze das zugrunde gelegte Kompetenzmodell sowie das Erhebungsdesign vorgestellt, um dann die Ergebnisse zur Validitätsprüfung zu berichten.

Zum Kompetenzbegriff und Kompetenzmodell

Im Rahmen der beruflichen Bildung werden Kompetenzen zumeist als berufliche Handlungskompetenzen sehr umfassend beschrieben (siehe im Überblick z. B. Breuer 2006 oder Spöttl 2011). Immer wenn es um die large-scale-Erfassung von Kompetenzen mittels objektiver Tests geht, werden Kompetenzen dagegen in Anlehnung an Klieme/Leutner (2006) enger gefasst „als kontextspezifische kognitive Leistungsdispositionen, die sich funktional auf Situationen und Anforderungen in bestimmten Domänen beziehen“ (S. 4). *Enger* ist diese Definition in dreifacher Hinsicht:

- Sie schließt Bereitschaft und Gefühle aus und konzentriert sich allein auf Denkprozesse und Wissen („Kognition“).
- Sie konzentriert sich auf Fähigkeiten, die zweckgebunden sind („funktional“). Kompetenzen werden demnach eingesetzt, um (berufliche) Anforderungen zu bewältigen, die durch Dritte formuliert werden (etwa durch Kunden).
- Sie konzentriert sich auf inhaltlich definierte Bereiche („Domänen“) und ist damit verschieden von allgemeinen kognitiven Fähigkeiten wie Intelligenz.

Kompetenz im Rahmen der hier beschriebenen Untersuchung beruht auf diesem eng gefassten Begriff. Für die Domäne *Kfz-Service und Reparatur* wird Kompetenz definiert als die Fähigkeit, berufliche Aufgaben in Kfz-Servicewerkstätten zu bewältigen. Ein Modell kann nun auf zweierlei Weise konkretisiert werden. Entweder es benennt die objektiven Anforderungen, die zu bewältigen ein kompetenter Facharbeiter in der Lage sein muss (äußere Welt) oder aber die psychischen Voraussetzungen, die nötig sind, um den von Dritten formulierten Anforderungen zu genügen (innere Welt).

Tabelle 1: Arbeitsprozessbezogenes Kompetenzmodell für Facharbeit im Kfz-Service (Becker 2009, S. 243)

Schwierigkeitsniveau / Subdomäne	1	2	3	4
Service	Standardservice: Pflege und Wartung	Inspektion	Inspektion mit Zusatzarbeiten	Inspektion/ Sicherheitsprüfungen/ Abnahmen
Reparatur	Austauschreparatur	Verschleißreparatur	Schadensbehebung	Aggregateüberholung
Diagnose	Routinediagnose	Integrierte Diagnose	Regelbasierte Diagnose	Erfahrungsbasierte Diagnose
Installation	Zusatzinstallation/ Anbauteile	Zusatzinstallation/ Einbauteile	Erweiterungsinstallation	Systemerweiterung und –integration

Die Formulierung von externen Anforderungen hat dabei den Vorteil, dass diese in stärkerem Maße objektiv erhoben werden können und dass eine solche Erhebung für den Kfz-Service-Sektor bereits vorliegt: Im Rahmen der Neuordnung der Kfz-Berufe im Jahr 2003 wurden die Arbeitsprozesse in Werkstätten systematisch erhoben (vgl. Becker/Spöttl/Hitz/Rauner 2002). Diese lassen sich vier Subdomänen zuordnen, die zusammen das Arbeitsfeld von Fachkräften in Kfz-Werkstätten beschreiben: Service, Reparatur, Diagnose und Installation (vgl. die Zeilen in Tab. 1).

Das Anspruchsniveau der Arbeitsprozesse wurde auf der Grundlage von Expertenurteilen und Arbeitsprozessanalysen beurteilt (vgl. Becker u. a. 2002) und unter Rückgriff auf das Entwicklungsmodell vom Neuling zum Experten von Dreyfus/Dreyfus (1987) den verschiedenen Ausbildungsjahren zugeordnet (vgl. Spalten in Tab. 1). Die Testaufgaben (Items) wurden jeweils mit Bezug auf diese Arbeitsprozesse entwickelt, um so die Itemschwierigkeiten ex ante abschätzen zu können.

Forschungsdesign

Das Kompetenzerhebungsinstrument wurde in vier Phasen entwickelt, die jeweils durch eine Erprobung an unterschiedlich großen Stichproben abgeschlossen wurden. Nach der Fertigstellung zweier Rohtests für die zwei Subdomänen *Diagnose* und *Service* wurde mit 33 Probanden ein Pretest mit

anschließender Gruppendiskussion durchgeführt. Fünf dieser Probanden wurden zudem ausführlich interviewt, um zu überprüfen inwieweit die Aufgaben richtig verstanden werden und für die Zielgruppe angemessen sind. Die dabei identifizierten Unstimmigkeiten in den Items wurden behoben.

In der Finalerhebung wurde insgesamt 492 Probanden mindestens ein Test vorgelegt (siehe Tab. 3). Zwei Drittel dieser Stichprobe (330 Probanden) haben beide Instrumente ausgefüllt, 11 % haben nur den Servicetest und 22 % nur den Diagnosetest bearbeitet. Die befragten Auszubildenden in der Finalerhebung befanden sich allesamt kurz vor Abschluss ihrer Ausbildung zum Kfz-Mechatroniker und wurden über eine Gelegenheitsstichprobe an acht nordrhein-westfälischen Standorten rekrutiert.

Tabelle 2: Stichprobenzusammenstellung der Finalerhebung

	Häufigkeit	**Prozent**
SERVICE+DIAGNOSE	330	67%
nur SERVICE	53	11%
nur DIAGNOSE	109	22%
Gesamt	492	100%

Unter den 109 Probanden, die nur zur Diagnose befragt wurden, befanden sich auch 51 Kfz-Facharbeiter mit mindestens zweijähriger Berufserfahrung im Kfz-Service. Sie befanden sich allesamt am Beginn einer Fortbildung zum Kfz-Servicetechniker, die den technischen Teil der Ausbildung zum Kfz-Technikermeister umfasst.

Die Bearbeitung der zwei Tests dauerte jeweils 60 Minuten, die eines zusätzlichen Hintergrundfragebogens etwa 10 Minuten. Inklusive Testinstruktion bedeutete dies für die Probanden einen Aufwand von 3 Schulstunden (135 Minuten), die in aller Regel an unterschiedlichen Tagen aufgebracht wurden.

Ergebnisse

Im Folgenden werden anhand der aufgestellten Hypothesen zunächst die Ergebnisse zur internen und dann zur externen Validierung vorgestellt:

Zu Hypothese I) Die Leistung Serviceaufgaben zu lösen ist auf ein eindimensionales Konstrukt der „Servicekompetenz“ zurück zu führen, wenn der erhobene Datensatz mit den Annahmen des Raschmodells im Einklang

steht. Dazu wurden die Antworten der Probanden zunächst mit der Software ConQuest (Wu u. a. 2007) raschskaliert und anschließend geprüft, wie gut das Raschmodell zu den empirischen Daten passt. Nach einer ersten Skalierung der Daten wurde deutlich, dass 6 der 31 Service-Testitems nicht ausreichend zwischen kompetenten und weniger kompetenten Probanden unterscheiden konnten (mangelnde Trennschärfe der Items). Die verbleibenden 25 Items wiesen jedoch allesamt eine ausreichende Modellpassung auf. Zur Beurteilung des Misfits wurde ein Konfidenzintervall für den Weighted-Means-Square (MNSQ) auf der Grundlage des Standardfehlers berechnet. Dieser berücksichtigt die Größe der Stichprobe und liegt für 383 Probanden zwischen 0,93 und 1,07. Es zeigte sich, dass vier Items einen leichten Overfit zwischen 0,89 und 0,92 aufweisen. Diese Items zu entfernen hätte jedoch die inhaltliche Validität reduziert und zudem besonders trennscharfe Items entfernt. Da die Intervallgrenzen im Vergleich zu anderen Studien (bspw. Geißel 2008; Nickolaus/Gschwendtner/Geißel 2008) verhältnismäßig eng gezogen sind, ist die Beibehaltung der Items zu rechtfertigen.

Zusätzlich wurde geprüft, ob andere Testmodelle die erhobenen Daten evtl. besser erklären als das Raschmodell. Beispielsweise wäre denkbar, dass nicht für alle 383 Probanden ein einheitliches Raschmodell gilt, sondern in Untergruppen (so genannten Klassen) die Testaufgaben unterschiedlich schwierig ausfallen. Um dies auszuschließen, wurden mit der Software WINMIRA alternativ ein Zwei-, Drei- und Vierklassenmodell berechnet und mit dem Einklassenmodell verglichen. Es zeigt sich, dass das Einklassenmodell relativ am besten zu den Daten passt, wenn man Wert auf ein sparsames[2] Testmodell legt[3]. Zugleich ergeben sich für die Dimension Service Hinweise darauf, dass durch die Anwendung eines Zweiklassenmodells weitergehende Informationen über die Testteilnehmer gewonnen werden können, die aber aus Platzgründen an anderer Stelle berichtet werden müssen. Der auf 25 Items reduzierte Test ist für wissenschaftliche Zwecke ausreichend messgenau (Cronbachs $\alpha = 0{,}71$).

Zu Hypothese II) Die Eindimensionalität der Diagnosekompetenz wurde in gleicher Weise geprüft, wie die der Servicekompetenz. Nach einem ersten Skalierungsdurchgang wiesen 4 der insgesamt 31 Items zu geringe Trennschärfen auf und wurden aus dem Test entfernt. Die übrigen 27 Diagnose-

2 Sparsamkeit ist ein Zielkriterium bei der Anwendung von Testmodellen auf Leistungsdaten. Sparsamkeit bedeutet, dass die Varianz in den Daten durch möglichst wenige Modell-Parameter aufgeklärt werden sollte, weil dadurch die Interpretierbarkeit des Tests steigt.

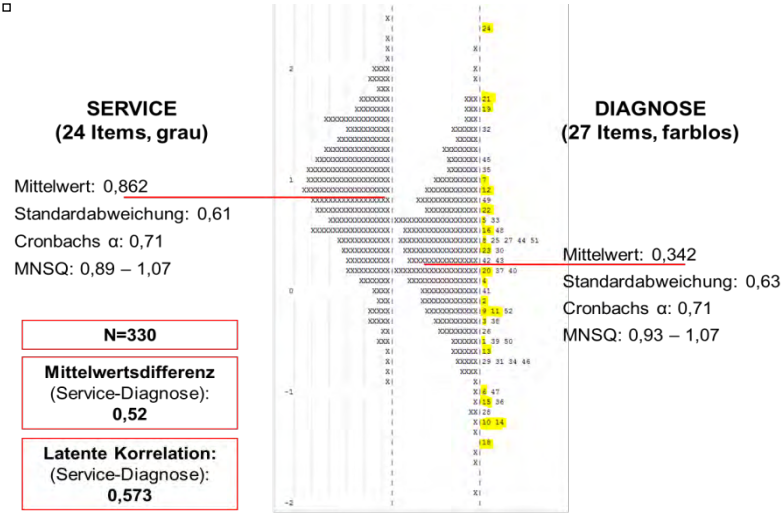
3 Für den Modellvergleich wurden so genannte Informationskriterien herangezogen, die angeben, welches Modell die vorliegenden Daten von allen angewendeten Modellen am besten beschreibt. Sowohl für das Bayes Information Criterion (BIC) als auch für das Consistent Akaike Information Criterion (CAIC) weisen die mehrklassigen Raschmodelle höhere Werte als das einklassige auf, was für die bessere Passung des einfachen Raschmodells spricht.

Items wiesen in einem erneuten Skalierungsdurchgang MNSQ-Werte zwischen 0,9 und 1,1 auf. Die interne Konsistenz beläuft sich auf 0,71 (Cronbachs α). Ebenso wie bei der Dimension Service konnten Zwei-, Drei- und Vierklassenmodelle die Daten nicht wesentlich besser erklären als das Einklassenmodell, so dass die Geltung des Raschmodells als gegeben angesehen werden kann.

Zu Hypothese III) Mithilfe einer multidimensionalen Skalierung, die eine Verallgemeinerung des eindimensionalen Raschmodells darstellt (vgl. Adams/Wu 2007), wurden die Diagnose- und Serviceitems jeweils einer eigenen Dimension zugewiesen und geschätzt, wie hoch die zwei Dimensionen miteinander korrelieren. Dabei handelt es sich um eine so genannte latente Korrelation, die vom Messfehler der Instrumente unabhängig ist. Wie angenommen, lassen sich beide Dimensionen empirisch trennen und korrelieren mit 0,573. Dies ist ein verhältnismäßig niedriger Wert, der bedeutet, dass Service- und Diagnosekompetenz zwar nicht unabhängige aber eigenständige Konstrukte darstellen. Die Eigenständigkeit der Dimensionen äußert sich auch in einem großen Mittelwertunterschied von 0,52 Logits. Um diesen Betrag fällt die Diagnose schwieriger aus als der Service. Dieser Unterschied ist signifikant und praktisch bedeutsam, da er nahezu eine Standardabweichung umfasst ($SD_{Service}$=0,61 bzw. $SD_{Diagnose}$=0,63 Logits)[4].

4 Der Mittelwertunterschied zwischen beiden Dimensionen zeigt, dass der Servicetest den Probanden leichter gefallen ist, als der Diagnosetest. *Beide Dimensionen sind also direkt miteinander vergleichbar!* Ob diese Aussage auf die Grundgesamtheit der Personen (Auszubildende können besser Service als Diagnose!) oder das Item-Universum (Service ist leichter als Diagnose!) verallgemeinert werden kann, hängt jeweils von der Stichprobenziehung ab. Die Testteilnehmer wurden über eine Gelegenheitsstichprobe akquiriert, was zu Einschränkungen der Verallgemeinerbarkeit führt. Für die Testinhalte wurde jedoch größte Sorgfalt auf die Auswahl und Gestaltung der Iteminhalte gelegt. Da dem Autor für die Berufsbildung keine Modellierungen an Zufallsstichproben bekannt sind, unterliegt die Aussagekraft dieser Modellierung damit den gleichen Einschränkungen wie alle bisher in der Berufsbildung vorgelegten empirischen Modellierungen.

Abbildung 2: Wright-Map zur multidimensionalen Skalierung von Service- und Diagnosetest



Zu Hypothese IV) Zur Prüfung der vierten Hypothese wurde ermittelt, wie hoch die Vorhersagekraft der theoretisch angenommenen Aufgabenmerkmale für die empirisch ermittelten Aufgabenschwierigkeiten ist. Lässt sich auf der Grundlage einer Theorie im Vornherein bestimmen, welche Aufgaben den Probanden schwerer bzw. leichter fallen, so ist dies ein entscheidender Schritt zum Verständnis der erfassten Kompetenz (vgl. auch Hartig 2007). Aufgabenmerkmale sind demnach das Verbindungsglied zwischen Theorie und Empirie und wurden hinsichtlich der folgenden Kriterien klassifiziert:

- Zuordnung zum Kernarbeitsprozess (vgl. Abb. 2)
- Zuordnung zum Expertisemodell nach Dreyfus/Dreyfus (1987)
- Häufigkeit des Vorkommens von den Items entsprechenden Aufgaben in Kfz-Werkstätten

Darüber hinaus gibt es Aufgabenmerkmale, die möglicherweise deren Schwierigkeit beeinflussen, ohne dass dies einen sinnvollen theoretischen Beitrag darstellt. Dies betrifft vor allem formale Aspekte der Items wie die

- Anzahl der Zeichen je Item,
- Aufgabendarstellung mit Bild/ ohne Bild.

Insgesamt haben fünf Beurteiler alle 27 Diagnose-Items bezüglich dieser Merkmale eingeschätzt. Bei fehlender Übereinstimmung wurde ein Konsens im Dialog hergestellt, bzw. das arithmetische Mittel der Beurteilungen her-

angezogen. In Regressionsanalysen zeigte sich, dass weder die theorierelevanten noch die theorieirrelevanten Aufgabenmerkmale einen signifikanten Beitrag zur Aufklärung der Itemschwierigkeiten leisten. Dieser Befund bedarf einer eingehenden Diskussion, die an dieser Stelle nicht geleistet werden kann.

Zur Hypothese V und VI) Für das vorliegende Testinstrument wurde als Anforderung formuliert, dass es ein Konstrukt erfasst, welches sowohl für frühe als auch für späte Stadien der Kompetenzentwicklung von Bedeutung ist. Ist das gegeben, kann gefolgert werden, dass der Test mit Blick auf das Endziel einer beruflichen Kompetenz valide ist und nicht nur schulisches Wissen ohne praktische Relevanz für den Arbeitsprozess erfasst. Eine Prüfung ist auf der Grundlage des Raschmodells möglich, da dort die grundlegende Annahme getroffen wird, dass für Probanden *mit gleicher Fähigkeit* alle Testitems die gleiche Lösungswahrscheinlichkeit aufweisen, also jeweils gleich schwierig sind. Ist dies für einzelne Items nicht gegeben, spricht man von Differential Item Functioning (DIF, Osterlind/Everson 2009). Lässt sich DIF für die Gruppe der Auszubildenden im Vergleich zu erfahrenen Fachkräften nachweisen, wäre dies ein Hinweis darauf, dass der Test in beiden Gruppen ein unterschiedliches Konstrukt erfasst, welches für die Arbeit von Fachkräften wenig Relevanz besitzt. Die Prüfung auf DIF zwischen Facharbeitern und Auszubildenden wurde nur anhand der Diagnose-Dimension vorgenommen, weil für den Servicetest bisher keine Facharbeiter gewonnen werden konnten.

Drei von 27 Items funktionieren bei Facharbeiten nicht in gleicher Weise wie bei Auszubildenden. Diese abweichenden Schwierigkeitsparameter sind bei einem 5%-Signifikanzniveau von Null verschieden. Alle drei Items sind dabei für die Auszubildenden schwieriger zu lösen als für die Facharbeiter, benachteiligen dementsprechend die Auszubildenden. Zur Prüfung der Hypothese, dass Fachkräfte im Test besser abschneiden als Auszubildende, sind die DIF aufweisenden Items auszuschließen. Durch den Ausschluss produziert die Skalierung faire Kompetenzwerte für beide Statusgruppen[5] bei einer vertretbaren Reduktion der Reliabilität von 0,71 auf 0,69 (Cronbachs α). Für den in dieser Weise reduzierten Test kann die Hypothese V als bestätigt angesehen werden. Zur Diskussion der ausgeschlossenen Items muss aus Platzgründen auf folgende Publikationen verwiesen werden.

Zur Analyse der Unterschiede in den mittleren Kompetenzausprägungen von Auszubildenden und Facharbeitern (Hypothese VI) wurde eine so genannte latente Regression berechnet, die den Einfluss der Statusvariablen (Azubi vs. Facharbeiter, unabhängige Variable) auf die latente Variable (Diagnosekompetenz, abhängige Variable) bestimmt. Trotz Ausschluss der drei

5 Ein Chi^2 Test auf Parametervergleichbarkeit für beide Statusgruppen ergibt einen Wert von 20.27 bei 23 Freiheitsgraden und wird damit nicht signifikant ($p = 0.625$).

DIF aufweisenden Items liegen die Kompetenzunterschiede zwischen Auszubildenden und erfahrenen Facharbeitern bei 0,525 Logits. Bei einem Standardfehler von 0,124 ist dieser Wert auf dem 99%-Niveau signifikant (z=4,23 p=0,00). Somit kann auch die Hypothese VI bestätigt werden, was insgesamt für die externe Gültigkeit des entwickelten Tests im Sinne der Argumentation spricht.

Schlussfolgerung und Ausblick

Zur Validierung des Multiple-Choice-Tests zur Erfassung von Kompetenzen in der Domäne Kfz-Service & Reparatur wurden sowohl interne als auch externe Strategien eingesetzt. Die Ergebnisse belegen im Großen und Ganzen die Validität des Instruments für Auszubildende am Ende des dritten Lehrjahres und Facharbeiter mit mindestens 2-jähriger Berufserfahrung.

Neben den beschriebenen statistischen Möglichkeiten der Validitätsprüfung ist es von besonderer Bedeutung, die inhaltliche Gültigkeit des Instruments zu belegen. Zu diesem Zwecke sind zahlreiche Anstrengungen unternommen worden, die im Rahmen dieses Beitrags jedoch keinen Platz fanden. Die Bedeutung dieses Aspekts ist aber umso größer, als einige der theoretisch für wichtig erachteten Items dem Testoptimierungsprozess „zum Opfer fielen".

Darüber hinaus ist zu klären, was die Items mit DIF zwischen Facharbeitern und Auszubildenden über die Natur von Kompetenzen im Kfz-Service verraten, insbesondere welche Art von Wissen eher für Facharbeiter und weniger für Auszubildende relevant ist. Im Zuge dessen finden sich evtl. auch Hinweise, die die mangelnde Erklärungsmächtigkeit von theoretischen Parametern für die empirischen Itemschwierigkeiten betreffen.

Mittelfristig scheint lohnenswert, zu prüfen, inwiefern sich der hier verwendete Test auch als stabil bei der Erfassung von Kompetenzentwicklungen erweist. Durch den „pseudo"-Längsschnitt im Experten-Novizen-Vergleich bestehen gute Chancen, dass sich eine große Zahl der verwendeten Items als gute Ankeritems herausstellen.

Literatur

Adams, R. / Wu, M. L. (2007). The Mixed-Coefficients Multinomial Logit Model: A Generalized Form of the Rasch Model. In: M. von Davier; C. H. Carstensen (Hrsg.): Multivariate and Mixture Distribution Rasch Models. Extensions and Applications. New York, NY: Springer Science + Business Media LLC, S. 57–75.

Becker, M. (2009). Kompetenzmodell zur Erfassung beruflicher Kompetenz im Berufsfeld Fahrzeugtechnik. In: C. Fenzl; G. Spöttl; F. Howe; M. Becker (Hrsg.): Berufsarbeit von morgen in gewerblich-technischen Domänen - Forschungsansätze und Ausbildungskonzepte für die berufliche Bildung. Bielefeld: Bertelsmann, S. 239–245.

Becker, M./ Spöttl, G./ Hitz, H./ Rauner, F. (2002). Wissenschaftliche Begleitung zur Neuordnung der fahrzeugtechnischen Berufe. Aufgabenanalyse für die Neuordnung der Berufe im Kfz-Sektor. Abschlussbericht. Bremen, Flensburg.

Borsboom, D. (2005). Measuring the Mind. Conceptual issues in modern psychometrics. Cambridge: Cambridge University Press.

Borsboom, D. / Mellenbergh, G. J. / Heerden, J. van (2004). The Concept of Validity. In: Psychological Review, Jg. 111, H. 4, S. 1061–1071.

Breuer, K. (2006). Kompetenzdiagnostik in der beruflichen Bildung - eine Zwischenbilanz. In: Zeitschrift für Berufs- und Wirtschaftspädagogik, 102. Band, Heft 2, S. 194–210.

Dreyfus, H. L./ Dreyfus, S. E. (1987). Künstliche Intelligenz. Von den Grenzen der Denkmaschine und dem Wert der Intuition. Reinbek bei Hamburg: Rowohlt.

Geißel, B. (2008). Ein Kompetenzmodell für die elektrotechnische Grundbildung: Kriteriumsorientierte Interpretation von Leistungsdaten. In: R. Nickolaus; H. Schanz (Hrsg.): Didaktik der gewerblich-technischen Berufsbildung. Konzeptionelle Entwürfe und empirische Befunde. Baltmannsweiler: Schneider-Verl. Hohengehren.

Gruber, H./ Mandl, H./ Renkl, A. (2000). Was lernen wir in Schule und Hochschule: Träges Wissen. In: H. Mandl; J. Gerstenmaier (Hrsg.): Die Kluft zwischen Wissen und Handeln. Empirische und theoretische Lösungsansätze. Göttingen: Hogrefe Verl. für Psychologie, S. 139–156.

Gschwendtner, T. (2008). Ein Kompetenzmodell für die Kraftfahrzeugtechnische Grundbildung. In: R. Nickolaus; H. Schanz (Hrsg.): Didaktik der gewerblich-technischen Berufsbildung. Konzeptionelle Entwürfe und empirische Befunde. Baltmannsweiler: Schneider-Verl. Hohengehren, S. 103–119.

Hartig, J. (2007). Skalierung und Definition von Kompetenzniveaus. In: B. Beck; E. Klieme (Hrsg.): Sprachliche Kompetenzen. Konzepte und Messung. Weinheim: Beltz, S. 79–95.

Klieme, E./ Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Überarbeitete Fassung des Antrags an die DFG auf Einrichtung eines Schwerpunktprogramms.

Musekamp, F. (2009). Entwicklung eines standardisierten Instruments zur Kompetenzerhebung im Kfz-Service. In: C. Fenzl; G. Spöttl; F. Howe; M. Becker (Hrsg.): Berufsarbeit von morgen in gewerblich-technischen Domänen - Forschungsansätze und Ausbildungskonzepte für die berufliche Bildung. Bielefeld: Bertelsmann, S. 246–251.

Musekamp, F./ Spöttl, G./ Becker, M. (2010). Schriftliche Arbeitsaufträge zur Erfassung von Differenzen in der Expertise von Facharbeitern und Auszubildenden. In: Zeitschrift für Berufs- und Wirtschaftspädagogik, 106. Band, Heft 3, S. 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. Stuttgart.

Osterlind, S. J. / Everson, H. T. (2009). Differential item functioning. Thousand Oaks, Calif: Sage.

Rost, J. (2004). Lehrbuch Testtheorie - Testkonstruktion. Psychologie Lehrbuch. Bern u. a.: Huber.

Spöttl, G. (2011). Kompetenzmodelle als Grundlage für eine valide Kompetenzdiagnostik – Anforderungen an Theoriebildung und Empirie. In: M. Fischer; M. Becker; G. Spöttl (Hrsg.): Kompetenzdiagnostik in der beruflichen Bildung. – Probleme und Perspektiven. Frankfurt am Main: Peter Lang, S. 13–39.

Spöttl, G. (2009). Kompetenzmodelle in der beruflichen Bildung - Grenzen und Chancen. In: C. Fenzl; G. Spöttl; F. Howe; M. Becker (Hrsg.): Berufsarbeit von morgen in gewerblich-technischen Domänen - Forschungsansätze und Ausbildungskonzepte für die berufliche Bildung. Bielefeld: Bertelsmann, S. 233–238.

Spöttl, G./ Becker, M./ Musekamp, F. (2011). Anforderungen an Kfz-Mechatroniker und Implikationen für die Kompetenzerfassung. In: R. Nickolaus; G. Pätzold (Hrsg.): Lehr-Lernprozesse in der gewerblich-technischen Berufsbildung. Stuttgart: Steiner, S. 37–53.

Wu, M. L./ Adams, R./ Wilson, M. R. (2007). ACER ConQuest version 2.0. Generalised item response modelling software. Camberwell Vic.: ACER Press.