

**Universität Stuttgart**  
Institut für Erziehungswissenschaft  
und Psychologie



**Georg-August-Universität  
Göttingen**

**Abele, Stephan / Achtenhagen, Frank / Gschwendtner, Tobias /  
Nickolaus, Reinhold / Winther, Esther**

## **Die Messung beruflicher Fachkompetenz im Rahmen eines Large Scale Assessments im Bereich beruflicher Bildung (VET-LSA) – Vorstudien zur Validität von Simulationsaufgaben**

Integrierte **Kurzfassung der Studien** von

**Achtenhagen, Frank / Winther, Esther: Konstruktvalidität von Simulationsaufgaben: Computergestützte Messung berufsfachlicher Kompetenz am Beispiel der Ausbildung von Industriekaufleuten**

und

**Nickolaus, Reinhold / Gschwendtner, Tobias / Abele, Stephan: Die Validität von Simulationsaufgaben am Beispiel der Diagnosekompetenz von Kfz-Mechatronikern**

### **Kontakt:**

Seminar für Wirtschaftspädagogik der  
Georg-August-Universität Göttingen  
Platz der Göttinger Sieben 5  
37073 Göttingen (E-Mail: [fachten@uni-goettingen.de](mailto:fachten@uni-goettingen.de); [ewinthe@uni-goettingen.de](mailto:ewinthe@uni-goettingen.de))

Universität Stuttgart  
Institut für Erziehungswissenschaft und Psychologie  
Abteilung Berufs-, Wirtschafts- und Technikpädagogik (BWT),  
Geschwister-Scholl-Str. 24 D, 3. OG, 70174 Stuttgart  
Tel.: 0711/685-83181, Fax: 0711/685-83130,  
(E-Mail: [nickolaus@bwt.uni-stuttgart.de](mailto:nickolaus@bwt.uni-stuttgart.de); [abele@bwt.uni-stuttgart.de](mailto:abele@bwt.uni-stuttgart.de);  
[gschwendtner@bwt.uni-stuttgart.de](mailto:gschwendtner@bwt.uni-stuttgart.de))

Fassung vom 14. Januar  
2010

## 1. Zielsetzung der Studien, grundlegende Prinzipien und Messverfahren

Vorgestellt werden hier Ergebnisse von zwei Studien zur Klärung der Frage, auf welche Weise berufliche Fachkompetenz in einer internationalen Vergleichsstudie so erfasst werden kann, dass sowohl das für berufliches Handeln notwendige Wissen als auch die **Leistung in realen Anforderungssituationen** verlässlich gemessen wird.

Dabei erfolgt eine Konzentration auf fachliche Anforderungssituationen in den Berufen Industriekauf- frau/-mann und Kfz-Mechatroniker. Zur Erfassung sozialer und personaler Kompetenzen sind im Rahmen der internationalen Vergleichsstudie spezielle Messverfahren vorgesehen, die auch zu diesen Kompetenzdimensionen eine verlässliche Messung gewährleisten. Gleichwohl sind auch in den auf die Erfassung fachlicher Kompetenz ausgerichteten Aufgaben zugleich soziale und personale Kompetenzen relevant, beim kaufmännischen Handeln ist dies offensichtlich aber auch bei der beruflichen Facharbeit von Kfz-Mechatronikern hat deren vordergründig fachliche Leistungsfähigkeit (z. B. bei der Fehlerdiagnose) soziale und personale Implikationen, wenn z. B. aufgrund von Fehldiagnosen funktionsfähige Teile ausgetauscht und den Kunden in Rechnung gestellt werden. **Insgesamt wird mit dem vorgesehenen Messverfahren eine verlässliche Abschätzung von zentralen Kernen beruflicher Handlungskompetenz gewährleistet.** Der Anspruch, alle möglichen situationsspezifischen Kompetenzfacetten abzubilden, kann zwar formuliert jedoch vermutlich nie eingelöst werden, da die Anforderungssituationen ein nahezu unbegrenztes Spektrum umfassen.

### 1.1 Ansprüche an die Messverfahren

Zentral ist der Anspruch Messverfahren einzusetzen, die eine verlässliche Abschätzung beruflicher Leistungsfähigkeit (Performanz) gewährleisten. Das bedeutet zugleich, dass diese Verfahren objektiv (nicht durch subjektive Einschätzungen verfälscht), reliabel (genaue Messung), valide (messen das, was erfasst werden soll) und skalierbar sind.

### 1.2 Prinzipiell mögliche Messverfahren

Prinzipiell kommen folgende Messverfahren in Frage:

(1) Konfrontation mit realen, standardisierten Aufgaben (Arbeitsproben), (2) Einschätzungen der Performanz im Arbeitsalltag, (3) Erfassung beruflicher Leistungsfähigkeit in realitätsnahen Simulationen beruflicher Anforderungssituationen, (4) Selbsteinschätzungen zur Ausprägung von Teilkompetenzen und paper-pencil-Tests mit offenen oder geschlossenen Fragestellungen.

#### *Eignung der Verfahren für eine (internationale) Vergleichsstudie*

zu (1): Konfrontation mit realen Aufgaben: Diese prinzipiell denkbare Variante ist mit massiven Problemen verbunden, die Aufgaben in betrieblichen Anforderungssituationen im internationalen Kontext

zu standardisieren und damit die Ergebnisse vergleichbar zu machen. Zugleich wäre dieses (unpraktikable) Verfahren sehr teuer.

Bewertung: unpraktikabel, erhebliche Finanzierungsprobleme, die Standardisierung im realen Arbeitsvollzug scheint unrealistisch; für den kaufmännisch-verwaltenden Bereich existieren neben dem Gebrauch betriebsindividueller Software auch die Probleme der Geheimhaltung von Daten (Preise, Kalkulationen, Kundenkontaktdaten, etc.); ohne Standardisierung sind keine verlässlichen Ergebnisse zu erwarten.

Zu (2): Einschätzungen der Performanz im Alltag: Dieses Verfahren wäre bei Einsatz örtlicher Ausbilder/Vorgesetzter zwar kostengünstig, die Einschätzungen würden sich jedoch notgedrungen auf unterschiedliche Anforderungssituationen beziehen, zudem ist - auch bei indikatorengestützten Verfahren – mit subjektiv gefärbten Einschätzungen zu rechnen, sofern die Einschätzungen von den Ausbildern vor Ort vorgenommen werden. Der Einsatz von unabhängigen Beobachtern vor Ort ist hoch aufwändig, da eine längere Beobachtung notwendig wäre.

Bewertung: für übergreifende Vergleiche ungeeignet.

Zu (3): Simulation realer Anforderungen: Dieses Verfahren gewährleistet vergleichbare Ergebnisse und kann realitätsnah gestaltet werden. Die Prüfung, ob dieses Verfahren verlässliche Abschätzungen der beruflichen Leistungsfähigkeit ermöglicht, wird in den Vorstudien geprüft.

Bewertung: geeignet (s.u.), am besten in Verbindung mit (5).

Zu (4): Selbsteinschätzungen liefern keine verlässlichen Daten zur realen Ausprägung der Kompetenz, sind in hohem Grade subjektiv verzerrt und beziehen sich auf differente Anforderungssituationen.

Bewertung: ungeeignet

Zu (5) Paper-pencil-Tests: Paper-pencil-Tests können sehr variantenreich gestaltet werden. Die Aufgaben können offen oder geschlossen und mehr oder weniger fach- oder auch handlungssystematisch zugeschnitten werden. Zweifellos kann durch diese Testform das für berufliches Handeln relevante Wissen gut abgeschätzt werden. Inwieweit diese Testform auch geeignet ist Performanz abzuschätzen wäre gegebenenfalls zu prüfen.

Bewertung: beim gegenwärtigen Kenntnisstand vor allem geeignet in Verbindung mit den Simulationen.

## **2. Das der Messung zugrunde liegende bereichsspezifische Kompetenzkonstrukt und dessen Bezüge zu Wissen und Performanz**

Ein Vergleich der unterschiedlichen Kompetenzdefinitionen in der beruflichen Bildung zeigt, dass das Definitionselement „auf spezifische (berufliche) Anforderungsbereiche bezogene Fähigkeiten, die eine eigenständige, gegebenenfalls auch kooperative Bewältigung variierender berufstypischer Anforderungssituationen ermöglichen—, durchgängig herangezogen wird. In den meisten Fällen werden auch Bereitschaften einbezogen diese Fähigkeiten einzusetzen, was allerdings messtechnisch insoweit prob-

lematisch ist, als Bereitschaften gegebenenfalls starken, auch im Tagesverlauf schwankenden Ausprägungen (hohe Bereitschaft am Morgen, nachlassende Bereitschaft gegen 16.30 Uhr) unterworfen sind. Zur **Kompetenzstruktur** wird meist angenommen, dass die Kompetenzdimensionen Fachkompetenz, Sozialkompetenz und personale Kompetenz zu unterscheiden sind und unterhalb dieser Ebene weitere Ausdifferenzierungen vorgenommen werden können (z. B. **Fachkompetenz**: Wissen und Verständnis von fachlichen Zusammenhängen, Fähigkeit fachliche Probleme eigenständig zu lösen; **Sozialkompetenz**: Dialogfähigkeit, Koordinationsfähigkeit, Kooperationsfähigkeit; **Personalkompetenz**: Wissen um die eigenen Stärken und Schwächen, Fähigkeit die eigene Entwicklung selbst zu steuern). Zum Teil wird auch angenommen, dass es eine übergreifende Methodenkompetenz gibt, was bisher allerdings empirisch nicht bestätigt werden konnte. Vielmehr scheinen fachliche und darauf bezogene methodische Fähigkeiten eng verwoben.

Versuche alle möglichen Kompetenzaspekte mit einer oder mehreren Aufgaben simultan zu erfassen führten bisher zu keinen verlässlichen Abschätzungen, ein befriedigendes Ergebnis ist auf diesem Wege auch nicht zu erwarten. Vor diesem Hintergrund sind verlässliche, auf zentrale Kompetenzkerne ausgerichtete Verfahren, die an dem obigen Strukturmodell ausgerichtet sind, für die internationalen Vergleichsuntersuchungen adäquat. Damit können über Fachkompetenzen hinaus auch wichtige, gesellschaftlich als bedeutsam erachtete Kompetenzausprägungen wie z. B. Bereitschaften zur Mitgestaltung erfasst werden.

#### *Verhältnis von Wissen, Kompetenz und Performanz*

Es besteht Einigkeit, dass Wissen eine notwendige Voraussetzung für ein zieladäquates und reflektiertes Handeln, die Fähigkeit Wissen zu reproduzieren jedoch keine hinreichende Bedingung für erfolgreiches Handeln darstellt. Empirisch zeigt sich im gewerblich-technischen Bereich, dass das Fachwissen der stärkste Einflussfaktor auf die fachspezifische Problemlösefähigkeit zu sein scheint (vgl. z.B. NICKOLAUS/HEINZMANN/KNÖLL 2005; NICKOLAUS/KNÖLL/GSCHWENDTNER 2006). Generell ist davon auszugehen, dass der Zusammenhang sowohl von der Art des Wissens als auch den Anforderungsstrukturen der Problemstellungen abhängig ist.

Die Stärke des Zusammenhangs zwischen der Wissensausprägung und der beruflichen Performanz ist davon abhängig, ob das Wissen mehr oder weniger handlungsbezogen erfasst wird bzw. davon, inwieweit die Testaufgaben zur Wissensausprägung die kognitiven Anforderungen der Problemsituation repräsentieren.

### **3. Neue Messverfahren und ihre Vorteile gegenüber üblichen Prüfungsvarianten**

Im Rahmen der internationalen Vergleichsstudien wie z. B. PISA wurden neue Messverfahren eingesetzt, die die Möglichkeit geben zu prüfen, ob sich Annahmen zur Kompetenzstruktur bestätigen. Des Weiteren werden so genannte Niveaumodellierungen möglich, die Informationen bereit stellen, welches Fähigkeitsniveau die Auszubildenden erreicht haben, d.h. welchen Anforderungen die Auszubil-

denden gewachsen sind und welchen nicht.

Um Tests entwickeln zu können, die den Anforderungen dieser neuen Verfahren genügen, ist es notwendig zu wissen, welche Anforderungen in einem Beruf relevant sind und welche Aufgabenmerkmale Aufgaben mehr oder weniger schwierig machen. Der inzwischen erreichte Forschungsstand gewährleistet die Einlösung beider Bedingungen in den beiden gewerblich-technischen Berufen und dem kaufmännischen Beruf.

#### **4. Ergebnisse der Studien**

Durchgeführt wurden insgesamt drei Vorstudien: zwei Validierungsstudien zur Frage, ob mit Simulationen realer Anforderungen die Leistungsfähigkeit in realen Situationen verlässlich abgeschätzt werden kann, und eine Studie, in der geprüft wurde, ob in den für eine Vergleichsstudie zunächst ausgewählten Berufen ein internationaler Vergleich möglich ist. Berichtet wird hier von den Ergebnissen der Validierungsstudien.

##### **4.1 Anlage und Ergebnisse der Validierungsstudie im Kfz-Bereich**

Die Anlage der Validierungsstudie wird hier knapp skizziert; zu deren Einbindung in den Forschungsstand sei auf die Langfassung des Berichts verwiesen. Auch bei der Ergebnisdarstellung beschränken wir uns hier auf jene Kernaussagen, die für die Klärung der Validitätsfrage hinreichend erscheinen. Dem an Details interessierten Leser sei auch dazu die Lektüre der Langfassung empfohlen.

###### **4.1.1 Anlage der Studie**

Ziel der Studie war die Klärung der Frage, ob die Abschätzung beruflicher Performanz auf der Basis von simulierten Anforderungssituationen verlässlich möglich ist. Geklärt wurde diese Frage im Kfz-Bereich über die Untersuchung der Übereinstimmung in den Fehleranalyseleistungen innerhalb authentischer Fehlerfälle in realen kraftfahrzeugtechnischen Systemen einerseits und computerbasierten Repräsentationen dieser Systeme andererseits. Dazu wurden acht komplexe Aufgabenstellungen entwickelt, die je Aufgabe in Zwillingenform realisiert wurden, einmal in Form einer Computersimulation und zum anderen im realen Fahrzeug.

###### **4.1.2 Erhebungsdesign**

Für die Überprüfung der Übereinstimmung kommen prinzipiell verschiedene Erhebungsdesigns in Frage, die mit je spezifischen Implikationen verknüpft sind: Z.B. (1) Allen Probanden werden alle Aufgaben beider Settings vorgelegt, (2) es werden randomisierte und hinreichend große Gruppen gegeneinander getestet, für die unterstellt werden kann, dass die Fähigkeit zur Lösung der Aufgaben gleich verteilt ist, (3) ein multi-matrix-design mit einer hinreichenden Stichprobengröße und (4) ein cross-over-design, für das wir uns unter den gegebenen Bedingungen entschieden haben (ausführlicher siehe die Langfassung).

Die Auszubildenden wurden randomisiert auf zwei Versuchsgruppen aufgeteilt, die in einem cross-over-design alle acht Aufgaben verteilt auf beide Settings zu lösen hatten (siehe Abb. 1). Durch den Einsatz von Fehlerfall-hochaffinen und anwendungsorientierten Wissensaufgaben planten wir, die Problemfälle von Gruppe 1 und Gruppe 2 gemeinsam zu verankern und somit vergleichbar zu machen (siehe Abb. 1).

	Gruppe 1 (N = 134)	Gruppe 2 (N = 123)
<b>Fehlerfall 1</b>	Am realen Kfz ( $R_1$ )	Bearbeitung des Fehlers in Simulation ( $S_1$ )
<b>Fehlerfall 3</b>	$R_3$	$S_3$
<b>Fehlerfall 5</b>	$R_5$	$S_5$
<b>Fehlerfall 7</b>	$R_7$	$S_7$
<b>Fehlerfall 2</b>	$S_2$	$R_2$
<b>Fehlerfall 4</b>	$S_4$	$R_4$
<b>Fehlerfall 6</b>	$S_6$	$R_6$
<b>Fehlerfall 8</b>	$S_8$	$R_8$
<b>Übergreifend</b>	<b>Anwendungsorientierte Wissensaufgaben</b>	

**Abb. 1: Erhebungsdesign für den Vergleich realer und simulierter Fehlerfalllösungen**

#### 4.1.3 Stichprobe und Durchführungsobjektivität

Die Stichprobengesamtgröße beträgt  $N = 294$ . Die Gesamtstichprobe untergliedert sich in  $N = 202$  Schüler aus dem 3. und  $N = 92$  Schüler aus dem 4. Ausbildungsjahr. Die Auszubildenden des 4. Lehrjahrs sind allesamt Lehrlinge aus Handwerksbetrieben. Die Auszubildenden des 3. Lehrjahres bestehen aus  $N = 63$  Auszubildenden aus Berufskollegklassen,  $N = 78$  Auszubildenden aus Handwerksklassen und  $N = 61$  Auszubildenden aus Industrieklassen. Diese Konstellation eröffnet zusätzlich zu der Fragestellung der Studie weitere Analysemöglichkeiten.

Neben der starken Affinität der Fehlerfälle zum beruflichen Alltag sorgte ein Anreizsystem für eine hohe Motivation der Probanden und damit für eine vergleichsweise hohe Durchführungsobjektivität. Die Innung des Kraftfahrzeuggewerbes Region Stuttgart und die Handwerkskammer Region Stuttgart verfassten separat Anschreiben an alle Auszubildenden des Ausbildungsberufs Kraftfahrzeugmechatroniker der Region und die dazugehörigen Betriebe und betonten den hohen Stellenwert der Untersuchung im Sinne der Prüfungsvorbereitung. Zusätzlich wurden attraktive Preise in Höhe von 1000 EURO für die „Besten—in Aussicht gestellt. In Verbindung mit dem personellen Aufwand ergab sich damit eine sehr hohe Objektivität der erhobenen Daten.

Die Probanden wurden auf der Ebene von Klassen und gemäß dem Rücklauf aus den Anschreiben den beiden Settings randomisiert zugeordnet. Der Versuchsgruppe 1 gehörten  $N = 134$  Auszubildende an, der Versuchsgruppe 2  $N = 123$ .

#### 4.1.4 Erhebungsinstrumente

Im Anschluss an vorliegende Studien zu Tätigkeitsanforderungen und Tätigkeitsbereichen von Kfz-Mechatronikern, die vor allem in den Bereichen Service sowie Diagnose- und Reparaturarbeiten angesiedelt sind (BECKER 2005; HÄGELE 2002), wurde entschieden, in dieser Studie den Fokus auf Diagnosearbeiten und damit auf einen zentralen und zugleich relativ anspruchsvollen Tätigkeitsbereich zu legen. Die Entwicklung und Selektion der Diagnosefälle und Wissensitems erfolgte in enger Anlehnung an die in der Praxis auftretenden Fehlerfälle und in enger Kooperation mit Experten (Zentralverband des deutschen Kraftfahrzeuggewerbes, Innung des Kraftfahrzeuggewerbes Region Stuttgart, Bildungszentrum der Handwerkskammer Region Stuttgart, Hotlinemitarbeiter eines Anbieters von Diagnosesoftware und Hilfen für alltägliche Werkstattprobleme, berufliche Schulen, Fachleiter, Ausbildungsmeister und Auszubildende). Alle Inhalte der Erhebungsinstrumente sind sowohl im 3. als auch 4. Ausbildungsjahr curricular abgesichert. Die Inhalte beziehen sich auf zwei große Fahrzeugsysteme, die Beleuchtungsanlage und das Motormanagement. Folgende Erhebungsinstrumente kamen zum Einsatz:

##### **(1) Authentische Fehlerfälle im realen Fahrzeug und in der Computersimulation**

Ausgewählt bzw. entwickelt wurden acht komplexe Diagnoseaufgaben, die im Bereich des Motormanagements (sechs Aufgaben) und der Beleuchtungsanlage (zwei Aufgaben) angesiedelt sind. Angestrebt wurde neben der hohen Authentizität eine möglichst große Varianz der Schwierigkeitsgrade, wobei zur Abschätzung einerseits auf eigene Vorarbeiten zu schwierigkeitsbestimmenden Merkmalen (GSCHWENDTNER/GEIBEL/NICKOLAUS 2007; GSCHWENDTNER 2008) und andererseits auf die Erfahrungen der Experten zurückgegriffen werden konnte. Die Pilotierung der Aufgaben, im Rahmen derer mit den Auszubildenden auch Interviews zur Bearbeitung durchgeführt und deren Herangehensweise erfasst wurden, führte zu sukzessiven Optimierungen der Aufgaben. Alle Testelemente wurden von Seiten der Experten als inhaltlich valide und im Anspruchsgrad als angemessen und variabel eingeschätzt. Die Testzeit wurde je Fehlerfall auf 30 Minuten normiert. Damit ergab sich eine Gesamttestzeit von 4 Stunden je Proband. Aus den Pilotierungen und Durchführungen in der Hauptstudie lässt sich sagen, dass für die meisten Probanden die Fehlerfallbearbeitung günstigerweise als Power- und nicht als Speedtest anzusehen ist. Die Auswertungen der Fehleranalysefähigkeit der Probanden (sowohl für die realen als auch die simulierten Fehlerfälle) erfolgte an Hand eines jedem Fehlerfall beigelegten Dokumentationsbogens. Auf diesem waren mittels drei Fragen im offenen Antwortformat der realisierte Arbeitsplan zur Fehlersuche (Fehlersuchstrategie), die genaue Benennung des defekten Bauteils und eine Begründung abzugeben, warum es nicht auch ein anderer Fehler sein könnte (schlussfolgerndes Denken, Messwertinterpretationen, Kontrollmessungen). Der Bearbeitung der Fehlerfälle, die in der Computersimulation repräsentiert wurden, gingen eine 20-minütige Einführung und

eine 10-minütige Übungsphase voraus, in der durch die Bearbeitung eines Übungsblattes mit exemplarischen Funktionalitäten abgesichert werden konnte, dass ein jeder das Handwerkszeug der Simulationsbedienung beherrschte, bevor mit dem ersten Fehlerfall begonnen wurde.

## **(2) Wissenstest**

Neben den Diagnoseaufgaben, anhand derer die fachspezifische Problemlösefähigkeit in diesem Tätigkeitssegment abgeschätzt werden kann, wurde ein handlungsorientierter Fachwissenstest entwickelt, der inhaltlich auf das im Kontext der Fehlerdiagnosen notwendige Wissen ausgerichtet ist. Alle Testelemente des Wissenstests wurden von Seiten der Experten als inhaltlich valide und im Anspruchsgrad als angemessen eingeschätzt. Der Wissenstest besteht aus 16 Items, wovon zwei als multiple-choice-Items und 14 im offenen Antwortformat formuliert sind. Der Wissenstest hat zwei Teile. Der erste Teil bezieht sich auf die Fahrzeugbeleuchtungsanlage, der zweite Teil auf das Motormanagement. Beiden Teilen liegt ein Stromlaufplan als Analysemedium zu Grunde. Ebenso beiden Teilen gemein ist die Fragestruktur der Items: Es werden funktionale Zusammenhänge, sowie systemische Kenntnisse im Sinne von Wissen über Veränderungen im Systemoutput durch Variation von Eingangsgrößen und Fehlersuchstrategien erfragt. Die Testzeit wurde auf 60 Minuten normiert. Die Erfahrungen mit den Pilotierungen und den Durchführungen der Hauptstudie zeigen, dass - wie für die Fehlerfälle - die Fehlerfallbearbeitung für die meisten Probanden günstigerweise ein power- und kein speedtest ist. Der Wissenstest ist analog dazu konzipiert, wie die Fehlersuche in diesen Systemen gedanklich vollzogen werden kann. Der Anspruch ist, den gesamten Test oder eine Itemauswahl als Ankeritems und damit Vergleichsmedium zu benutzen (siehe oben).

### **4.1.5 Kodierung der Daten**

Die Items wiesen bis auf wenige im Wissenstest allesamt ein offenes Antwortformat auf. Der Anreiz für die Schüler, möglichst präzise Antworten zu geben, war aufgrund des gegebenen Anreizsystems hoch. Damit sollten auch Teillösungen mit in das Modell aufgenommen werden, d.h. Lösungen, die im Ansatz (Lösungsweg) zwar (weitgehend) richtig, aber sprachlich so unpräzise waren, dass eine genaue Lokalisierung des Fehlers nicht hinreichend präzise getroffen werden konnte. Die Beurteilung der Adäquanz solcher partial-credit items sollte anhand psychometrischer Kriterien erfolgen (siehe unten). Die Tests wurden von uns in doppelter und getrennter Korrektur kodiert, wobei es zu einer hohen Übereinstimmung gekommen ist. In den wenigen Fällen diskrepanter Beurteilung wurde zusammen mit einem externen Experten eine Entscheidung getroffen.

### **4.1.6 Validitätsprüfung**

Zu prüfen war, ob zur Lösung der realen und simulierten Aufgaben gleiche Fähigkeitsbündel, die in einer Fähigkeitsdimension konvergieren, benötigt werden und wenn ja, ob die Aufgaben gleich schwierig (komplex) und damit vergleichbar sind. Zur Prüfung dieser Frage bieten sich mehrere Ver-

fahren an: (1) ein auf Basis klassischer Testtheorie angelegter Vergleich der simulierten (S1...S8) und realen (R1...R8) Aufgaben und (2) eine Skalierung der Aufgaben auf der Basis der Item-Response-Theorie.

#### a) **Prüfschritte zur Beantwortung der Forschungsfrage**

Die Forschungsfrage wird beantwortet, indem wir zunächst prüfen, ob wir unterstellen können, dass für die Lösung realer und simulierter Aufgaben die gleichen Fähigkeiten nötig sind. Trifft das zu, so ist es legitim, fehlende Schwierigkeitsdifferenzen zwischen einzelnen realen und simulierten Fehlerfällen auf einer gemeinsamen Skala beider Gruppen (die die Realitäts- und Simulationsaufgaben alternierend gelöst haben) so zu interpretieren, dass es keinen Unterschied für die reliable Verortung einer Person auf einem Fähigkeitskontinuum macht, ob die hinter dem Item stehende Anforderung in der Simulation oder der Realität erfolgt. Durchgeführt wurden dazu drei ineinander verschränkte Prüfungen: (1) Die Dimensionalitätsprüfung mittels latenter Korrelationen zwischen der Lösung realer und simulierter Aufgaben und der Prüfung der Itemfitwerte in einzelnen Gruppenskalierungen. (2) Die Prüfung der randomisiert zusammengesetzten Untersuchungsgruppen (Gruppe 1 und Gruppe 2) auf gleiche Verteilungen und (3) eine Differenzwertbeurteilung etwaiger Itemschwierigkeitsverzerrungen auf der Basis von ANOVA, DIF-Analysen und *scale linking*.

#### b) **Beantwortung der Forschungsfrage**

**Zu (1):** Wir haben durch die oben skizzierten Qualitätsmaßnahmen die Simulation hoch authentisch und damit hoch realitätsparallel gestaltet. Experten aus Werkstätten, Verbänden, Schulen und letztlich die Auszubildenden selbst verstärken unsere Einschätzung. Diese gilt es nun, über eine Dimensionalitätsanalyse empirisch zu prüfen. In der Gruppe 1 verzeichnen wir sehr hohe latente Korrelationen zwischen den Itempaketen aus der Realität (R1, R3, R5, R7) und Simulation (S2, S4, S6, S8) in Höhe von  $r=.94$ . Die gleiche Höhe (.94) erhalten wir in der Gruppe 2 zwischen den Itempaketen aus der Realität (R2, R4, R6, R8) und Simulation (S1, S3, S5, S7). Die sehr hohen Korrelationen zwischen den einzelnen Itempaketen und die Tatsache, dass dies wechselseitig in den Gruppen zutrifft, in denen die Itempakete im Sinne des Settings (Realität und Simulation) über Kreuz (*cross-over*) realisiert wurden, liefern starke Hinweise für eine eindimensionale Fähigkeitsstruktur. Auf Grund der relativ geringen Itemanzahl je Dimension und der Tatsache, dass auf Grund der Untersuchungsanlage die Korrelationen nicht auf Basis aller Probanden-Itemfacetten (eine Person löste nicht alle beide Facetten je Fehlerfall, sondern nur eine) zustande kam, werden zusätzlich die Fitwerte der je Gruppe einzeln durchgeführten Skalierungen beurteilt. Hier sind sehr gute Itemfitwerte zu konstatieren (vgl. die Langfassung). Zusammen genommen können diese Befunde als starke Indizien gesehen werden, dass zur Lösung der realen und simulierten Aufgaben gleiche Fähigkeitsbündel benötigt werden.

**zu (2):** Um Gruppendifferenzen in der abhängigen Variable (Fehleranalysefähigkeit) zu kontrollieren und damit die Fehleranalysen der einzelnen Fehlerfälle bezüglich Schwierigkeitsdifferenzen ver-

gleichbar zu machen, hatten wir in der Untersuchungsanlage eine Verankerung beider Versuchsgruppen durch den anwendungsorientierten Wissenstest vorgesehen. Die Prämisse für eine Verankerung beider Fehlerfallskalen durch den Wissenstest ist, dass die Ankeritems zur Lösung die gleiche latente Fähigkeit voraussetzen wie die zu skalierenden Fehlerfälle (auch hier: Eindimensionalitätsannahme). Als Bedingung hierfür gelten eine sehr hohe Korreliertheit der Testergebnisse aus den Fehlerfällen mit den Leistungen im Wissenstest und ferner eine günstigere Passung eines eindimensionalen Fähigkeitsmodells auf die Daten als die eines zweidimensionalen Modells (Wissen und Fehleranalyse). Das Fachwissen korreliert mit den Ergebnissen der Fehleranalyse von Gruppe 1 mit .76; in Gruppe 2 ergibt sich eine Korrelation von .80. Dieser Befund bestätigt auch die Annahme, dass für die Erfassung der Fachkompetenz in einem VET-LSA beide Testformen in Kombination eingesetzt werden sollten. Devianzstatistisch passt darüber hinaus das zweidimensionale Modell besser auf die Daten. Daraus schlussfolgern wir, dass mit dem Wissenstest eine eigenständige Facette von Fachkompetenz erhoben wurde, die zwar mit der Fehleranalyse hoch korreliert, aber nicht in ihr aufgeht und deshalb für eine Verankerung nur bedingt geeignet scheint<sup>1</sup>. Vor diesem Hintergrund sind Verzerrungen nicht auszuschließen. Eine Prüfung mittels Verankerung durch den Wissenstest wird dennoch den Befunden weiter unten ergänzend (vergleichend) gegenüber gestellt, womit das Ergebnis dieser Untersuchung multimethodisch abgesichert werden kann.

Zur Prüfung der beiden randomisiert gebildeten Untersuchungsgruppen (Gruppe 1 und Gruppe 2) auf gleiche Verteilung nutzen wir als Kriterium die Fähigkeiten im Wissenstest. Dazu wurde zunächst der Wissenstest psychometrisch beurteilt. Die Wissenstestitems erwiesen sich auf Anhieb als psychometrisch gut. Kein Item hat einen signifikant schlechten Fit. Die Reliabilität ist mit .65 (Verhältnis von EAP zu PV) bzw. .67 (Cronbach's Alpha) ausreichend. Auch wenn dies die Beantwortung unserer Fragestellung nicht beeinflusst, so sei doch mit Blick auf ein *large-scale assessment* bemerkt, dass sichere Personenverortungen (z.B. auf Kompetenzstufen) erst mit einem wesentlich höheren Reliabilitätswert vorgenommen werden können. Notwendig dafür ist eine substantielle Erweiterung des Instrumentariums, die im Rahmen der Aufgabenentwicklung für ein VET-LSA zu leisten ist.

### *Gruppenvergleich*

Gruppe 1 hat im Wissenstest einen Summenscore<sup>2</sup> von 14.82 (von 26 erreichbaren Punkten) bei einer Standardabweichung von 4.27, eine Schiefe von -0.17 und eine Kurtosis von -0.54. Die Gruppe 2 hat eine annähernd gleiche Verteilungsstruktur wie Gruppe 1: einen Summenscore von 14.46 (von 26 erreichbaren Punkten) bei einer Standardabweichung von 4.18, einer Schiefe von -0.39 und eine Kurtosis von -0.13. Für beide Gruppen können zusätzlich Normalverteilungen konstatiert werden (Kolmogorov-Smirnov-Test). Somit können die beiden Versuchsgruppen hinsichtlich des Kriteriums

---

<sup>1</sup> Selbst bei vorheriger regressionsanalytischer (mit schrittweiser Integration) Ermittlung günstigster Itempakete (sechs Wissensitems) erhöhte sich die Korrelation zwischen Gruppe 1 und Wissenstest lediglich auf .81.

Wissen als „gleiche— Gruppen aufgefasst werden und hiermit die Itemschwierigkeiten, gestützt auf die Indizien zur Eindimensionalität (s. o.), direkt verglichen werden.

**Zu (3):** Die möglichen Prüfverfahren sind (1) ANOVA, (2) DIF-Analyse und (3) scale linking. Auf Grund der bisherigen Befunde unter „zu (1)— und „zu (2)— müssten alle drei Verfahren zu ähnlichen Aussagen kommen. Alle Verfahren sollen hier dennoch in einem sich gegenseitig stützenden multime-  
thodischen Verfahren zur Anwendung kommen. Die ANOVA baut auf der gleichen Verteilung der beiden Versuchsgruppen auf (s. o.). Die DIF-Analyse berücksichtigt für die Differenzwertbeurteilung der einzelnen Fehlerfälle in der Realität und in der Simulation ergänzend die Differenz der mit den (nicht parallelisierten) Testteilen ermittelten Gruppendifferenz in der Fehleranalysefähigkeit. Das scale linking kontrolliert etwaige Gruppendifferenzen in dem Ankertestmerkmal (hier Wissenstest) zur Skalierung der Fehlerfälle.

#### *Auswertung mittels Lösungshäufigkeiten (bessere Anschaulichkeit) und ANOVA*

Im folgenden Schaubild (siehe Abb. 2) sind aufgabenspezifisch die Anteile der Auszubildenden gegenübergestellt, die keine zutreffende Diagnose stellen konnten<sup>3</sup>. Die Verteilung der Diagnoseleistungen auf der Skala verdeutlicht eine gut gelungene Verteilung der Fehlerfälle. D.h. es war auf der Basis der Erkenntnisse aus den Vorstudien und den Kooperationen mit den Experten möglich, den Schwierigkeitsgrad der Aufgaben relativ gut abzuschätzen.

Die erzielten Übereinstimmungen der Diagnoseleistungen am realen und simulierten Kfz sind bemerkenswert hoch; leichte Abweichungen weisen die Fehlerfälle 4 (8,1%-Punkte), 6 (10%) und 7 (9,9%) auf. Lediglich der Fehlerfall 3 weist eine größere Abweichung (16,3%-Punkte) auf.

---

<sup>2</sup> Die weiteren Analysen basieren auf dem Summenscore. Diese Darstellungsform wird an manchen Stellen auf Grund einfacherer Verständlichkeit und damit Kommunizierbarkeit gewählt, verändert jedoch nichts an den Aussagen.

<sup>3</sup> Der Bezug zu „Falschlösungen— ist auf Grund der Formulierung von *partial credit items* vorzuziehen. Erst die Auswertung im nächsten Kapitel wird klären, inwieweit die *partial credits* mit dem Setting interagieren.

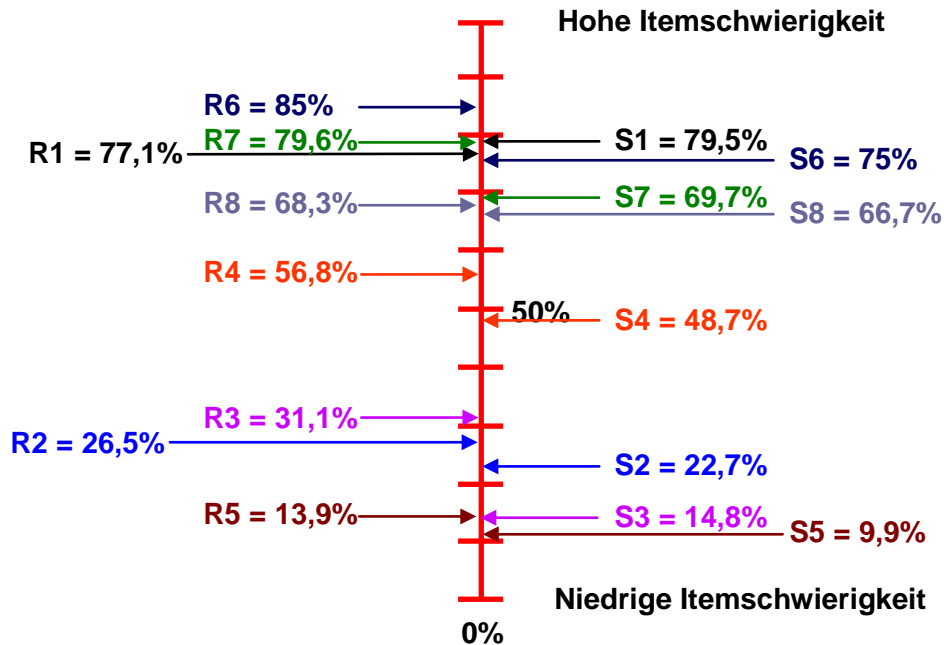


Abb.2: Lösungsquoten (falsche Lösung) aller 8 Fehlerfälle in beiden Zwillingversionen (links reale Fehlerfälle und rechts computersimulierte Fehlerfälle)

Die Auswertung mittels ANOVA erbringt in zwei Fehlerfällen (Fehlerfall 3 u. 4) signifikante Unterschiede.<sup>4</sup> In einem weiteren Fehlerfall liegt der Unterschied kurz vor der Signifikanzgrenze (0.08). Damit ist es bei 6 Fehlerfällen gelungen, gleich schwierige Items in der Realität und der Simulation zu erzeugen.

#### Auswertung mittels DIF-Analyse

Eine DIF-Analyse untersucht, ob Personen gleicher Fähigkeiten in unterschiedlichen Subgruppen sich ähnlich zu den Items verhalten. Sie dürfen sich danach hinsichtlich der Wahrscheinlichkeit, mit der einzelne Probanden einzelne Items richtig lösen, nicht unterscheiden. Das heißt, dass die Item-Responsefunktionen bzw. die Itemparameter in den Subgruppen gleich sein müssen (Hambleton/Swaminathan/Rogers 1991, S. 110). Diese Untersuchung erfolgt, nachdem die Itemschwierigkeiten zwischen den Gruppen im Hinblick auf den Gruppenfähigkeitsunterschied kontrolliert wurden.

Geprüft wurde zunächst, um welches Maß die Schwierigkeitsdifferenz der Zwillingitems korrigiert werden muss: Die beiden Versuchsgruppen Gruppe 1 und Gruppe 2 unterscheiden sich in der Dimension Fehleranalysefähigkeit um nichtsignifikante 0.074 Logits (Standardschätzfehler = 0.079; sign. = 0.643), was in etwa einem Unterschied von 8 % der Stichprobenstandardabweichung entspricht.

Auch bei der DIF-Analyse weisen sechs Items (Fehlerfälle 1, 2, 5, 6, 7, 8) keine signifikante Schwierigkeitsverzerrung auf! Bei den Items 3 und 4 zeigen sich jedoch signifikante Schwierigkeitsverzerrungen, zu deren Ursachen wir begründete Vermutungen anstellen können (s. u.).

<sup>4</sup> Wobei immer die Stichprobengröße zu berücksichtigen bleibt!

Auch mit der DIF-Analyse zeigt sich wie bei der Analyse der Lösungsquoten nahezu durchgängig, dass die Simulationsaufgaben etwas leichter ausfallen als die gleichen Aufgaben in der Realität. Dies ist sicherlich auf die höheren Komplexitäten der realen Anforderungssituationen zurückzuführen. Jedoch zeigen die gering ausfallenden Differenzen zwischen Realität und Simulation, dass es keinen systematischen Einfluss auf die Diagnoseleistung durch z.B. manuelle Anforderungen der Realität (Stecker und Abdeckungen lösen, Adapterleitungen anbringen, Messgerät einstellen und anschließen etc.) gibt, die in einer Simulation aus der Sache heraus entfallen müssen. Zusammenfassend konnten wir für sechs von acht Items zeigen, dass eine sorgfältig gestaltete Simulation sehr ähnliche Aussagen zur Leistungsfähigkeit von Auszubildenden zulässt, wie das normalerweise nur Aufgaben in der Realität zugeschrieben wird. Dies ist in Anbetracht der immensen Komplexität moderner Fahrzeugarchitekturen nicht trivial. Bei einem Item konnten wir leichte Abweichungen und bei einem anderen Item größere Abweichungen feststellen. Die leichten Abweichungen von Fehlerfall 4 (die Realität ist schwieriger) können wir uns dadurch erklären, dass wir in Anbetracht von Kostenrestriktionen eine Innenraumkomponente im Motorraum (in der Simulation der Ort aller Diagnoseschritte) visualisierten, was zu einer Vereinfachung der Realität geführt hat. Die großen Abweichungen bei Item 3 (die Realität ist schwieriger) erklären wir uns dadurch, dass wir in der Instruktionsphase für die Simulation Übungsmessungen an einem Bauteil durchgeführt haben, das für die Lösung von Item 3 relevant war. Wahrscheinlich konnte hierdurch das Bauteil leichter im Motorraum aufgefunden werden (ein notwendiger Schritt, um überhaupt diagnostizieren zu können).

*Auswertung mittels scale linking der beiden Versuchsgruppen durch den Wissenstest*

Die letzte Prüfmethode führt zu einer vergleichbaren Einschätzung wie die vorangegangenen Analysen. Eine Verankerung der beiden Gruppen durch den Wissenstest ergab keine signifikanten (mittels Einschätzung durch die Standardmessfehler) Differenzen zwischen den Items R1-S1, R2-S2, R5-S5, R7-S7 und R8-S8. Signifikante Differenzen bestehen nur zwischen den Items R3-S3, R4-S4 und R6-S6. In dieser Auswertungsvariante wird die „annähernd—ignifikante Differenz bei Fehlerfall 6, die sich auch im ersten Analyseverfahren gezeigt hat, jedoch signifikant.

**c) Zusammenfassung der Befundlage**

**Zusammenfassend kann festgehalten werden: (1) Bei konservativer Einschätzung ergeben sich bei fünf von acht komplexen Fehleranalyseaufgaben zwischen den Tests in realen und simulierten Anforderungskontexten keine signifikanten Unterschiede. Bei einer weniger strengen, jedoch noch vertretbaren Einschätzung würde diese Aussage für sechs von acht Fehlerfällen gelten. Für die bei zwei Items bestehenden Schwierigkeitsverzerrungen gibt es nahe liegende Erklärungen, die deren Vermeidung mit hoher Wahrscheinlichkeit ermöglichen. (2) Bei annähernd allen Fehlerfällen (bis auf Fehlerfall 1) scheint die Simulationsvariante trotz großen Bemühens um Authentizität und damit Komplexitätsbezug zur Realität etwas leichter als die Realität zu sein. Bei**

sorgfältiger Entwicklung ist nach unseren Ergebnissen jedoch nur mit leichten Verschätzungen zu rechnen. (3) Die Analysen zur Kompetenzstruktur weisen das Fachwissen und die Fehlerdiagnoseleistung als eigenständige Kompetenzfacetten aus, die bei der Testkonstruktion für ein VET-LSA zu berücksichtigen sind. (4) Es ist gelungen, auf der Basis der in vorausgegangenen Studien gewonnenen Erkenntnisse zu den Schwierigkeitsparametern der Aufgaben sowohl für den Fachwissenstest als auch die Fehlerdiagnosen gezielt ein wünschenswertes Schwierigkeitspektrum zu generieren. Eine verlässliche Niveaumodellierung setzt eine substantielle Erweiterung des Tests zur Fehleranalysefähigkeit voraus.

**Damit sind insgesamt sehr günstige Voraussetzungen geschaffen, um fachliche Kompetenzen valide zu erfassen.**

## **4.2 Anlage und Ergebnisse der Validierungsstudie im kaufmännischen Bereich**

Im Rahmen der Validierungsstudie des kaufmännischen Bereichs werden Struktur und Graduierung berufsfachlicher Kompetenz dargestellt. Die Ergebnisse beziehen sich auf die Erfassung kaufmännischer Kompetenz mit Hilfe authentischer beruflicher Simulationen. Für einen umfassenden Bericht sei auf die Langfassung der Studie verwiesen.

### **4.2.1 Anlage der Studie**

Für die Validierungsstudie wurde kaufmännische Kompetenz als Fähigkeit definiert, auf Grundlage eines systemischen Verstehens betrieblicher Teilprozesse und deren Rekonstruktion aus realen Unternehmensdaten in berufsrealen Situationen unternehmerische Entscheidungen treffen und diese validieren zu können, um damit das eigene Wissens- und Handlungspotential vor dem Hintergrund der Entwicklung individueller beruflicher Regulationsfähigkeit auszubauen (WINTHER/ACHTENHAGEN 2008). Mit dieser Definition wird vor allem die Bedeutung der kompetenten Erstellung angemessener Handlungspläne hervorgehoben.

Zur Messung kaufmännischer Kompetenz werden domänenspezifische Geschäftsvorfälle herangezogen. Unter Geschäftsvorfall wird in diesem Zusammenhang eine über konkrete Arbeitsprozesse definierbare Anforderungssituation verstanden, die sich in unternehmensspezifischen Geschäftsprozessen verorten lässt. Ein Geschäftsvorfall stellt damit eine inhaltlich-systematisch ausgestaltete arbeitsplatzspezifische Situation dar, die sich detailliert im Hinblick auf das Anforderungsniveau, die Handlungsspielräume sowie die intendierten Zielsetzungen beschreiben lässt (WINTHER, in Vorbereitung). Im Hinblick auf die Intentionen eines Large Scale-Assessment in der kaufmännischen beruflichen Bildung wurden solche Inhalte ausgewählt, die auch für einen internationalen Vergleich geeignet wären: Geschäftsvorfälle aus dem Bereich der betrieblichen Wertschöpfung sowie Geschäftsvorfälle, die betriebliche Steuerungsprozesse abbilden. Mit diesen Geschäftsvorfällen wird eine große Spannweite der Tätigkeiten von Industriekaufleuten abgedeckt, die es gestatten, sowohl von der Tiefe als auch von der Breite der Anforderungen her das Konstrukt der beruflichen Handlungskompetenz objektiv, reliabel und valide abzubilden. Für den Entwurf der Geschäftsvorfälle war dabei wichtig, welche Art von Ar-

beits- und Geschäftsprozessen abgebildet werden sollten. Grundlage für diese Aufteilung war die Annahme, dass Wertschöpfungs- bzw. Steuerungsprozesse eine unterschiedliche kognitive Verarbeitungsstruktur aufweisen. Aufgrund der Unterschiede in den Anforderungen darf vermutet werden, dass die Auszubildenden zu ihrer Bewältigung auf unterschiedliche berufsfachliche Fähigkeiten zurückzugreifen haben. Daneben wurde versucht zu erfassen, welche verstehensbasierten Kompetenzen gegeben sind. Die entsprechenden Items sind auf die Adaptation ausgewählter betriebswirtschaftlicher Konzepte in den zwei betrieblichen Geschäftsprozessen bezogen.

Über das gewählte Testformat wurden

- reale Arbeitsprozesse (Verhandlungen, Arbeitsverteilung in der Interaktion),
- reale Geschäftsprozesse (z. B. die konkrete Abwicklung einer Bestellung),
- reale kaufmännische Entscheidungen (z. B. Ermittlung eines Liefertermins)

abgebildet. Dies geschieht mit Hilfe der Unternehmenssimulation ALUSIM. Diese ist web-basiert und enthält

- eine allgemeine Einführung in die Unternehmensstruktur (einschließlich der Unternehmenshistorie);
- einen animierten Arbeitsplatz (Schreibtisch mit Zugriff auf sämtliche Ablagen und die simulierte Software) sowie
- ein animiertes Sideboard, über die die web-basierten Situationen gesteuert werden;
- allgemeine Zusammenstellungen zur geschäftlichen Lage des Unternehmens (textbasierte Analyse der Geschäftszahlen sowie Darstellungen der Bilanz, Erfolgsrechnung und Geldflussrechnung).

In der Unternehmenssimulation sind die Bereiche Vertrieb, Einkauf als Beispiele für Wertschöpfungsprozesse und der Bereich der Arbeitsvorbereitung als Bestandteil eines betrieblichen Steuerungsprozesses verarbeitet. Für diese Bereiche wurden typische Arbeitssituationen als Modelle zur Abbildung von domänenspezifischen Geschäftsvorfällen herausgegriffen (Für die Darstellung des Vorgehens sei auf die Langfassung des Berichts verwiesen).

Die zentrale Herausforderung des Projekts war es, ein Testformat zu entwickeln, über das zum einen verschiedene berufliche Inhaltsbereiche mit ihren komplexen Zugriffs- und Handlungsebenen und zum anderen der Aktionsraum der Auszubildenden im Hinblick auf vollständige Arbeits- und Geschäftsprozesse in ihrem ganzheitlich zeitlichen und organisationalen Umfang abgebildet werden. Mit der Berücksichtigung dieser Anforderungen sollte der Anspruch abgesichert werden, berufliche Handlungskompetenz in ihren verstehensbasierten sowie – und vor allem – handlungsbasierten Strukturen abzubilden. Hierfür wurden (1) konzeptuale verstehensbasierte Anwendungsaufgaben und (2) prozessuale handlungsbasierte Simulationsaufgaben konstruiert, die in der web-basierten Testumgebung – ALUSIM – eingebunden sind.

#### 4.2.2 Beschreibung der Stichprobe und der Testitems

Es wurden insgesamt 264 Auszubildende (3. Ausbildungsjahr, sieben Berufsschulen, 61 Betriebe, drei Bundesländer) erfasst; 56,5 Prozent von ihnen sind weiblich. Das Alter der Auszubildenden variiert zwischen 18 und 34 Jahren, wobei 83,4 Prozent der Befragten in der Altersgruppe von 20 bis 23 Jahren liegen – dies ist typisch für eine Abschlussklasse im Rahmen einer dreijährigen kaufmännischen Berufsausbildung (für eine ausführliche Charakteristik des Samples sei auf die Langfassung der Studie verwiesen). Von den 264 Auszubildenden, die an der Erhebung teilgenommen haben, haben 43 ausschließlich den Testbereich „Einkauf—Testgruppe 2) und 39 ausschließlich die Testbereiche „Vertrieb—und „Arbeitsvorbereitung—Testgruppe 3) im Rahmen der Simulation ALUSIM gelöst; die von den Auszubildenden in den Testgruppen 2 und 3 jeweils nicht behandelten betrieblichen Bereiche waren in Form von authentisch inszenierten Arbeitsproben in ihrem Ausbildungsbetrieb zu bearbeiten. Beide Testformate sind durch Ankeritems miteinander verbunden, so dass durch eine Verlinkung der jeweiligen Teilstichproben generalisierbare Aussagen möglich sind. Hierfür ist allerdings der Rücklauf der Arbeitsproben aus den Betrieben noch zu gering, so dass über den Erfolg des Einsatzes der betrieblichen Arbeitsproben zur Zeit keine Aussagen möglich sind. Allerdings zeichnet sich deutlich ab, dass – wie schon in 1.2 erwähnt – sich betriebliche Arbeitsproben kaum als Erhebungsinstrument eignen.

	Testbereich Einkauf	Testbereich Vertrieb	Testbereich Arbeitsvorbereitung	
Testgruppe 1	182	182	182	182
Testgruppe 2	43	---	---	43
Testgruppe 3	---	39	39	39
INSGESAMT	225	221	221	264

**Abb. 3: Design der Erhebung (mit Angabe der Zahl der Testpersonen)**

Die einzelnen Testbereiche sind wie folgt über Items der handlungsbasierten Simulationsaufgaben und verstehensbasierten Anwendungsaufgaben repräsentiert (Abb. 47):

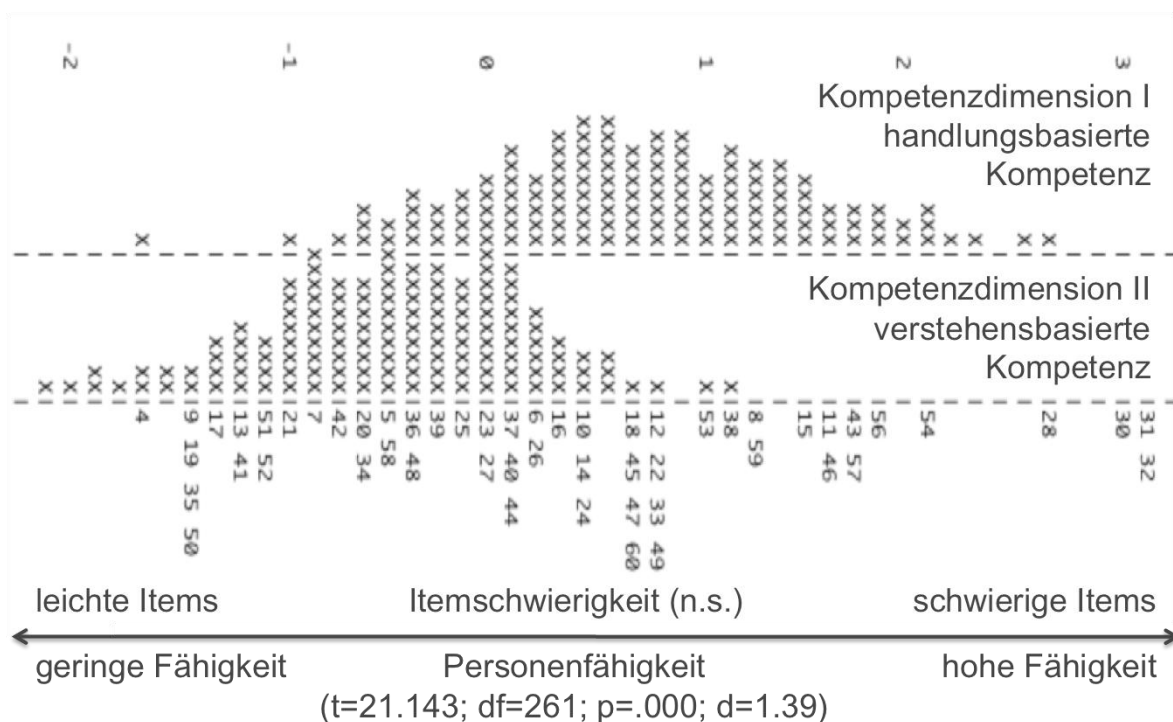
	Testbereich Einkauf	Testbereich Vertrieb	Testbereich Arbeitsvorbereitung	
Itemanzahl der handlungsbasierten Simulationsaufgaben	18	12	4	34
Itemanzahl der verstehensbasierten Anwendungsaufgaben	11	12	3	26
INSGESAMT	29	24	7	60

**Abb. 4: Itemanzahl in den zwei verschiedenen Testformen**

### 4.2.3 Zentrale Ergebnisse

#### a) Unterschiede zwischen den Dimensionen beruflicher Handlungskompetenz

Zur Abbildung der Unterschiede zwischen den Dimensionen beruflicher Handlungskompetenz werden zwei latente Variablen miteinander verglichen: die Personenfähigkeit, handlungsbasierte Simulationen innerhalb betrieblicher Situationen lösen zu können (handlungsbasierte Kompetenz), und die Personenfähigkeit, verstehensbasierte Anwendungsaufgaben vor dem Hintergrund betrieblicher Situationen zu bewältigen (verstehensbasierte Kompetenz). Die Ergebnisse auf Basis des Multidimensional Random Coefficients Multinomial Logit Model (ADAMS/WILSON/WANG, 1997) bestätigen das Vorliegen dieser unterschiedlichen Kompetenzdimensionen. Damit steht für ein VET-LSA für die Industriekaufleute ein empirisch geprüftes Kompetenzstrukturmodell zur Verfügung.



**Abb. 5: Dimensionen beruflicher Handlungskompetenz**

Die Abbildung verdeutlicht, dass die verschiedenen Testaufgaben unterschiedliche Fähigkeitsstrukturen der Auszubildenden ansprechen. Eine zweidimensionale Lösung, die zwischen handlungsbasierter und verstehensbasierter Kompetenz differenziert, dominiert auf Basis der Chi-Square-Statistik eine eindimensionale Kompetenzstruktur (Differenz der Devianzen=101,91;  $df=3$ ;  $p<.001$ ). Ein Vergleich der Verteilungen der Personenfähigkeiten zeigt, dass Anforderungssituationen, die auf die handlungsbasierte Kompetenz der Auszubildenden abstellen, besser gelöst werden als diejenigen, die das konzeptuale Verständnis der Auszubildenden in beruflichen Anforderungssituationen erfassen. Die praktische Signifikanz dieses Befundes ist mit  $d=1,39$  extrem hoch.

### **b) Kompetenzstufen beruflicher Handlungskompetenz**

Ein zentrales Problem bei der Bewertung der Skalen beruflicher Handlungskompetenz besteht darin, ob sie sich – wie das auch in den PISA-Studien geschehen ist – in einzelne, sinnvoll interpretierbare Stufen im Hinblick auf die Bewältigung der Anforderungssituationen zerlegen lassen. Für den Zusammenhang zwischen Anforderungssituationen und Antwortverhalten wurden Merkmale von Anforderungssituationen identifiziert, mit deren Hilfe sich unterschiedliche Komplexitätsgrade der betrieblichen Anforderungssituationen bestimmen lassen, so nach dem Umfang und der Qualität der Modellierungsschritte zum Verständnis der Arbeitssituation, nach der inhaltlichen Beherrschung der Arbeitssituation sowie nach der kognitiven Beanspruchung durch die Arbeitssituation

(WINTHER/ACHTENHAGEN 2008, 2009; WINTHER, in Vorbereitung). Unter Berücksichtigung dieser Parameter lassen sich mit Hilfe von Regressionsanalysen Stufen in der Ingesamtskala bestimmen, die inhaltlich sinnvoll interpretierbar sein sollten. Für die Validitätsstudie ist dieses gegeben. Es lassen sich eindeutig vier Kompetenzstufen bestimmen; dabei entspricht die Kompetenzstufe II: „Kaufmännisches Handlungs- und Aktionswissen“ den Anforderungen, wie sie in der Ausbildungsordnung und im Rahmenlehrplan für den Ausbildungsberuf „Industriekaufmann/Industriekauffrau“ vorgegeben sind.

In Abb. 69 sind die Kompetenzstufen über die Logit-Grenzen definiert; zusätzlich zu den Schwellenwerten ist die prozentuale Zuordnung der Auszubildenden auf die einzelnen Kompetenzstufen angegeben. Der Anteil der Auszubildenden in den jeweiligen Kompetenzstufen im Hinblick auf die zwei verschiedenen Kompetenzdimensionen unterscheidet sich deutlich: Wenn die Kompetenzstufe 2: Kaufmännisches Handlungs- und Aktionswissen, als Anspruch des Arbeitsmarktes an Absolventen des Dualen Systems in Anlehnung an die Vorgaben anzusehen ist, können im Bereich der handlungsbasierten Kompetenz 22,6 Prozent und im Bereich der verstehensbasierten Kompetenz 78,2 Prozent der getesteten Auszubildenden diesen Anspruch nicht adäquat erfüllen. Zu einer Relativierung der Ergebnisse sei an dieser Stelle erwähnt, dass (1) die Stichprobe ausschließlich Auszubildende im Rahmen des 3-jährigen Ausbildungsganges enthält, so dass die erwartet leistungsstarken Auszubildenden insbesondere im Bereich der verstehensbasierten Kompetenz nicht in der Masse erfasst werden konnten, und dass (2) die Ergebnisse im handlungsbasierten Kompetenzbereich weit positiver ausgefallen sind, als dies vor dem Hintergrund vergleichbarer Studien (ULME III: LEHMANN/SEEBER 2007; TOSCA: KÖLLER/WATERMANN/TRAUTWEIN/LÜDTKE, 2004) hätte angenommen werden können.

Kompetenzstufe	Stufenschwelle der Kompetenzbereiche	Anteil der Auszubildenden auf den Stufen der handlungsbasierten Kompetenz (in %)	Anteil der Auszubildenden auf den Stufen der verstehensbasierten Kompetenz (in %)
unter Kompetenzstufe 1	-1,479	3,10	35,50
Kompetenzstufe 1: Kaufmännisches Grund- und Regelwissen	- 0,723	19,50	42,70
Kompetenzstufe 2: Kaufmännisches Handlungs- und Aktionswissen	0,117	52,60	21,40
Kompetenzstufe 3: Kaufmännisches Analysewissen	1,410	22,50	0,40
Kompetenzstufe 4: Kaufmännisches Entscheidungswissen	2,703	2,30	---

**Abb. 6: Kompetenzstufen und prozentuale Verteilung der Auszubildenden**

Im Hinblick auf die **verstehensbasierte Kompetenz** ist das Ergebnis absolut erwartungskonform; es repliziert in hohem Maße die Befunde des ULME III-Projektes (Untersuchung von Leistungen, Motivation und Einstellungen in der beruflichen Bildung; LEHMANN/SEEBER, 2007). Die Befunde zeigen, dass die Leistungen im Ausbildungsberuf Industriekaufmann/Industriekauffrau tendenziell eine bimodale Verteilung mit einem Gipfel im unteren Leistungsbereich und einem flachen Kurvenverlauf zum oberen Leistungsspektrum hin aufweisen. Die Leistungsschwächen beziehen sich dabei insbesondere auf Testaufgaben, die das Verstehen und Interpretieren ökonomischer Beziehungen zum Inhalt haben (LEHMANN/SEEBER, 2007, S. 140). Auch für die vorliegende Studie zeigt sich, dass ein Großteil der Auszubildenden (42,70 Prozent) zwar über kaufmännisches Grund- und Regelwissen verfügt, das jedoch zu unflexibel ausgebildet ist, um es variabel im Rahmen beruflicher Anforderungssituationen einsetzen zu können.

Für den Testbereich der **handlungsbasierten Kompetenz** fällt insbesondere positiv auf, dass nur drei Prozent der Auszubildenden unterhalb der Kompetenzstufe 1 liegen und dass 52,6 Prozent die Kompetenzstufe 2, die von den Autoren als Mindestanspruchsstufe des Arbeitsmarktes definiert wurde, und weitere 24,80 Prozent die darüber liegenden Kompetenzstufen erreichen konnten.

#### 4.2.4 Schlussfolgerung

Bezogen auf die Entwicklung der Testbereiche und der Testitems ist Folgendes zusammenfassend festzuhalten:

- Im Zentrum der Arbeit stand die Konstruktion von Testsituationen, deren Angemessenheit sowohl im Hinblick auf betriebliche als auch auf schulische Anforderungen mehrfach geprüft wurde.
- Diese Anforderungen sind in der Simulation angelehnt an Bedingungen, wie sie im Hinblick auf die Erfüllung betriebs- und arbeitsspezifischer Vorgaben als berufstypisch gelten. Zudem wurden die generellen Zielsetzungen für die Berufsausbildung berücksichtigt.

Die Analysen haben gezeigt, dass die konstruierten Testitems eine sehr umfassende Skala zur Erfassung handlungsbasierter Kompetenzen repräsentieren und dass sich handlungsbasierte Kompetenz empirisch von verstehensbasierter Kompetenz trennen lässt. Es ist folglich möglich, empirisch gesichert unterschiedliche Fähigkeitsstrukturen der Auszubildenden in Entsprechung zu den Zielen der Dualen Ausbildung separiert zu erfassen und aufeinander zu beziehen. Bei der Auswertung kam es darauf an zu prüfen, ob sich die Anforderungssituationen mit dem gewählten Format dafür eignen, in einem international vergleichenden VET-LSA Verwendung zu finden. Die Aussage, die wir aufgrund des gesamten Vorgehens sowie der erhaltenen Daten treffen können, lautet uneingeschränkt: Ja!

## Literatur

- Adams, R. J./Wilson, M./Wang, W.-C. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement*, 21 (1), 1—23.
- Becker, M. (2005): Einbindung von Facharbeiterkompetenzen in IKT-dominante Diagnoseabläufe im Kfz-Service. In: Pangalos, J./Spöttl, G./Knutzen, S./Howe, F. (Hrsg.): Informatisierung von Arbeit, Technik und Bildung. Münster: LIT, S. 45-54
- Gschwendtner, T. (2008): Ein Kompetenzmodell für die kraftfahrzeugtechnische Grundbildung. In: Nickolaus, R./Schanz, H. (Hrsg.): Didaktik gewerblich-technischer Berufsbildung. Hohengehren: Schneider, S. 103-119
- Gschwendtner, T./Geißel, B./Nickolaus, R. (2007): Förderung und Entwicklung der Fehleranalysefähigkeit in der Grundstufe der elektrotechnischen Ausbildung. In: bwp@, Ausgabe 13
- Hägele, T. (2002): Modernisierung handwerklicher Facharbeit am Beispiel des Elektroinstallateurs. Hamburg, Univ., Diss. (<http://www.sub.uni-hamburg.de/opus/volltexte/2002/787>)
- Hambleton, R. K./Swaminathan, H./Rogers, H. J. (1991): Fundamentals of Item Response Theory. Newbury Park (CA): SAGE Publications
- Köller, O./Watermann, R./Trautwein, U./Lüdtke, O. (Hrsg.) (2004). Wege zur Hochschulreife in Baden-Württemberg. TOSCA – Eine Untersuchung an allgemein bildenden und beruflichen Gymnasien. Opladen: Leske und Budrich
- 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
- Nickolaus, R./Heinzmann, H./Knöll, B. (2005): Ergebnisse empirischer Untersuchungen zu Effekten methodischer Grundentscheidungen auf die Kompetenz- und Motivationsentwicklung in gewerblich-technischen Berufsschulen. In: ZBW, 101. Bd., H. 1, S. 58-78
- Nickolaus, R./Knöll, B./Gschwendtner, T. (2006): Methodische Präferenzen und ihre Effekte auf die Kompetenz- und Motivationsentwicklung – Ergebnisse aus Studien in anforderungsdifferenten elektrotechnischen Ausbildungsberufen in der Grundbildung. In: ZBW, 102. Bd., H. 4, S. 552 – 577
- Winther, E. (in Vorbereitung). Kompetenzgestützte Bildungsmobilität. Habilitationsschrift. Professur für Wirtschaftspädagogik. Georg-August-Universität Göttingen
- Winther, E./Achtenhagen, F. (2008). Kompetenzstrukturmodell für die kaufmännische Bildung. Zeitschrift für Berufs- und Wirtschaftspädagogik, 104, 511 – 538
- Winther, E./Achtenhagen, F. (2009). Measurement of Vocational Competencies – A Contribution to an International Large-Scale-Assessment on Vocational Education and Training. Empirical Research in Vocational Education and Training, 1, 88–106