Tobias Gschwendtner / Stephan Abele / Reinhold Nickolaus

Computersimulierte Arbeitsproben: Eine Validierungsstudie am Beispiel der Fehlerdiagnoseleistungen von Kfz-Mechatronikern

Kurzfassung: Im Kontext verschiedener Arbeiten zur Vorbereitung eines Large-Scale Assessments in der beruflichen Bildung (VET-LSA) gehen wir in unserem Beitrag der Frage nach, wie valide computerbasierte Simulationen realer Arbeitsproben zur Kompetenzmessung sind. Dabei wollen wir prüfen, was bei einer inhaltstreuen Überführung beruflicher Realität in computersimulierten Umwelten genau passiert, d.h. wie ändern sich die Anforderungen und damit die Schätztreue von Kompetenzen? Diese Fragestellung untersuchen wir an Hand eines Zweigruppenversuchsplans mit randomisierter Gruppenbildung und N = 257 Kraftfahrzeugmechatronikern des dritten und vierten Ausbildungsjahres. Im Rahmen des Projekts wurden in enger Kooperation mit Experten aus dem Feld des Kraftfahrzeugwesens acht authentische Fehlerfälle entwickelt, die sowohl im Kraftfahrzeug selbst als auch in simulierter Form realisiert wurden. Beiden Versuchsgruppen wurden jeweils vier reale und vier simulierte Fehlerfälle zur Bearbeitung gegeben und anschließend die Daten mittels klassischer Analyseverfahren (Häufigkeitstabellen, Chi-Quadrat-Statistiken), ergänzt um Verfahren mit latenten Variablen (latente Korrelationen, Analyse von Itemfitwerten) ausgewertet. Dabei zeigen sich nur geringe Testmoduseffekte zwischen den beiden Darbietungsformen „Realität“ und „Virtualität“. Mit diesen Erkenntnissen empfehlen sich computerbasierte Messverfahren zur validen Erfassung zentraler Kompetenzaspekte.

Abstract: Within the preparation for a Large-Scale Assessment in Vocational Education and Training (VET-LSA), we address in our contribution the question how valid computer-based simulations of real work samples are to measure competencies. We want to examine what exactly happens when we transfer real-life content to a computer-based virtual environment. More precisely, how do the requirements and thus the fidelity of abilities alter? We investigate this question on the basis of a randomised equivalent-two-group experimental design with N = 257 car mechatronics apprentices attending the 3rd and 4th training year. Within this study we developed eight authentic malfunction scenarios in close cooperation with car mechatronic experts. All malfunction scenarios were implemented in real cars and in a computer-based simulation. Both experimental groups were confronted with four real and four simulated tasks. The data was analysed using classical analytical techniques (frequency tables, chi-square statistics), supplemented by procedures with latent variables (latent correlations, analysis of item-fit). In sum, only small test-mode effects between reality and virtuality arose. According to these findings computer-based measurement procedures can be recommended for a valid acquisition of main competency aspects.

1. Einleitung, Grundgedanken zu Assessmentverfahren, Zielsetzung und Fragestellung der Validierungsstudie

Die Frage, die sich gegenwärtig vor allem in Anbetracht der Anforderungen eines europäischen Large-Scale-Assessments in der beruflichen Bildung (VET-LSA oder „Berufsbildungs-PISA“) (Baethge et al. 2006) stellt, ist die nach möglichst validen, reliablen, hoch standardisierten (objektiven), praktikablen und gleichzeitig kosteneffizienten Messinstrumenten zur Messung beruflicher Handlungskompetenzen, speziell für große Stichproben. Nach Luecht und Clauser (2002) werden computerbasierte Testverfahren zur Messung solcher komplexen und dynamischen Kompetenzbündel für günstig erachtet. Sie „konkurrieren“ mit gängigen Messanordnungen wie

Beobachtungsverfahren von Tätigkeiten am Arbeitsplatz (im Arbeitsprozess) oder retrospektiver Begutachtung von Arbeit an Hand von Dokumentationen sowie Produkten getaner Arbeit, realen Arbeitsproben, paper-pencil Tests oder Interviews (Fachgespräche). Computerbasierte Testverfahren können im einfachsten Fall als computerisierte Duplikate von paper-pencil Tests aufgefasst werden, bei umfangreicher Abbildung von realen Kontexten und Interaktionsmöglichkeiten des Nutzers mit diesen spricht man von einer Simulation. Dabei können simulierte Arbeitsproben oder auch, bei sehr umfassender Gestaltung, simulierte Arbeitsprozesse mit komplexen Kommunikationsinterakten realisiert werden.

Klar ist, dass es, mit welchem Zugang auch immer, generell nicht möglich ist, simultan und gleich hoch ausgeprägt alle Testgütekriterien (Validität, Reliabilität, Objektivität und im weiteren Sinne Praktikabilität und Ökonomie) zu realisieren, womit notwendigerweise Messkompromisse einzugehen sind. Dabei stellen sich Probleme auf zwei Ebenen: (1) Wie verhält es sich mit den Gütekriterien in den gängigen Messverfahren? Besonders virulent ist sicherlich das Problem mangelnder Kenntnis darüber, welche Testzuschnitte wie valide hinsichtlich beruflicher Performanz sind. (2) Wie ist das Verhältnis (im Sinne der Validität) zwischen computerbasierten Simulationen und einem gängigen Messverfahren (Arbeitsprobe) und wie sind die andere Gütekriterien hierbei einlösbar? Auf dieser Ebene stellen wir Anlage und Ergebnisse eines eigenen Forschungsprojekts vor, das der Frage nachgeht, ob sich zentrale fachliche Kompetenzaspekte auf der Basis von Computersimulationen technischer Systeme valide abschätzen lassen.

Zuerst einige Gedanken zur ersten Betrachtungsebene: In unserer Disziplin existieren unterschiedliche Verständnisse und damit Ansätze für Testinstrumente zur Erfassung beruflicher Handlungskompetenzen, die als Abbild individuell verschieden bewältigter Unsicherheiten im Umgang mit der messtechnischen Abbildbarkeit der Komplexität ganzheitlicher beruflicher Handlungen verstanden werden können. Generell zeigen sich die Unterschiede in drei Richtungen: (1) Es werden unterschiedlich viele Kompetenzaspekte (berufsfachlich, sozial-kommunikativ, personal) betrachtet, (2) die Kompetenzaspekte werden entweder gleichzeitig (in einem Instrument integrativ) oder additiv (mit verschiedenen Instrumenten) erfasst und (3) zielen diese Betrachtungen entweder auf externe Tätigkeiten (Handlungen resp. Performanzen) oder interne Bedingungen (Dispositionen resp. Kompetenzen) der Probanden. Gleichgültig um welchen Ansatz es sich handelt, ist die Schlüssigkeit der Ableitung von Messkonzepten aus der Konzeptualisierung beruflicher Handlungskompetenzen bisher nicht gegeben, da deren Rückbindung (empirische Validierung) an breitere Ausschnitte der realen Arbeit hinsichtlich der Zielvariable tatsächlicher beruflicher Handlungskompetenzen unseres Wissens noch nicht befriedigend angegangen wurde. Es reicht nicht hin, zu propagieren, man messe berufliche Handlungskompetenz ohne es eigentlich zeigen zu können. Somit sind „blinde Flecken" bisher sichtbarer als Bemühungen, diese in Form von Grundlagenforschung zu erhellen. Im Folgenden wollen wir kurz die vorherrschenden Messvorstellungen hinsichtlich des Kriteriums Validität diskutieren und auf die obigen Einschätzungen beziehen.

In der Untersuchung KOMET werden beispielsweise sprachproduktionslastige paper-pencil Tests („Fachaufsätze")[1] zu komplexen Problemstellungen mit „Gestal-

1 In Form von kurzen Beschreibungen eines Kundenwunschszenarios zu einer technischen Realisierung wird den Auszubildenden die Anweisung gegeben, eine möglichst umfassende Umsetzung des Kundenauftrags schriftlich in „Aufsatzform" vorzunehmen.

tungsspielräumen“[2] eingesetzt, die mit einem Schreibaufwand von 2 Stunden je Arbeitsauftrag einhergehen (Haasler/Erdwien 2009, S. 153). Die Fachaufsätze werden anschließend einem kategorialen Rating unterzogen[3]. Die ausgesprochen offen gestalteten Items[4] und die den Auszubildenden nicht explizit gemachten Bewertungskriterien werfen die Fragen auf, ob es forschungsethisch und vor allem inhaltlich überhaupt vertretbar ist, so zu verfahren[5]. Ganz zu schweigen von Sprachproduktionsbarrieren (d.h. motivational ungünstige Attributionen, sich schreibend nicht hinreichend mitteilen zu können) der gewerblich-technischen Klientel und der Einflüsse von sprachlichen Kompetenzen auf Leistungen in Fachkompetenztests (Gschwendtner 2008), die das eigentlich zu messende Merkmal mit unnötigen Messfehlern aufladen und zusammen mit den weiter oben angeführten inhaltlichen Bedenken wahrscheinlich invalide gegenüber dem eigentlich zu messenden Gegenstand werden lässt[6].

2 Gestaltungsspielräume von Aufgaben ermöglichen nach Rauner die Aufdeckung mehrerer Aspekte beruflicher Handlungskompetenz.

3 Die Ratingkategorien sind Funktionalität, Anschaulichkeit/Präsentation, Wirtschaftlichkeit, Gebrauchswertorientierung, Geschäfts- und Arbeitsprozessorientierung, Umweltverträglichkeit, Kreativität der Lösung, Sozialverträglichkeit (Rauner et al. 2009b, S. 8).

4 Die Begründung von Rauner et al. (2009b, S. 10ff.) für das von ihnen gewählte Testformat ist: „Offene Testaufgaben eignen sich in besonderer Weise für das Erheben beruflicher Kompetenzen, da sie die berufliche Wirklichkeit insofern repräsentieren, als berufliche Aufgaben in der Regel mehr oder weniger *zweckmäßig* (Hervorhebungen im Original; d.V.) gelöst werden.“ Ferner gelte: „Das Konzept der *beruflichen Validität* (Hervorhebungen im Original; d.V.) der Testaufgaben legt methodisch den Einsatz offener Testaufgaben nahe. (ebd., S. 11)“. Aus unserer Sicht sollte zwischen dem Grad an Offenheit des Itemstamms und des Antwortformats differenziert werden. Für das Antwortformat mag eine gewisse Öffnung angemessen sein, für den Itemstamm wäre ein höherer Grad an Konkretisierung sicherlich zielführend.

5 Rauner et al. (2009a) bewerten offene Formate mit z.T. unklaren Kategorien, wie bspw., ob eine technische Zeichnung „farbig“ angefertigt wurde (ebd., S. 181), ob eine Kostenrechnung aufgeführt wurde oder ob expressis verbis umweltfreundliche, sozialverträgliche und kreative Lösungen angesprochen wurden (ebd., S. 182ff.). Aus der Aufgabenstellung, die lediglich nach „möglichst vollständige(n) Unterlagen zur Realisierung der Steuerung“ und einer umfassend und detaillierten Begründung fragt (ebd., S. 180), kann der Auszubildende diese Bewertungskriterien jedoch nicht extrapolieren. Unklar bleibt, auf welcher Basis die Testperson die ökonomischen Überlegungen anstellen kann, stehen doch nur technische Tabellenbücher, Fachbücher, eigene Mitschriften und der Taschenrechner zur Verfügung (ebd., S. 180). Außerdem stellt sich die Frage: Wie wichtig bzw. schlüssig ist es, „sozialverträgliche“ Maßnahmen, wie das Erwähnen, dass man Arbeitskleidung (ebd., S. 182) tragen würde oder eine kreative Lösung in der Bewertung zu berücksichtigen? Und wie geschieht dies eigentlich genau? Zusammengenommen scheint es angebracht, trotz der adretten Robe von Handlungsorientierung und Beruflichkeit der Aufgaben und des Kategorienschemas erstens zu reflektieren, ob die vom Probanden eingeforderten Leistungen zur Messung von Facharbeit adäquat oder überhöht sind und zweitens wie sich solche eher planenden (konstruktiven) Aufgaben auf konkretere, alltägliche und (aus der Sicht des Experten) repetitivere berufliche Handlungen übertragen lassen.

6 Hinweise darauf könnte ein in KOMET dargestellter Vergleich zwischen zwei Querschnittserhebungen liefern (Rauner et al. 2009b, S. 25ff.): Auszubildende des 2. und des 3. Lehrjahres (Zeitdifferenz von 10 Monaten) wurden verglichen und ein mangelnder „Zugewinn“ an Testleistung festgestellt. Die Autoren entwerfen eine Stagnations- oder Durchhängerhypothese und erklären diese durch die gedehnte Prüfungspraxis der Abschlussprüfung zwischen Teil 1 und 2 und den damit einhergehenden, nur punktuell vor den Prüfungsteilen vorhandenen lernhaltigen Übungszeiten und einem „Lorbeereffekt“ (sich auf den eigenen Leistungen nach der Abschlussprüfung Teil 1 auszuruhen). Wir halten eine andere Erklärung für mindestens ebenso plausibel: Das Testformat ist wahrscheinlich nicht sensitiv genug gegenüber Veränderungen der abhängigen Variablen, damit nicht „kompetenzindikativ“ und somit invalide.

In einer anderen Untersuchung, dem Modellversuch GAB, wurde sehr ähnlich zu KOMET mit Evaluationsaufgaben gearbeitet (siehe hierzu Bremer 2009). Auf einer eher testphilosophischen Ebene sprechen sich Becker und Spöttl (2008) dafür aus, dass „berufliche Kompetenz holistisch zu erfassen" (ebd., S. 213) sei und nicht isoliert (in Itemform) getestet werden dürfe, „da eine Variablenisolation zu einer Demontage der untersuchten Zusammenhänge und damit zu einem Nachweis von Validität eines unzulässigen Gegenstandsbereichs führen würde. (…) Jedes Merkmal wäre für sich genommen erfassbar und würde dennoch die berufliche Kompetenz, über die ein Facharbeiter verfügt, nicht angemessen kennzeichnen" (ebd., S. 204). Damit sprechen sie sich gegen Tests zur Erfassung kognitiver Leistungsdispositionen und ebenso gegen Simulationen aus. So schreiben die Autoren: „Ebenso ist eine Messaufgabe unter Laborbedingungen nicht vergleichbar mit derjenigen, die direkt an einer technischen Anlage durchgeführt wird, die sich gänzlich anders verhält und zugänglich ist, wie ein Laborobjekt (ebd., S. 204)". Leider bleibt bei den Autoren weitgehend unklar, was sie konstruktiv andenken, evtl. Beobachtungsverfahren im Arbeitsprozess kombiniert mit Fachgesprächen (ebd., S. 209 ff.) oder ein Zugang über berufliche Arbeitsaufgaben (ebd., S. 207)? Dies ist sicherlich ein valider Zuschnitt, jedoch wissen wir, dass sich in technischen Systemen die Problemcharakteristiken und die Anforderungen an die Probanden mit leichten Veränderungen im Problemraum zum Teil beträchtlich unterscheiden (Gschwendtner/Geissel/Nickolaus 2007), was bei einer betriebsspezifischen Messung mit dem Anspruch der Objektivität konfligiert, abgesehen von den sich ergebenden Reliabilitätsproblemen vieler verschiedener Rater vor Ort. Ferner bleibt offen, welche Kriterien der Evaluation beruflicher Arbeitsaufgaben unterliegen können und wie es dann um die Einlösung zentraler Gütekriterien bestellt ist.

Der Stuttgarter Arbeitskreis präferierte in seinen Untersuchungen bislang paper-pencil Tests und simulierte Arbeitsproben zur Erfassung von Fachwissen und fachspezifischer Problemlösefähigkeit und Selbsteinschätzungsskalen zur Erfassung affektiv-motivationaler, metakognitiver und arbeitsplatzbezogener Merkmale (Nickolaus/Knöll/Gschwendtner 2006; Geissel/Gschwendtner/Nickolaus 2009). Diese Messzuschnitte gewähren meist gute bis sehr gute Reliabilitäten. Umgekehrt gibt es auch Einschränkungen: Unsere eigenen Untersuchungen waren in der beruflichen Grundbildung angesiedelt und lassen eher Aussagen über die Messung beruflicher Basiskompetenzen als arbeitsprozessorientierter beruflicher Handlungskompetenzen zu. Eine Korrelationsstudie gibt dennoch Hinweise zur Relevanz beruflicher „Basiskompetenzen" hinsichtlich beruflicher Performanzen (Nickolaus/Gschwendtner/Geissel/Abele 2009). So korreliert beispielsweise das Fachwissen am Ende des ersten Ausbildungsjahrs mit den Ergebnissen der IHK-Abschlussprüfung mit $r = .49$ und mit den betrieblichen Praxisleistungen mit $r = .44$.

Die Validitätsfrage ist bisher jedoch von keinem Arbeitskreis hinreichend angegangen worden[7].

Nun zur zweiten Betrachtungsebene: Wie ist es um das Verhältnis gängiger Messverfahren und computerbasierten Simulationen bestellt? Zu zeigen wäre vor allem, ob simulierte berufstypische Arbeitsproben/Arbeitsprozesse und reale Arbeits-

7 In einer kleiner angelegten, parallel zur hier referierten Studie durchgeführten Validitätsstudie konnte allerdings inzwischen ebenfalls gezeigt werden, dass die im Elektrobereich entwickelten und als Testelemente zur Erfassung der Problemlösefähigkeit eingesetzten Simulationen gute Abschätzungen der einschlägigen Fehleranalysefähigkeit ermöglichen (Wiesner 2009).

proben/Arbeitsprozesse in den erzielten Testleistungen vergleichbar sind[8]. Jurecka und Hartig (2007, S. 45 ff.) vermuten, dass es eine hohe Übertragbarkeit von Testergebnissen aus computerbasierten Tests auf reale Situationen gibt. Wenn dies der Fall wäre, würde man durch Computersimulationen ein betriebsspezifisches und damit unstandardisiertes Testen durch Beobachtungen im Arbeitsprozess ebenso vermeiden können, wie die sehr kostenaufwändige Bereitstellung realer Arbeitsproben. Jude und Wirth (2007, S. 56) fordern in diesem Zusammenhang, „zu analysieren, inwieweit beispielsweise simulierte Situationen mit den realen übereinstimmen, um zu recht vom Testverhalten auf die tatsächliche Kompetenz schließen zu können".

Mit unserer Untersuchung wollen wir prüfen, wie valide computerbasierte Simulationen realer Arbeitsproben sind. Etwas mikroskopischer wollen wir fragen, was bei einer inhaltstreuen Überführung beruflicher Realität in computersimulierten Umwelten genau passiert, d.h. wie verändern sich die Itemcharakteristiken und damit die Schätztreue von Kompetenzen? Dabei steht zunächst nicht die Reliabilität des Gesamttests im Fokus dieser Arbeit, sondern die Vergleichbarkeit der Bausteine (Items) eines Tests. Hinsichtlich des Gütekriteriums Reliabilität lässt sich jedoch festhalten, dass computerbasierte Verfahren, die ein computergestütztes, automatisiertes Beurteilungsverfahren mit einem IRT-basierten adaptiven Messzuschnitt unter Maximierung der Iteminformation verbinden, den klassischen Testzuschnitten in anderen Studien mindestens ebenbürtig zu sein scheinen (Stout 2002, S. 104).

Insgesamt versprechen wir uns mit den Befunden der Untersuchung[9] neben evidenzbasierten, steuerungsrelevanten Informationen zu adäquaten Testverfahren für die politischen Akteure eines möglichen Large-Scale-Assessments (VET-LSA) vor allem einen wissenschaftlichen Zugewinn.

2. Methode

2.1 Experimentelles Design, Stichprobe und Stichprobenziehung

Die Forschungsfrage klären wir an Hand der Simulation eines Kraftfahrzeugs. Wir entschieden uns auf Grund des dominierenden Stellenwerts von elektrischen und elektronischen Problemen in Kraftfahrzeugen für Fehler und deren Diagnose im Motormanagement und der Beleuchtungsanlage und entwickelten zusammen mit verschiedenen Praxisexperten acht unterschiedlich komplexe Fehlertypen. Diese Fehlertypen wurden in acht identische Autos implementiert und zusätzlich in Form einer Simulation realisiert. Damit entstanden 16 Fehlerfälle. Jeder Fehlertyp war also in beiden Settings (Realität und Simulation) als Zwillingsitem (ein Itempaket aus je einem realen und einem simulierten Fehlerfall) abgebildet. Die Bearbeitung der Fehlerfälle am realen Auto wurde in einer regulären Werkstatt mit allem verfügbaren Equipment einer Standardwerkstatt durchgeführt. Die Bearbeitung der Fehlerfälle

8 Untersuchungen zu Testmoduseffekten bezogen auf Korrelationen zwischen paper-pencil Tests und deren computerbasierten Derivaten liegen für die Übertragung klassischer paper-pencil Tests in das Computerformat vor. Hierbei waren keine großen Testmoduseffekte festzustellen, „wenn bei der Bearbeitung beider Formen dieselben Bedingungen bezüglich der Items und der Testvorgabe gegeben sind" (Jurecka/Hartig 2007, S. 43).

9 Die Ergebnisse entstammen einem Projekt, das vom BMBF finanziert wurde (Nickolaus/Gschwendtner/Abele 2009).

in der Simulation wurde in einem Computerraum an einer beruflichen Schule bzw. einer überbetrieblichen Ausbildungsstätte realisiert.

Um die reale Aufgabenbearbeitung mit der in der Simulation zu vergleichen, kommen verschiedene experimentelle Designs in Frage (von Davier/Carstensen/von Davier 2008)[10]. (1) Allen Probanden werden alle Aufgabenstellungen beider Settings (N = 16) dargeboten und anschließend die Abweichungen im Lösungsverhalten bzw. Lösungsergebnis jedes Itempakets (Fehlerfall 1 in Simulation & Fehlerfall 1 in Realität, …) miteinander verglichen. Das Problem ist in diesem Fall, dass es je Zwillingsitem zu einer Verzerrung in Abhängigkeit der Platzierung des Items kommt, die nicht mit Unterschieden in den Itemschwierigkeiten erklärbar ist, sondern als Resultat von Lerneffekten[11], motivationalen Verschlechterungseffekten etc. zu interpretieren wäre. Diese Variante schied damit von vornherein aus. (2) Die Probanden werden in 2 Gruppen unterteilt. Eine Gruppe löst alle Fehlerfälle in der Realität und eine andere Gruppe alle computerbasierten Fehlerfälle. Dazu müssen beide Gruppen bezogen auf die Fähigkeit, die der Lösung der Items zugrunde liegt, gleich verteilt sein. Das Problem ist, dass dieser Untersuchungszuschnitt nur mit sehr großen Stichproben funktioniert, die eine gleiche Verteilung der Fehleranalysefähigkeit erwarten lässt, was aus Ressourcengründen nicht in Frage kam. Ferner kann nicht geprüft werden, ob die latenten Fähigkeiten, die zur Lösung realer und simulierter Fehlerfälle die gleichen sind. (3) Sehr geschickt und vergleichsweise dem Ideal am nächsten lassen sich Auswertungen mit Ankertests und *multi-matrix design* (unter Rückgriff auf Verfahren der probabilistischen Testtheorie) durchführen. Diese Designs fungieren mit rotierten Itembündel. Durch eine geschickte Verschränkung der Items lassen sich die Daten mehrerer Gruppen gemeinsam skalieren. Zu berücksichtigen ist hierbei jedoch die reduzierte Stichprobengröße mancher gering verschränkter Items zur Schätzung der Itemparameter. Bei unserer gewählten Stichprobengröße würde aller Voraussicht nach ein Potential für größere Schätzfehler und damit für eine Nichtsignifikanz von Schwierigkeitsabweichungen provoziert werden. Auf Grund der enormen Planungsdichte der Erhebung und der Restriktionen auf Seiten der Organisationsstruktur der Durchführungspartner wählten wir (4) ein experimentelles Design, das eine randomisierte Zuteilung der Auszubildenden auf zwei Versuchsgruppen vorsieht (Zweigruppenversuchsplan mit Zufallsgruppenbildung mittels Zufallszahlentabelle bzw. Losentscheid) (siehe Tabelle 1), die jeweils vier Fehlerfälle in beiden Settings lösten.

10 Mit jedem Design sind spezifische Implikationen verknüpft. Die Implikationen sind primär mit den Designs konstruktiv verwobene Mess- bzw. Aussagefehler. Dabei müssen Mess- und Aussagefehler soweit als möglich in der Konstruktion des Designs reduziert werden. Mess- bzw. Aussagefehler können bspw. entstehen durch Vergleichsgruppen mit unterschiedlichen Fähigkeiten, durch Probanden, die ungeübt am Medium Computer und/oder computerängstlich sind oder durch die Itemplatzierung innerhalb des Testzyklus (Lerneffekte bzw. gegenläufig sinkende Anstrengungsbereitschaften/Motivationen bei Items nachfolgender Platzierungen). Ferner, wenn beim Vergleich von Aufgaben in der Realität und Simulation nicht folgende zwei Bedingungen erfüllt sind: (1) Jede Aufgabe, die in der Realität gelöst wurde, muss psychometrisch valide hinsichtlich der modellierten latenten Fähigkeit sein, Fehler in der Realität zu lösen. Gleiches gilt für Aufgaben, die in der Simulation gelöst wurden. (2) Jedes Item aus der Realität muss auch durch die latente Fähigkeit erklärt werden, die den Simulationsitems unterliegt, wenn sie gemeinsam skaliert werden. Wenn diese beiden Bedingungen nicht realisiert wären, würde man einen unzulässigen Vergleich zwischen unterschiedlichen Kompetenzdimensionen vornehmen.

11 Dies erscheint plausibel, da jeder Fehlertyp zweimal gelöst wird. In einer parallel durchgeführten Studie (Wiesner 2009) im Elektrobereich konnte das auch empirisch bestätigt werden.

Tab. 1: Erhebungsdesign für den Vergleich realer und simulierter Fehlerfalllösungen

	Gruppe 1 (N = 134)	**Gruppe 2 (N = 123)**
Fehlerfall 1	Bearbeitung des Fehlers am realen Kfz (R_1)	Bearbeitung des Fehlers in der Simulation (S_1)
Fehlerfall 3	**R_3**	**S_3**
Fehlerfall 5	**R_5**	**S_5**
Fehlerfall 7	**R_7**	**S_7**
Fehlerfall 2	**S_2**	**R_2**
Fehlerfall 4	**S_4**	**R_4**
Fehlerfall 6	**S_6**	**R_6**
Fehlerfall 8	**S_8**	**R_8**

Übergreifend	**Anwendungsorientierter Fachwissenstest ($N_{Gruppe1}$ = 121; $N_{Gruppe2}$ = 112)**
Übergreifend	**Intelligenztest CFT 20-R ($N_{Gruppe1}$ = 65; $N_{Gruppe2}$ = 78)**

Die Aufgaben selbst wurden den Probanden reihenfolgenrandomisiert zugewiesen (Randomisierung der Bedingung). Mittels eines zusätzlichen, den Fehlerfällen hochaffinen und anwendungsorientierten Fachwissenstests und des IQ-Tests CFT 20-R lassen sich die Gruppen auf Gleichheit in diesen gewöhnlich mit der abhängigen Variable Fehleranalyse- bzw. Problemlösefähigkeit hoch korrelierten Merkmalen (siehe Süß 2001; Gschwendtner 2008) gut beurteilen.

Die Stichprobengesamtgröße beträgt N = 294, die 202 Schüler aus dem dritten und 92 Schüler aus dem vierten Ausbildungsjahr umfasst. Die Auszubildenden des vierten Lehrjahrs sind allesamt Lehrlinge aus Handwerksbetrieben. Die Auszubildenden des dritten Lehrjahrs bestehen aus 63 Auszubildenden aus Berufskollegklassen, 78 Auszubildenden aus Handwerksklassen und 61 Auszubildenden aus Industrieklassen[12]. Vollständige Daten zu den Fehleranalysen und dem Fachwissenstest stehen von 257 bzw. 233 Probanden, IQ-Daten für 143 Fälle zur Verfügung. Aus Tabelle 1 wird die Verteilung auf die beiden Versuchsgruppen deutlich.

2.2 Testmaterialien und Prozedur

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 Fachwissensitems erfolgte in enger Anlehnung an die in der Praxis auftretenden Fehlerfälle und in enger Kooperation

12 Diese Konstellation eröffnet zusätzlich zu der Fragestellung der Studie ausbildungsspezifische Analysemöglichkeiten, die in weiteren Publikationen thematisiert werden.

mit Experten[13]. Alle Inhalte der Erhebungsinstrumente sind sowohl im dritten als auch vierten Ausbildungsjahr curricular abgesichert.

2.2.1 Authentische Fehlerfälle am realen Fahrzeug und in der Computersimulation

Ausgewählt bzw. entwickelt wurden acht komplexe Fehlertypen, die im Bereich des Motormanagements (sechs Aufgaben) und der Beleuchtungsanlage (zwei Aufgaben) angesiedelt sind. Entwickelt wurden Fehler am Injektor, am Ladedruckmagnetventil, an der Lichtanlage (2 Fehlerfälle), am Kraftstofftemperatursensor und an verschiedenen Sicherungen. Die Fehlerfälle wurden ausgehend von einem VW Golf V 1,9 TDI mit Pumpe-Düse-Technologie konstruiert. Das Lastenheft für die Konstruktionskriterien sah vor, dass

- der Fehlertyp realitätsgerecht ist;
- das Fehlersetting authentisch ist (z. B. werden die Fehlerfälle über Arbeitsaufträge eingeführt);
- die zentralen Arbeitsprozessschritte in der Fehlerdiagnose zu vollziehen sind [da in den Werkstätten die Auftragsannahme (und die Abnahme) durch den Meister erfolgt, beginnt die Diagnose erst mit dem Lesen und der Interpretation des Arbeitsauftrags] und
- das Schwierigkeitsspektrum möglichst umfassend abgedeckt ist: hierbei konnte auf eigene Vorarbeiten zu schwierigkeitsbestimmenden Merkmalen (Gschwendtner/Geissel/Nickolaus 2007; Gschwendtner 2008) und andererseits auf die Erfahrungen der Experten zurückgegriffen werden.

Die Pilotierung der Aufgaben, im Rahmen derer mit den Auszubildenden Interviews zur Bearbeitung durchgeführt und deren Herangehensweisen erfasst wurden, führte zu sukzessiven Optimierungen der Aufgaben. Alle Testelemente wurden zudem von Seiten der Experten als inhaltlich valide und im Anspruchsgrad als angemessen und variabel eingeschätzt. Die Testzeit betrug je Fehlerfall 30 Minuten. Damit ergab sich eine Gesamttestzeit von 4 Stunden je Proband[14]. 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 anzugeben, warum es nicht auch ein

13 Die Experten sind zwei Kfz-Ausbildungsmeister der Bildungsakademie Handwerkskammer Region Stuttgart in Weilimdorf und ein Kfz-Meister einer Hotline für Fehleranalysen im Kfz eines Stuttgarter Diagnosespezialisten. Weitere Experten vom Zentralverband des deutschen Kraftfahrzeuggewerbes (ZDK) und ein Fachleiter Kfz einer Stuttgarter Berufsschule validierten unsere Arbeit zusätzlich. Herzlichen Dank an dieser Stelle für die viele Arbeit, die sie investiert haben!

14 Trotz der ausgedehnten Testzeiten konnte eine hohe Durchführungsobjektivität realisiert werden durch die motivationale Kraft von Fehlerfällen, die eine starke Affinität zum beruflichen Alltag haben. Zusätzlich betonte ein Anschreiben der Innung des Kraftfahrzeuggewerbes Region Stuttgart und der Handwerkskammer Region Stuttgart an alle Mitgliedsbetriebe bzw. Auszubildenden der Region Stuttgart den hohen Stellenwert der Untersuchung im Sinne der Prüfungsvorbereitung für die anstehende praktische Gesellenprüfung. Ergänzend wurden attraktive Preise in Höhe von 1000 EURO für die „Besten“ ausgesetzt. Die Testzeit war so bemessen, dass die Fehlerfallbearbeitung für die meisten Auszubildenden kein Speedtest, sondern ein Powerttest war.

anderer Fehler sein könnte. Nicht erfasst wurde die Sprachproduktionskompetenz/Schreibfähigkeit zur Kontrolle der offenen Antwortformate. Die Sprachproduktionskompetenz/Schreibfähigkeit als potentiell konfundierendes Merkmal ist deshalb weitgehend zu vernachlässigen, weil unsere Auswertungen primär auf der Frage nach dem defekten Bauteil basieren. Die angegebene Messstrategie diente uns lediglich als messtechnischer Hinweis zur Validierung der Bauteilbenennung.

Der Bearbeitung der computersimulierten Fehlerfälle 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 jeder Auszubildende das Handwerkszeug der Simulationsbedienung beherrschte, bevor mit dem ersten Fehlerfall begonnen wurde. Insofern hielten wir es nicht für nötig, Computererfahrung oder Computerängstlichkeit zu kontrollierten.

Realisierung der Fehlerfälle in der Simulation

Alle Testmaterialien (Arbeitsaufträge, Bearbeitungsebenen, etc.) werden über den Bildschirm dargestellt, die Reaktion des Probanden erfolgt über die Anwahl der einzelnen Funktionalitäten via Maus. Das Ziel vollständiger inhaltlicher Abbildungstreue der Realität war in jedem Entwicklungsschritt leitend. Zur Vergegenwärtigung unserer Bemühungen soll im Folgenden mit einigen kommentierten Ausschnitten aus der Simulation angedeutet werden, wie die Fehlerfälle in der Computersimulation realisiert wurden und welche Möglichkeiten sie bieten. Mit der Simulation können bspw. neben Sichtprüfungen (Cockpitkontrollleuchten, Tankanzeige, Drehzahlmesser und Sicherungen [15]; Ansaugbereich, Kraftstoffleitungen, Motorluftfilter, Leuchtmittel) auch eine akustische Kontrolle des Motorlaufgeräusches sowie messtechnische Prüfungen elektrotechnischer Komponenten vorgenommen werden, wofür die erforderlichen Messgeräte wie Oszilloskop, Multimeter und Strommesszange (von links nach rechts in der oberen Tollleiste in Abbildung 1) zur Verfügung stehen.

Messungen können an insgesamt 12 Steckverbindungen, 3 Leuchtmitteln und am Buchsenkasten als Repräsentanten des Motorsteuergeräts (in Abbildung 1 zentral in der oberen Toolleiste) vorgenommen werden. Eine Adapterleitung (obere Toolleiste drittes Icon von rechts) dient der Überbrückung offener Steckverbindungen und damit der Messung von Signalen. Insgesamt wurden ca. 1500 Messwerte hinterlegt. Überall dort, wo mit den Messgeräten Messungen unternommen werden können, wurden Messstellen durch rote oder schwarze Kreise dargestellt.

Da an modernen, hochkomplexen Kraftfahrzeugen zur Fehlerdiagnose auch in den Werkstätten Expertensysteme herangezogen werden, implementierten wir ebenso ein Expertensystem. Die Entscheidung fiel dabei zugunsten der ESI[tronic] von BOSCH, die national und international weit verbreitet ist und deren Nutzung für Simulationszwecke von BOSCH gestattet wurde. Auch die Simulation der ESI[tronic] erfolgte in hohem Grade authentisch. Gleichwohl war es mit den verfügbaren Ressourcen nicht möglich, die Komplexität dieses Systems vollständig abzubilden. Die vorgenommene Begrenzung orientierte sich an den fehlerspezifischen Notwendigkeiten, wobei darauf geachtet wurde, dass bei allen Fehlern ein großes Spektrum an Diagnoseschritten eröffnet wurde, das auch zahlreiche Fehlwege einschloss. Mit einem Mausklick links unten auf den ESI[tronic] Button kommt man in diese zweite Programmebene (siehe Abbildung 2).

15 Bei Klick auf das Icon in der oberen Toolleiste rechts in Abbildung 1 erscheinen diese Features.

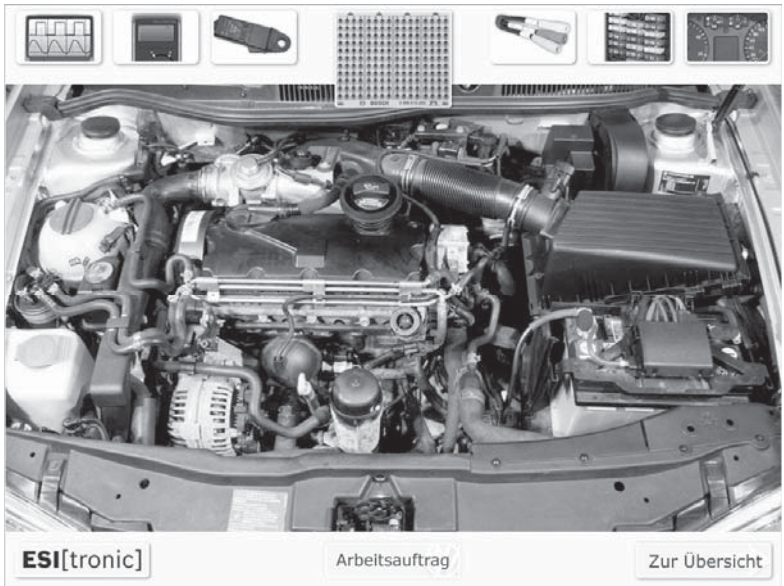

Abb. 1: Programmebene „Motorraumübersicht“

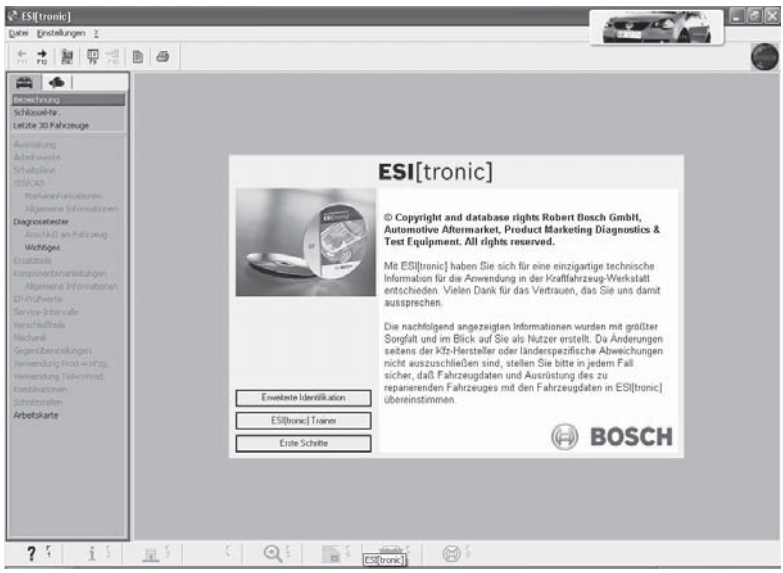

Abb. 2: Programmebene „ESI[tronic]“; mit Mausklick auf das Fahrzeug rechts oben kommt man wieder in die Programmebene „Motorraumübersicht“

Das Betreten der Programmebene wird für die Auszubildenden notwendig, sobald sie informationelle Unterstützung benötigen. Das Expertensystem stellt dem Auszubildenden auf unterschiedlichen Ebenen Informationen zur Verfügung, wie z.B.

die Einbaulage von Komponenten, Schaltpläne und insbesondere Informationen zu möglichen Ursachen des Fehlerfalles und Hinweise, wie ein Teil dieser möglichen Ursachen messtechnisch verifiziert bzw. falsifiziert werden kann. Zur Aktivierung des Systems ist es notwendig, das Fahrzeug mit den relevanten Informationen (Schlüsselnummer, Motortyp) aus dem Arbeitsauftrag genau zu spezifizieren.

Beispielhaft zeigen wir einen Auszug aus einem eher komplexen, schwierigen Fehlerfall. In diesem Arbeitsauftrag wurde angegeben: „Fahrzeug wurde vom ADAC angeliefert: ADAC Servicetechniker berichtet, dass der Wagen nicht mehr anspringt. Anlasser dreht aber noch durch“. Ließt der Auszubildende mit Hilfe der ESI[tronic] den Fehlerspeicher aus, bietet das Expertensystem verschiedene Messhinweise und dazu die Referenzdaten, wie das akzeptable Widerstandsspektrum des Innenwiderstandes oder zu erwartende Kennlinienverläufe an. Folgt man den Messhinweisen der ESI[tronic], sollte zuerst eine Widerstandsmessung am Drehzahlgeber erfolgen (siehe Abbildung 3).

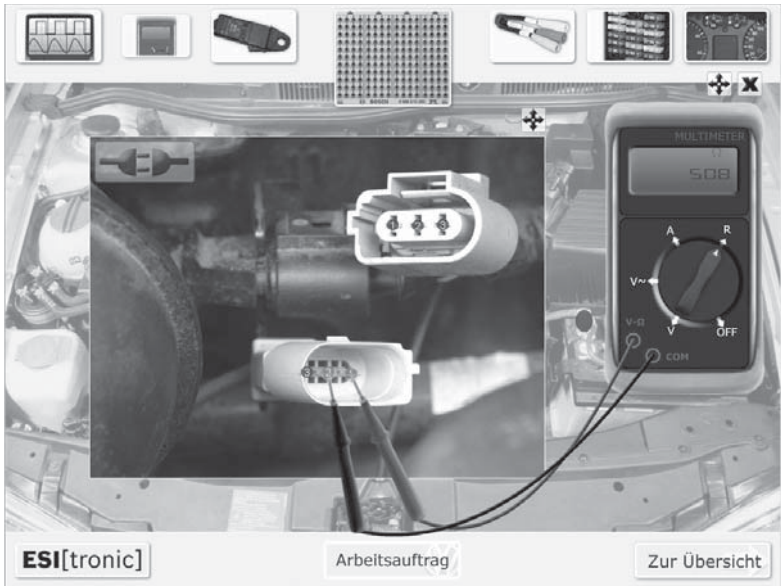

Abb. 3: Widerstandsmessung mit dem Multimeter an der Steckverbindung Bezugsmarkensensor in Richtung Drehzahlgeber

Die Messung von 508 Ω entspricht den Referenzdaten für einen intakten Drehzahlgeber. Somit muss die messtechnische Suche nach dem fehlerhaften Bauteil weitergehen. Auch die weiteren Überprüfungen am Drehzahlgeber (Kennlinienprüfung) zeigen keine Fehlfunktion an. Die Unterstützungsleistungen des Expertensystems enden in diesem Fall schließlich mit der Angabe weiterer möglicher Fehlerursachen, zu deren Überprüfung allerdings keine weitere Anleitung bereitgestellt wird.

Angegeben wurden bei diesem Fehler als weitere potentielle Ursachen:

- Leitung(en) mit Unterbrechung, Plus- oder Massenschluss
- Steckanschlüsse ohne oder mit schlecht leitender Verbindung

- Störende Einstreuungen infolge defekter oder nicht mit Masse verbundener Abschirmungen
- Impulsrad beschädigt, verschmutzt oder lose
- Drehzahlgeber trotz bestandener Prüfung defekt
- Steuergerät defekt

Spätestens an dieser Stelle müssen eigene Fehlersuchstrategien entwickelt werden, deren Umsetzung auch mit ökonomischen Implikationen verbunden sind. So wäre beispielsweise das Auswechseln des Steuergerätes sehr kostenträchtig. Die zwei zuerst genannten Fehlerursachen wären vermutlich weniger kostenträchtig, machen allerdings eine systematische Analyse notwendig, wobei auf die durch das Expertensystem bereitgestellten Schaltpläne zurückgegriffen werden kann.

Die Fehlercharakteristik variiert über die verschiedenen Fehlerfälle erheblich. In einem der Fälle bietet das Expertensystem beispielsweise keinerlei Anleitung für die Fehlersuche, so dass sofort eigene Fehlersuchstrategien entwickelt werden müssen. In einem sehr leichten Fall ist lediglich ein Leuchtmittel defekt, das einer Sichtprüfung zwar standhält, jedoch einen unendlich hohen Innenwiderstand besitzt.

Realisierung der Fehlerfälle in der Realität
Die Bearbeitung der Fehlerfälle am realen Auto wurde in einer regulären Werkstatt mit allem verfügbaren Equipment einer Standardwerkstatt durchgeführt. Insofern ist hierbei maximale inhaltliche Validität gesichert und somit eine gute Vergleichsbasis für die Validierung der Simulation gegeben.

2.2.2 Fachwissenstest und Intelligenzmessung (CFT 20-R)

Der Fachwissenstest (16 Items einer Mischung aus multiple-choice und halboffenen bis offenen Antwortformaten) ist analog zu den realen Arbeitsprozessen bei der Fehleranalyse in der Werkstatt konzipiert, d.h. handlungs- bzw. anwendungsorientiert. Die Items erfassen mehrere Facetten der Fehlersuche (Strategieschritte, Interpretationen von Messwerten, etc.) in den beiden Systemen Motormanagement und Beleuchtungsanlage. Für beide Fahrzeugsysteme wurde ein Testteil entwickelt. Beiden Testteilen 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. Ein Beispielitem, für dessen Bearbeitung ein Schaltplan zur Verfügung stand, soll den Grad der Handlungsorientierung verdeutlichen:

> „Der ausgelesene Fehlerspeicher eines Autos protokolliert: „Signal Fahrgeschwindigkeitssensor unplausibel.“ Führen Sie in der Tabelle unten alle Prüfschritte auf, die für eine eindeutige Fehleridentifikation nötig sind. Geben Sie dazu auch die Messstellen (Pins), das benötigte Messinstrument und den Messbereich an. Das Beispiel unten soll Ihnen das Vorgehen verdeutlichen.“

Prüfschritte	Messstellen (Pins)	Messinstrument	Messbereich
Beispiel: Prüfschritt 1	Zwischen Pin 85 an Bauteil K3 und Pin 6 am Steuergerät	Multimeter	Widerstand

Prüfschritte	Messstellen (Pins)	Messinstrument	Messbereich
1 . . . n			

Die Testzeit wurde auf 60 Minuten normiert. Die Erfahrungen mit den Pilotierungen und den Durchführungen der Hauptstudie zeigen, dass – wie für die Fehlerfälle ebenso gültig – der Test für die meisten Probanden günstigerweise ein Power- und kein Speedtest ist.

Um die beiden experimentellen Gruppen hinsichtlich ihrer Vergleichbarkeit zu beurteilen, wurde neben dem Fachwissenstest auch der nichtsprachliche IQ-Test CFT 20-R von Weiss (2006) verwendet.

2.3 Datenanalyse

Die Dokumentationsbögen zu den Fehlerdiagnosen und die Wissenstests wurden in doppelter und getrennter Korrektur kodiert, wobei es zu einer hohen Übereinstimmung (>.90) gekommen ist. In den wenigen Fällen diskrepanter Beurteilung wurde zusammen mit einem externen Experten eine Entscheidung getroffen. Die Kodierung erfolgte bei den Fehlerdiagnosen dichotom, bei dem Fachwissenstest sowohl dichotom als auch polytom.

Die Beantwortung der Forschungsfrage erfolgt in der Weise, dass wir in einem ersten Schritt die Legitimität prüfen, von einem der Lösung realer und simulierter Aufgaben gemeinsam zugrunde liegenden Fähigkeitsbündel auszugehen, das als eine psychometrische Dimension zum Ausdruck kommt. 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. Ergibt sich, dass die Eindimensionalitätsannahme zutrifft **und** dass die Gruppen hinsichtlich der Merkmale Fachwissen und IQ gleich verteilt sind (also keine systematischen Fähigkeitsverzerrungen zu erwarten sind), so ist es legitim, fehlende Schwierigkeitsdifferenzen und ähnliche Trennschärfen zwischen einzelnen realen und simulierten Fehlerfällen so zu interpretieren, dass es keinen Unterschied für die reliable Verortung einer Person auf einem Fähigkeitskontinuum macht, ob die Items *in der Simulation* **oder** *der Realität* dargeboten werden. Daraus ergeben sich drei ineinander verschränkt zu sehende Prüfungen: (1) Die Dimensionalitätsprüfung mittels latenten Korrelationen zwischen der Lösung realer und simulierter Aufgaben und der Prüfung der Itemfitwerte in unterschiedlichen Skalierungsrichtungen[16]. Aus der Analyse der Fit-

16 Zuerst wird je Gruppe eine Gesamtskalierung gerechnet, um das Verhalten aller Items zueinander zu beurteilen. Mit den Überlegungen zur Mehrdimensionalität von ICC-Verläufen von Wu (2004)

werte kann ersehen werden, ob einzelne Items nicht zur modellierten Fähigkeit der anderen Items passen. Diese Prüfverfahren können als gewichtige Indizien, wenn auch nicht als absolute Evidenzen (diese können immer nur angenähert werden), für eine Dimensionalitätsentscheidung angesehen werden. (2) Die Prüfung der randomisiert zusammengesetzten Untersuchungsgruppen (Gruppe 1 und Gruppe 2) auf gleiche Verteilungen in den Variablen Intelligenz und dem Fachwissen und (3) eine Differenzwertbeurteilung etwaiger Itemschwierigkeitsverzerrungen auf der Basis eines nonparametrischen Vierfelder-Chi-Quadrattests und der Beurteilung der Trennschärfen.

3. Ergebnisse[17] und Diskussion

Zu (1): 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 durch die Untersuchungsanlage die Korrelationen zwischen den Zwillingsitems nicht vorgenommen werden kann, werden zusätzlich die Fitwerte der je experimenteller Gruppe einzeln durchgeführten Skalierungen beurteilt. Hier sind sehr gute Itemfitwerte zu konstatieren (siehe Tabelle 2). Kein Item hat einen signifikant schlechten Fit (d.h. T-Wert von MNSQ > 2). Die MNSQ-Werte liegen durchgängig zwischen 0.78 und 1.25.

Bei einer wesentlich mikroskopischeren Prüfung (siehe Fußnote 16) ergeben sich keinerlei Modellverletzungen. Das heißt, dass sich insgesamt die Items und jedes einzelne Item aus der Realität gut zu jenen verhalten, die in der Simulation gelöst wurden und umgekehrt; dies gilt dazu noch in beiden Gruppen. Auch die mit Mplus (Version 5) durchgeführten konfirmatorischen Faktorenanalysen stützen die Eindimensionalitätsannahme. Zusammen genommen können diese Befunde als gewichtige Indizien für eine Dimensionalitätsentscheidung zu Gunsten eines eindimensionalen Fähigkeitsmodells gesehen werden, d.h. zur Lösung der realen und simulierten Aufgaben werden gleiche Fähigkeitsbündel benötigt.

wird anschließend folgendes Prozedere durchgeführt: Für Gruppe 1 wird zuerst ein Modell mit den vier Items aus der Simulation erstellt und dann geprüft, wie sich die Itemfitindizes verändern, wenn abwechselnd eines der vier Items aus der Realität mit modelliert wird. Dabei beobachten wir vor allem, wie sich die Fitindizes des Items aus der Realität zu jenen verhält, die das Grundmodell darstellen (Simulationsitems). Diese Prüfung wird ebenso für die vier Items aus der Realität vorgenommen, die dann sukzessive durch Items aus der Simulation angereichert werden. All diese Prüfungen wurden auch für Gruppe 2 vorgenommen.

17 Die Analysen wurden mit ConQuest 2.0 (Wu et al. 2007) durchgeführt.

Tab. 2: Itemwerte der Fehlerfälle gruppenweise skaliert (Gruppe 1 (G1), Gruppe 2 (G2)); Fehlerfälle G2: S7 und G1: S8 wurden zur Summennormierung constraint, womit keine Standardschätzfehler vorliegen

Item	Itemparameter	Schätzfehler	Weighted Fit (MNSQ, (T-Wert))	Punktbiseriale Korrelation
G1: R1	1.92	0.26	0.99 (T = 0.0)	0.62
G2: S1	1.62	0.25	0.99 (T = 0.0)	0.56
G2: R2	-1.56	0.22	1.18 (T = 1.6)	0.32
G1: S2	-2.05	0.24	1.16 (T = 1.2)	0.45
G1: R3	-1.50	0.22	0.97 (T = -0.3)	0.56
G2: S3	-2.44	0.26	1.06 (T = 0.4)	0.48
G2: R4	1.42	0.30	1.18 (T = 1.0)	0.42
G1: S4	0.01	0.26	1.11 (T = 0.9)	0.52
G1: R5	-2.77	0.27	1.01 (T = 0.1)	0.43
G2: S5	-2.94	0.30	0.84 (T = -0.7)	0.43
G2: R6	1.81	0.26	0.91 (T = -0.5)	0.53
G1: S6	1.34	0.29	1.25 (T = 1.4)	0.63
G1: R7	1.78	0.25	0.86 (T = -0.8)	0.67
G2: S7	1.13		0.78 (T = -2.0)	0.67
G2: R8	0.96	0.22	0.99 (T = 0.0)	0.54
G1: S8	1.28		0.85 (T = -1.1)	0.69

Zu (2): Für einen Schwierigkeitsvergleich muss auch die zweite Bedingung erfüllt sein, dass die experimentellen Gruppen gleich verteilt sind. Als Kriterien für eine gleiche Verteilung der Gruppen werden Fachwissen und IQ herangezogen. Diese Kriterien wählten wir, da sie gewöhnlich mit der Fehleranalysefähigkeit hoch korreliert sind (vgl. z.B. Süss 2001; Gschwendtner 2008)[18]. Die psychometrische Beurteilung des Fachwissenstests (skaliert mit N = 274) erwies sich auf Anhieb als gut. Kein Item hat einen signifikant schlechten Fit (d.h. T-Wert von MNSQ > 2). Die MNSQ-Werte liegen fast durchgängig (mit Ausnahme eines Items) zwischen 0.90 und 1.10. Die Reliabilität ist mit .67 (Cronbachs Alpha) noch 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, der z.B. über eine Erweiterung der Itembatterie (problemlos) realisierbar sein dürfte.

Im Wissenstest hat Gruppe 1 einen Summenscore von 14.82 (von 26 erreichbaren Punkten), SD = 4.27. Die Gruppe 2 hat eine annähernd gleiche Verteilungsstruktur wie Gruppe 1 (Summensore = 14.46, SD = 4.18). Für beide Gruppen können Normalverteilungen konstatiert werden (Kolmogorov-Smirnov-Test; p >.16) (siehe hierzu

18 Eine ausführliche Darstellung der konfirmatorischen Faktorenanalyse erfolgt in einer weiteren Publikation, die gegenwärtig in Bearbeitung ist.

auch Abbildung 4). Ein t-Test für unabhängige Stichproben ist der sehr ähnlichen Verteilungen entsprechend nicht signifikant (p = .19).

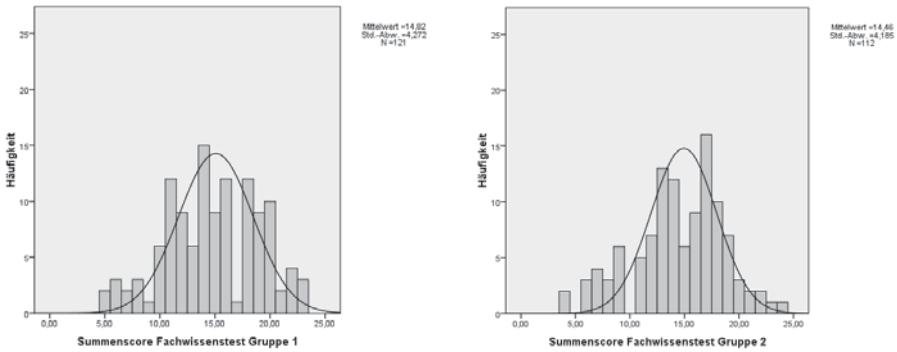

Abb. 4: Häufigkeitsverteilung des Fachwissenstest in den experimentellen Gruppen G1 und G2

Hinsichtlich des IQ sehen die Verteilungen ebenfalls sehr ähnlich aus[19]. In Gruppe 1 ergibt sich ein durchschnittlicher Intelligenzquotient von 107.88, Gruppe 2 erreicht einen annähernd gleichen Mittelwert von 105. Ähnliches gilt für die anderen Verteilungskennwerte. Für beide Gruppen können Normalverteilungen konstatiert werden (Kolmogorov-Smirnov-Test; p>.53) (siehe hierzu auch Abbildung 5). Ein t-Test für unabhängige Stichproben ist der sehr ähnlichen Verteilungen entsprechend ebenso wie beim Fachwissen nicht signifikant (p = .25).

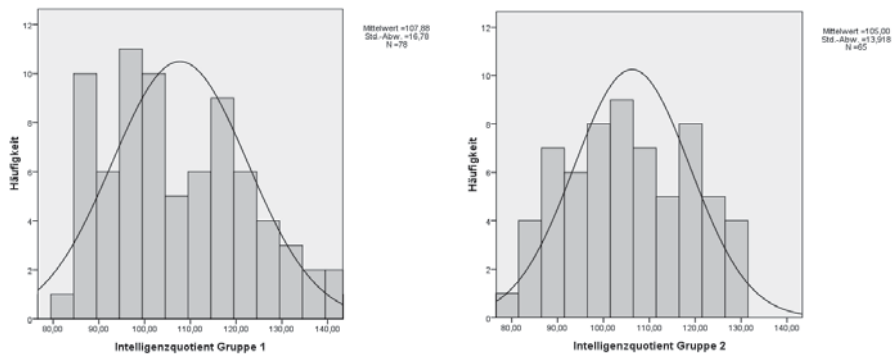

Abb. 5: Häufigkeitsverteilung des Intelligenzquotienten in den experimentellen Gruppen G1 und G2

19 Einschränkend ist hier zu erwähnen, dass wir zum IQ nur Daten von ca. 50 % der Probanden der beiden experimentellen Gruppen haben. Allerdings unterstellen wir, dass wir bei dieser zufällig zustande gekommenen Datenreduktion mit keinen wesentlichen Aussageeinschränkungen zu rechnen haben.

Somit können die beiden Versuchsgruppen hinsichtlich der Kriterien „Fachwissen“ und „IQ“ als „gleiche“ Gruppen aufgefasst werden und die Itemschwierigkeiten, gestützt auf die Indizien zur Eindimensionalität, direkt verglichen werden.

Zu (3): In Tabelle 3 ist je Fehlerfall eine Vier-Felder-Matrix mit den Richtig- und Falschlösungen differenziert je Setting aufgeführt. Zusätzlich befinden sich in der Tabelle asymptotische Signifikanzwerte der Chi-Quadrat Statistik zu den Unterschieden zwischen den Settings und eine Angabe zum relativen Risiko einer richtigen Lösung in der Realität gegenüber der Simulation, d.h. um welchen Faktor eine Richtiglösung in der Realität wahrscheinlicher ist als in der Simulation.

Tab 3: Vierfelder-Matrix der binär kodierten Itemantworten mit Chi-Quadrat-Statistik und dem relativen Risiko

Aufgabe		Setting		Asymptotische Signifikanz (2-seitig)	Relatives Risiko
		Real	Simulation		
Fehlerfall 1	Gelöst (in %)	16.0	17.2	.80	0.930
	Nicht gelöst (in %)	84.0	82.8		
Fehlerfall 2	Gelöst (in %)	73.6	77.3	.49	0.952
	Nicht gelöst (in %)	26.4	22.7		
Fehlerfall 3	Gelöst (in %)	68.9	85.3	.002**	0.808
	Nicht gelöst (in %)	31.1	14.7		
Fehlerfall 4	Gelöst (in %)	18.9	46.1	.001**	0.410
	Nicht gelöst (in %)	81.1	53.9		
Fehlerfall 5	Gelöst (in %)	85.4	90.1	.34	0.948
	Nicht gelöst (in %)	13.8	9.9		
Fehlerfall 6	Gelöst (in %)	15.0	25.0	.08	0.600
	Nicht gelöst (in %)	85.0	75.0		
Fehlerfall 7	Gelöst (in %)	17.4	23.8	.21	0.731
	Nicht gelöst (in %)	82.6	76.2		
Fehlerfall 8	Gelöst (in %)	26.7	23.5	.56	1.136
	Nicht gelöst (in %)	73.3	76.5		

Die Tabelle verdeutlicht verschiedene Befunde: Erstens kann festgehalten werden, dass die Aufgabenschwierigkeiten über das latente Schwierigkeitskontinuum der Fehleranalyseskala sehr gut verteilt sind. So konnten wir auf der Basis der Erkenntnisse aus den Vorstudien und der Kooperation mit den Experten den Schwierigkeitsgrad sehr variabel gestalten. Die Aufgabenschwierigkeiten bewegen sich zwischen ca. 16 % Lösungshäufigkeit im Fehlerfall 1 bis ca. 90 % im Fehlerfall 5. Einzig der Bereich zwischen ca. 40 % und 65 % ist nicht abgedeckt. Für die Grundlegung eines reliablen Instruments wäre an dieser Stelle ein Nachholbedarf vorhanden.

Weiterhin wird aus der Tabelle ersichtlich, dass die erzielten Übereinstimmungen der Diagnoseleistungen am realen und simulierten Kfz bemerkenswert hoch sind. In sechs von acht Fehlerfällen hängen die erzielbaren Lösungen **nicht** von den Settings

ab. Dafür sprechen die nichtsignifikanten Unterschiede der Lösungshäufigkeiten. Bei Fehlerfall 6 sind die Unterschiede kurz davor, die Signifikanzgrenze zu erreichen ($p = .08$). Die praktisch bedeutsamen Unterschiede zwischen den beiden Settings sind in der Regel nur minimal. So weichen die relativen Risiken meist gering vom Idealwert 1 ab. Beispielsweise unterscheiden sich die Lösungen zwischen den Settings im Fehlerfall 1 lediglich um den Faktor 0.93 (Messwert x 0.93; keine Abweichung bei Idealwert von 1). Bei zwei Fehlerfällen kommt es zu Abweichungen im Lösungsverhalten in Abhängigkeit vom Setting. Die erste Abweichung liegt bei Item 3 vor. Bei diesem Fehlerfall lösen lediglich 68.9 % der Probanden das Item in der Realität, hingegen 85.3 % in der Simulation. Dabei unterscheiden sich die beiden Settings um den Faktor 0.808. Gravierender stellen sich die Verhältnisse bei Fehlerfall 4 dar.

Der Tabelle kann auch entnommen werden, dass die Fehlerfälle annähernd durchgängig (bis auf Fehlerfall 8) in der Simulation leichter zu lösen sind. Dies ist sicherlich auf die höheren Komplexitäten der realen Anforderungssituationen zurückzuführen. Jedoch zeigen die meist geringen 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.

Wenn wir uns noch einmal Tabelle 2 in Erinnerung rufen, sehen wir auch, dass die Trennschärfen der Itemzwillinge relativ hoch sind und sehr nah beieinander liegen, wobei letzteres eine der Grundbedingungen für die Anwendungsfähigkeit des von uns verwendeten einparametrigen Raschmodells ist (vgl. Hambleton/Swaminathan/Rogers (1991).

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. Die Abweichungen bei Fehlerfall 4 können wir uns dadurch erklären, dass wir in Anbetracht von Kostenrestriktionen eine Innenraumkomponente im Motorraum (in der Simulation der Ort aller Diagnoseschritte) visualisierten, was (nicht nur auf der Darstellungsebene) zu einer (empirischen) Vereinfachung der Realität geführt hat. Die Abweichungen bei Item 3 erklären wir 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).

Die Analyse mit ambitionierteren statistischen Verfahren bzw. einem multimethodischen Ansatz haben wir an anderer Stelle angewandt, diskutiert und kritisch gewürdigt (Nickolaus/Gschwendtner/Abele 2009). Deshalb seien nur einige wenige Anmerkungen hierzu gemacht. Einen multimethodischen Zugriff hatten wir mit einem „blinden Fleck" des Untersuchungsdesigns begründet[20], der von drei zur Wahl stehenden Ansätzen

20 Der „blinde Fleck" ist selbst nicht in dieser Untersuchung geboren, sondern ein Problem von (berufspädagogischer) Empirie schlechthin: Fähigkeiten können generell als inhaltsgebunden angesehen werden und für diese spezifischen (fahrzeugtechnischen) Inhalte besitzen wir (noch) keine Diagnostik. Somit können wir nicht extern (mit einem anderen als dem hier zur Prüfung verwendeten Instrument) eine Verteilungsanalyse im Zielmerkmal „Fehleranalysefähigkeit" durchführen.

[Differenzwertbeurteilung von Lösungshäufigkeiten, Differential Item Functioning (DIF) und scale linking] unterschiedlich gesehen bzw. auszugleichen versucht wird: Wir wissen über die Fehleranalysefähigkeit der beiden Gruppen nur das, was wir mit dem letztlich in zwei Messinstrumente (je Gruppe eins) zerteilten Test erfasst haben. Wir wissen jedoch (noch) nichts über dessen Paralleltesteigenschaften (die es ja erst zu prüfen gilt). Die Häufigkeitsanalyse gründet in der Prämisse identischer Gruppen. Diese Prämisse hatten wir geprüft, indem wir die Häufigkeitsverteilungen der beiden experimentellen Gruppen in den – mit der Fehleranalysefähigkeit hoch korrelierten – Merkmalen „Fachwissen" und „IQ" untersucht haben. Trotz hoher Varianzaufklärungen verbleiben jedoch Restunsicherheiten. Die DIF-Analyse berücksichtigt für die Differenzwertbeurteilung der einzelnen Fehlerfälle in der Realität und in der Simulation die Differenz der mit den (nicht parallelisierten) Testteilen ermittelten Gruppendifferenz in der Fehleranalysefähigkeit. Das *scale linking*, bei dem die Fachwissensitems als Ankeritems zur Bildung einer gemeinsamen Skala benutzt werden, kontrolliert etwaige Gruppendifferenzen in dem Ankertestmerkmal (hier: Fachwissenstest) zur Skalierung der Fehlerfälle. Die Prämisse für eine Verankerung beider Fehlerfallskalen durch den Fachwissenstest ist, dass die Ankeritems zur Lösung die gleiche latente Fähigkeit voraussetzen wie die zu skalierenden Fehlerfälle (auch hier: Eindimensionalitätsannahme). Als Bedingung kann eine sehr hohe Korreliertheit der Testergebnisse aus den Fehlerfällen mit den Leistungen im Fachwissenstest und ferner eine günstigere Passung eines eindimensionalen Fähigkeitsmodells auf die Daten als die eines zweidimensionalen Modells (Fachwissen und Fehleranalyse) gelten. Das Fachwissen korreliert latent mit den Ergebnissen der Fehleranalyse von Gruppe 1 mit .76; in Gruppe 2 ergibt sich eine latente Korrelation von .80, was auf latenter Ebene zwar als relativ hoch, aber nicht zu hoch korreliert verstanden werden kann. Devianzstatistisch (Chi-Quadrat-Statistik) passt darüber hinaus das zweidimensionale Modell signifikant besser auf die Daten. Daraus schlussfolgern wir, dass mit dem Fachwissenstest 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[21]. Das Problem einer Verankerung mit diesen Items könnte sein, dass die zu skalierenden Fehlerfälle durch die Fachwissenstestitems fixiert und deshalb „inhaltsuntreu" verzerrt geschätzt würden. Die Korrelationen um .8 bestätigen ferner die Annahme, dass für die Erfassung von Fachkompetenz in einem VET-LSA beide Testformen in Kombination eingesetzt werden sollten.

Gleichgültig mit welchem Verfahren (DIF-Analyse, Vierfelder-Chi-Quadrat-Test) geprüft wird, die Befunde sind nahezu identisch. Allein das scale-linking weist den in der Chi-Quadrat-Statistik annähernd signifikanten Fehlerfall 6 als signifikant different aus.

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

4. Zusammenfassung und Ausblick

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 bedeutsamen Unterschiede. Bei einer weniger strengen, jedoch durchaus vertretbaren Einschätzung gilt diese Aussage für sechs von acht Fehlerfällen. 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 8) scheint die Simulationsvariante trotz großen Bemühens um Authentizität und damit Komplexitätsbezug etwas leichter als die Realität zu sein. **(3)** Die Daten zum relativen Risiko zeigen, dass die systematischen Verzerrungen zwischen Simulation und Realität in der Regel gering sind. Bei sorgfältiger Weiterentwicklung solcher Simulationen ist nach unseren Erkenntnissen mit der Minimierung solcher Verzerrungen zu rechnen. Weitere Entwicklungen sind ggf. durch zusätzliche Validierungsstudien abzusichern. **(4)** 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. **(5)** 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 Schwierigkeitsspektrum zu generieren. Eine verlässliche Niveaumodellierung setzt eine substantielle Erweiterung des Tests zur Fehleranalysefähigkeit voraus. **(6)** Bei Erweiterung unserer Software um Videosequenzen („Videovignetten“), weitere Audiodateien, Animationen und virtuellen Umgebungen können weitere Arbeitsprozesse wie Standardservice und weitere Diagnosetätigkeiten (z. B. in der Hydraulik) realisiert werden. Ferner können beispielsweise auch weitere Tätigkeitselemente wie Teilebestellungen integriert werden. Insgesamt wären damit zusätzliche Kompetenzbereiche miterfassbar. Weitere Forschung wäre zu der Frage angezeigt, inwiefern Reparaturen („Hands-on-Tätigkeiten“) simulierbar sind oder ob die Erfassung solcher erfolgskritischen Tätigkeiten wie Diagnose/Reparaturfähigkeit gleichzeitig via hoher korrelativer Verknüpfung, auch zur Abschätzung anderer Kompetenzfacetten geeignet sind.

Der hier beschrittene Weg scheint vor dem Hintergrund der erzielten Ergebnisse aussichtsreich, um zentrale Kompetenzaspekte mit computerbasierten Simulationen valide zu erfassen und andere „Gütekriterien“ (Reliabilität, Objektivität und im weiteren Sinne Praktikabilität und Ökonomie) gleichzeitig positiv mit zu beeinflussen. Inwieweit und in welcher Form weitere, gängige Testverfahren zur Erfassung eventuell computer-insensitiver Bereiche (wie z. B. motorisch artikulierte Tätigkeiten) ergänzend eingesetzt werden können, müssen weitere Forschungen, z. B. im Rahmen eines VET-LSA, zeigen.

Literatur

Baethge, M./Achtenhagen, F./Arends, L./Babic, E./Baethge-Kinsky, V./Weber, S. (2006): Berufsbildungs-PISA. Machbarkeitsstudie. Stuttgart: Franz Steiner

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.

Becker, M./Spöttl, G. (2008): Berufswissenschaftliche Forschung. Ein Arbeitsbuch für Studium und Praxis. Frankfurt a. M.: Peter Lang

Bremer, R. (2009): Erfassung beruflicher Kompetenzentwicklung und Identitätsbildung im Milieu großindustrieller Berufsausbildung (http://www.itb.uni-bremen.de/fileadmin/Download/kolloquien/2009/RB_ZBW_KOMPETENZENTWICKLUNGSVERLAUFSMESSUNG.pdf)

Geißel, B./Gschwendtner, T./Nickolaus, R. (2009): Betriebliche Ausbildungsqualität in der Wahrnehmung von Auszubildenden. In: Fürstenau, B./Tenberg, R./Wuttke, E. (Hrsg.): Tagungsband zur Herbsttagung der Sektion Berufs- und Wirtschaftspädagogik in Mannheim 2009. Opladen und Farmington Hills: Barbara Budrich (im Druck)

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@, H. 13 (http://www.bwpat.de/ausgabe13/gschwendtner_etal_bwpat13.pdf)

Haasler, B/Erdwien, B. (2009): Vorbereitung und Durchführung der Untersuchung. In: Rauner, F./Haasler, B./Heinemann, L./Grollmann, P. (2009a), S. 142–173

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

Hartig, J./Klieme, E. (Hrsg.) (2007): Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik. Eine Expertise im Auftrag des Bundesministeriums für Bildung und Forschung. BMBF

Jude, N./Wirth, J. (2007): Neue Chancen bei der technologiebasierten Erfassung von Kompetenzen. In: Hartig, J./Klieme, E. (Hrsg.), S. 49–56

Jurecka, A./Hartig, J. (2007): Computer- und netzbasiertes Assessment. In: Hartig, J./Klieme, E. (Hrsg.), S. 37–48

Luecht, R. M./Clauser, B. E. (2002): Test models for complex computer-based testing. In: Mills, C. N./Potenza, M. T./Fremer, J. J./Ward, W. C. (Hrsg.), S. 67–88

Mills, C. N./Potenza, M. T./Fremer, J. J./Ward, W. C. (Hrsg.) (2002): Computer-Based Testing: Building the Foundation for Future Assessments. Mahwah, NJ: Lawrence Erlbaum.

Nickolaus, R./Gschwendtner, T./Abele, S. (2009): Abschlussbericht für das Bundesministerium für Bildung und Forschung zum Projekt: Die Validität von Simulationsaufgaben am Beispiel der Diagnosekompetenz von Kfz-Mechatronikern. Vorstudie zur Validität von Simulationsaufgaben im Rahmen eines VET-LSA. Stuttgart

Nickolaus, R./Gschwendtner, T./Geißel, B./Abele, S. (2009): Konzeptionelle Vorstellungen zur Kompetenzerfassung und Kompetenzmodellierung im Rahmen eines VET-LSA bei Kfz-Mechatronikern und Elektronikern. In: AGBFN-Tagungsband (im Druck)

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.

Rauner, F./Haasler, B./Heinemann, L./Grollmann, P. (2009a): Messen beruflicher Kompetenzen. Band I: Grundlagen und Konzeption des KOMET-Projekts. Münster: LIT-Verlag

Rauner, F./Haasler, B./Heinemann, L./Piening, D./Maurer, A. (2009b): Berufliche Kompetenzen messen: Das Projekt KOMET der Bundesländer Bremen und Hessen. Zwischenbericht der wissenschaftlichen Begleitung. (http://www.ibb.uni-bremen.de/fileadmin/user/Kompetenzentwicklung/Zwischenbericht_KOMET_Final.pdf)

Stout, W. (2002): Test Models for Traditional and Complex CBTs. In: Mills, C. N./Potenza, M. T./Fremer, J. J./Ward, W. C. (Hrsg.), S. 103–112

Süß, H.-M. (2001). Prädiktive Validität der Intelligenz im schulischen und außerschulischen Bereich. In: Stern, E./Guthke, J. (Hrsg.): Perspektiven der Intelligenzforschung Lengerich: Pabst Science Publishers, S. 109–135

von Davier, A. A./Carstensen, C.H./von Davier, M. (2008): Linking Competencies in Horizontal, Vertical, and Longitudinal Settings and Measuring Growth. In: Hartig, J./Klieme, E./Leutner, D. (Hrsg.): Assessment of Competencies in Educational Contexts. Göttingen: Hogrefe.

Weiß, R. (2006): CFT 20-R. Grundintelligenztest Skala 2. Revision. Manual. Göttingen: Hogrefe.

Wiesner, K. (2009): Erstellung eines Simulationsprogramms zur Überprüfung von Leistungen in computersimulierten Umwelten unter besonderer Berücksichtigung der Einsetzbarkeit in Bildungseinrichtungen (Diplomarbeit Universität Stuttgart)

Wu, M. L. (2004): Item Response Modelling with ConQuest. Vortrag auf der Max-Planck-Summerschool 2004.

Wu, M. L./Adams, R./Wilson, M. R./Haldane, S. A. (2007): ACER ConQuest Version 2.0: Generalised Item Response Modelling Software. Melbourne: ACER Press.

Anschrift der Autoren: Dipl.-Gwl. Tobias Gschwendtner, Institut für Erziehungswissenschaft und Psychologie, Abteilung Berufs-, Wirtschafts- und Technikpädagogik, Geschwister-Scholl-Straße 24D, 70174 Stuttgart, Tel. 0711/6858-2998, Fax. 0711/6858-3130, gschwendtner@bwt.uni-stuttgart.de
Dipl.-Gwl. Stephan Abele, Institut für Erziehungswissenschaft und Psychologie, Abteilung Berufs-, Wirtschafts- und Technikpädagogik, Geschwister-Scholl-Straße 24D, 70174 Stuttgart, Tel. 0711/6858-2991, Fax. 0711/6858-3130, abele@bwt.uni-stuttgart.de
Prof. Dr. Reinhold Nickolaus, Institut für Erziehungswissenschaft und Psychologie, Abteilung Berufs-, Wirtschafts- und Technikpädagogik, Geschwister-Scholl-Straße 24D, 70174 Stuttgart, Tel. 0711/6858-3181, Fax. 0711/6858-3130, nickolaus@bwt.uni-stuttgart.de