

Universität Stuttgart
Institut für Erziehungswissenschaft
und Psychologie



Stuttgart, 4. Mai 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

Prof. Dr. Reinhold Nickolaus
Dipl.-Gwl. Tobias Gschwendtner
Dipl.-Gwl. Stephan Abele

Universität Stuttgart
Institut für Erziehungswissenschaft und Psychologie
Abteilung Berufs-, Wirtschafts- und Technikpädagogik (BWT)
Geschwister-Scholl-Str. 24 D
70174 Stuttgart
nickolaus@bwt.uni-stuttgart.de
Tel. 0711/685 83181

Inhaltsverzeichnis

I	KURZDARSTELLUNG	3
1.	AUFGABENSTELLUNG	3
1.1	Ansprüche an die Messverfahren	3
1.2	Prinzipiell mögliche Messverfahren und ihre Eignung für eine (internationale) Vergleichsstudie	3
2	DURCHFÜHRUNGSVORAUSSETZUNG DES VORHABENS	5
3	PLANUNG UND ABLAUF DES VORHABENS	6
3.1	Planung und Design der Validierungsstudie	6
3.2	Stichprobe und Durchführungsobjektivität	8
4	FORSCHUNGSSTAND UND DESSEN IMPLIKATIONEN FÜR DIE VALIDIERUNGSSTUDIE	9
II	ERGEBNISSE.....	14
1	TESTVERFAHREN	14
1.1	Authentische Fehlerfälle im realen Fahrzeug und in der Computersimulation	14
1.2	Wissenstest.....	22
1.3	Kodierung der Daten	23
1.4	Anforderungen der Aufgabenstellungen und notwendige Fähigkeiten zu deren Bewältigung	24
2	VALIDITÄTSPRÜFUNG.....	27
2.1	Prüfschritte zur Beantwortung der Forschungsfrage	27
2.2	Beantwortung der Forschungsfrage.....	27
2.3	Zusammenfassung der Befundlage	35
2.4	Weitere Befunde jenseits der gestellten Forschungsfrage	36
3	VERWERTBARKEIT DER ERGEBNISSE.....	36
3.1	Die Nutzung der Ergebnisse für diagnostische Zwecke	36
3.2	Die Nutzung der Ergebnisse für didaktische Bereiche	37
3.3	Während der Durchführung des Vorhabens bekannt gewordene Fortschritte auf dem Gebiet bei anderen Stellen	37
4	GEPLANTE VERÖFFENTLICHUNG	38
III	ANLAGE ZUM SCHLUSSBERICHT	39
1	BEITRAG DER STUDIE ZU DEN FÖRDERPOLITISCHEN ZIELEN.....	39
2	WISSENSCHAFTLICHE ERGEBNISSE.....	39
3	WEITERE VERWENDUNGSMÖGLICHKEITEN	39
4	WIRTSCHAFTLICHE ERFOLGSAUSSICHTEN.....	39
5	WISSENSCHAFTLICHE ERFOLGSAUSSICHTEN	39
6	ANSCHLUSSFÄHIGKEIT/WEITERENTWICKLUNG.....	40
7	ARBEITEN, DIE ZU KEINER LÖSUNG FÜHRTEN.....	40
8	KOSTEN UND ZEITPLANUNG	40
	LITERATUR.....	41

I Kurzdarstellung

1. Aufgabenstellung

Im Mittelpunkt des Projekts stand die Frage, ob die in simulierten beruflichen Anforderungssituationen erbrachten Leistungen eine zuverlässige Abschätzung fachspezifischer Performanz ermöglichen. Vorzunehmen war diese Validitätsprüfung für den Ausbildungsberuf Kfz-Mechatroniker, dessen Einbezug in ein VET-LSA vorgesehen ist. Die Fokussierung dieser Fragestellung erfolgte vor dem Hintergrund, dass für ein *large-scale assessment* die Konfrontation mit realen Aufgaben aus Praktikabilitätsgründen, die Einschätzung der Leistung am Arbeitsplatz durch die Vorgesetzten/Ausbilder wegen mangelnder Objektivität und Vergleichbarkeit ausscheiden und Paper-pencil Tests allein vermutlich nicht geeignet sind, die Performanz verlässlich abzuschätzen. Die Aufgabenstellung bzw. Zielsetzung der Studie ist vor dem Hintergrund wissenschaftlicher und bildungspolitischer Erwartungen an ein *large-scale assessment* zu sehen.

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) und valide (messen das, was erfasst werden soll) sind.

Im bildungspolitischen Raum besteht zumindest in jenen Ländern mit einem hohen Stellenwert dualer Ausbildungsvarianten der Anspruch, die Kompetenzmessung arbeitsprozessbezogen bzw. handlungsorientiert auszugestalten.

1.2 Prinzipiell mögliche Messverfahren und ihre Eignung für eine (internationale) Vergleichsstudie

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 (5) Paper-pencil Tests mit offenen oder geschlossenen Fragestellungen.

(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: eher unpraktikabel, verbunden mit erheblichen Finanzierungsproblemen, die Standardisierung im realen Arbeitsvollzug scheint unrealistisch, ohne Standardisierung sind keine verlässlichen Ergebnisse zu erwarten. Denkbar wäre eventuell die Integration kleinerer Elemente zur Abschätzung der Fertigkeiten.

(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, vor allem, wenn 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.

(3) Erfassung beruflicher Leistungsfähigkeit in realitätsnahen Simulationen beruflicher Anforderungssituationen

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 vorgenommen.

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

(4) Selbsteinschätzungen zur Ausprägung von Teilkompetenzen

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

Bewertung: ungeeignet

(5) Paper-pencil Tests

Paper-pencil Tests können sehr variantenreich gestaltet werden. Die Aufgaben können z. B. 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 in Verbindung mit den Simulationen geeignet.

2 Durchführungsvoraussetzung des Vorhabens

Die Durchführung der Studie setzte voraus, dass im Anschluss an die beruflichen Aufgaben Anforderungssituationen ausgewählt bzw. zugeschnitten wurden, die (1) dem Kern beruflicher Aufgaben entsprechen, zugleich (2) simulierbar und (3) in einer vertretbaren Dauer testbar sind sowie (4) möglichst die verschiedenen Kompetenzniveaus abdecken. Für die Generierung der Aufgaben war die enge Kooperation mit Experten unabdingbar. Kooperiert wurde u. a. mit dem Zentralverband des deutschen Kraftfahrzeuggewerbes, der Kfz-Innung Region Stuttgart, dem Bildungszentrum der Handwerkskammer Stuttgart, der Wilhelm-Maybach-Schule Bad Cannstatt, verschiedenen Experten zur Kfz-Fehlerdiagnostik, Bosch, Heinzmann-Training, verschiedenen Automobilunternehmen der Region und nicht zuletzt mit den Auszubildenden.

Benötigt wurden für die Untersuchungsdurchführung acht baugleiche Kraftfahrzeuge, in die Fehler implementiert wurden, acht Expertensysteme zur Fehlerdiagnose und vor allem die Bereitschaft der Kooperationspartner und der Auszubildenden zur andauernden Unterstützung des Vorhabens.

Trotz des sehr engen Zeitrahmens des Projektes (15.5.08 – 15.3.09), in dem die Kooperationsbeziehungen aufgebaut, die Kfz-Fachtests und Simulationen der Kraftfahrzeuge und Diagnosetools entwickelt, das Feld erschlossen, die Untersuchung durchgeführt und ausgewertet werden musste, konnte das Projekt zu einem guten Abschluss gebracht werden. Die ursprünglich angestrebte Samplegröße von 200 Probanden wurde deutlich überschritten. Die Datenerhebung erfolgte in der Zeit vom 10.11.2008 bis 04.02.2009. Die Haupterhebung zur Fehleranalysefähigkeit konzentrierte sich auf die Zeit vom 07.01.2009 bis zum 04.02.2009.

3 Planung und Ablauf des Vorhabens

3.1 Planung und Design der Validierungsstudie

Im Anschluss an relevante Vorarbeiten war vorgesehen,

- 1) die Tests und Simulationen in enger Kooperation mit den Experten zu entwickeln,
- 2) zu prüfen, ob zur Lösung der realen und simulierten Aufgaben gleiche Fähigkeitsbündel (eine Fähigkeitsdimension) benötigt werden und wenn ja, ob die Aufgaben gleich schwierig und damit vergleichbar sind und
- 3) damit zu klären, ob berufliche Fachkompetenz mit Simulationen beruflicher Anforderungssituationen verlässlich abgeschätzt werden kann.

Mit den Simulationen war vorgesehen, die realen Anforderungen möglichst authentisch nachzubilden. Ausgewählt wurden dazu aus dem Tätigkeitsspektrum der Kfz-Mechatroniker Fehlerdiagnosen am Kraftfahrzeug, die einen gewichtigen und den fachlich eher anspruchsvolleren Teil des Aufgabenspektrums abdecken. Die Entwicklung der Aufgaben und die Programmierung der Simulationsaufgaben erwiesen sich als hoch aufwändig, die Ergebnisse der Pilotierungen wurden sukzessive in den Optimierungsprozess eingespeist. Die ursprüngliche, sehr ambitionierte Zeitplanung (Fertigstellung der Simulationen bis 30.10.08) ließ sich nicht völlig einhalten. Erhebliche Probleme bereitete auch die Untersuchungsdurchführung, da die realen Kraftfahrzeuge nicht wie ursprünglich in Aussicht gestellt von Seiten des Zentralverbandes zur Verfügung gestellt werden konnten. Zugleich waren organisatorische Probleme zu bewältigen, da das Testdesign (parallele Tests an acht realen Kraftfahrzeugen und Tests an den Simulationen in PC-Räumen) in Verbindung mit den erheblichen Testzeiten (insgesamt 147h reine Erhebungszeit zur Erfassung der Fehleranalysefähigkeit) den Regelbetrieb der kooperierenden Einrichtungen erheblich einschränkte.

Durch diese Schwierigkeiten verschoben sich die Datenerhebungen bis in den Februar 2009, was auch zu Verzögerungen der Ergebnisvorlage führte.

Insgesamt ist festzuhalten, dass trotz der angedeuteten Probleme in hohem Grad authentische Simulationen entwickelt werden konnten, die inhaltliche Validität der Simulationsaufgaben in vollem Umfang gewährleistet ist, die parallel eingesetzten Fachwissenstests von den Experten als curricular und tätigkeitsbezogen ebenfalls als valide eingeschätzt werden, die Schwierigkeitsgrade der Aufgaben, wie angestrebt, das gesamte Spektrum des Leistungsniveaus abdecken, die Aufgaben selbst von den Experten und Auszubildenden als herausfordernd und motivierend

wahrgenommen werden und die Untersuchungsdurchführung in hoher Güte gewährleistet werden konnte.

Um die Untersuchungsfrage zu klären, kommen prinzipiell verschiedene Erhebungsdesigns in Frage, die mit je spezifischen Implikationen verknüpft sind (VON DAVIER/CARSTENSEN/VON DAVIER 2008): (1) Allen Probanden werden alle Aufgabenstellungen beider Settings ($N = 16$) dargeboten und anschließend die Abweichungen im Lösungsverhalten bzw. Lösungsergebnis jedes Itempakets (Arbeitsauftrag 1 in Simulation & Arbeitsauftrag 1 in Realität, ...) miteinander verglichen. Das Problem ist in diesem Fall, dass es je Itempaket zu einer Verzerrung in Abhängigkeit der Platzierung des Items kommt, die nicht mit Unterschieden in den Itemschwierigkeiten erklärbar ist sondern als Resultat von Lerneffekten (positive wie negative), motivationale Verschlechterungseffekte, 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. Dabei tauchen 2 Probleme auf: (a) Für solch ein Gruppenmatching existieren bisher keine geeigneten Messinstrumente und ein möglicher Ausweg aus diesem Dilemma, (b) die Untersuchung mit einer sehr große Stichprobe zu realisieren, die eine gleiche Verteilung des Zielmerkmals erwarten lässt, kam aus Ressourcen Gründen nicht in Frage. (3) Sehr geschickt lassen sich Auswertungen mit Ankertests und *multi-matrix design* (unter Rückgriff auf Verfahren der probabilistischen Testtheorie) durchführen. Diese Designs fungieren mit rotierten Itembündeln. Durch eine geschickte Verschränkung der Items lassen sich die Daten 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. (4) Somit wählten wir den Weg, Auszubildende randomisiert auf zwei Versuchsgruppen zuzuweisen, die jeweils vier Aufgaben in beiden Settings lösen. Durch den Einsatz von den Fehlerfällen hoch affinen, anwendungsorientierten Wissensaufgaben planten wir, die Problemfälle von Gruppe 1 und Gruppe 2 gemeinsam zu verankern und somit vergleichbar zu machen (siehe Abbildung 1).

	Gruppe 1 (N = 134)	Gruppe 2 (N = 123)
Fehlerfall 1	Am realen Kfz (R₁)	Bearbeitung des Fehlers in Simulation (S₁)
Fehlerfall 3	R₃	S₃
Fehlerfall 5	R₅	S₅
Fehlerfall 7	R₇	S₇
Fehlerfall 2	S₂	R₂
Fehlerfall 4	S₄	R₄
Fehlerfall 6	S₆	R₆
Fehlerfall 8	S₈	R₈
Übergreifend	Anwendungsorientierte Wissensaufgaben	

Abbildung 1: Erhebungsdesign für den Vergleich realer und simulierter Fehlerfalllösungen

3.2 Stichprobe und Durchführungsobjektivität

Die Stichprobengesamtgröße beträgt $N = 294$. Die Gesamtstichprobe untergliedert sich in $N = 202$ Schüler aus dem dritten und $N = 92$ Schüler aus dem vierten Ausbildungsjahr. Die Auszubildenden des vierten Lehrjahrs sind allesamt Lehrlinge aus Handwerksbetrieben. Die Auszubildenden des dritten Lehrjahrs 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 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“ ausgesetzt. 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 Forschungsstand und dessen Implikationen für die Validierungsstudie

Der Forschungsstand zur Kompetenzmodellierung und Kompetenzmessung konnte in den letzten Jahren substantielle Fortschritte verzeichnen, wengleich der großen Anzahl von hypothetischen Kompetenzmodellen (im Überblick z. B. BREUER 2006) immer noch eine relativ kleine Anzahl empirischer Überprüfungen gegenüber steht (LEHMANN/SEEBER 2007; GEIBEL 2008; GSCHWENDTNER 2008; NICKOLAUS 2008; NICKOLAUS/GSCHWENDTNER/GEIBEL 2008; SEEBER 2008; WINTHER 2008).

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 Aufgaben ermöglichen“ durchgängig herangezogen wird.

In den meisten Fällen, wie z. B. der Definition der KMK (2000), werden auch Bereitschaften einbezogen, diese Fähigkeiten einzusetzen, die allerdings starken, auch im Tagesverlauf schwankenden Ausprägungen unterworfen sein können.

Zur Kompetenzstruktur wird im Anschluss an die KMK-Definition meist angenommen, dass die Kompetenzdimensionen Fachkompetenz, Sozialkompetenz und personale Kompetenz zu unterscheiden sind und unterhalb dieser Ebene weitere Ausdifferenzierungen vorgenommen werden können.

Alternative Modelle finden sich beispielsweise bei OTT (2008), der zusätzlich eine eigene Methodenkompetenz postuliert, bei GREENO et al. (1984), die *conceptual*, *procedural* und *utilizational competence* unterscheiden oder auch im EQR.

Speziell zur Ausdifferenzierung der Fachkompetenz spricht die Befundlage dafür, dass, bei entsprechender Ausgestaltung der Testinstrumente, in der Regel zwei Dimensionen unterschieden werden können, eine Dimension, die eher Wissen und Verständnis und eine weitere, die eher Anwendungs- bzw. Problemlösefähigkeit abdeckt (ACHTENHAGEN/WINTHER 2009; SEEBER 2008; GEIBEL 2008; GSCHWENDTNER 2008; NICKOLAUS/GSCHWENDTNER/GEIBEL 2008). Zum Teil ergaben sich bisher allerdings auch eindimensionale Skalierungen. Vor diesem Hintergrund scheint hier auch von Interesse, ob sich durch die simulationsgestützte Neuentwicklung der Tests zur fachspezifischen Problemlösefähigkeit eine zweidimensionale Struktur ergibt. In beruflichen Anforderungsstrukturen, die (partiell) mathematische Fähigkeiten erfordern, wie im kaufmännischen Bereich im Rechnungswesen, kann davon ausgegangen werden, dass hier eine eigene Dimension zu einem stimmigeren Kompetenzmodell führt.

Für die Entwicklung der Aufgaben konnte neben den Erkenntnissen zur Dimensionalität auch auf Vorarbeiten zurückgegriffen werden, die Aussagen zu den schwierigkeitsbestimmenden Merkmalen bereitstellen. Über die verschiedenen Domänen hinweg erwiesen sich bisher folgende Merkmale als schwierigkeitsbestimmend:

- die Lernzielstufen nach Bloom
- die Komplexität bzw. der Vernetzungsgrad der relevanten Wissens Elemente
- die Anforderungen an eigene Modellierungsleistungen (z. B. Entwicklung einer Vorstellung wie etwas zusammenhängt oder funktioniert)
- die Vertrautheit der Anforderungen
- die Anzahl der Lösungsschritte und
- der mehr oder weniger übergreifende Charakter von Aufgaben (GEIBEL 2008; GSCHWENDTNER 2008; NICKOLAUS/GSCHWENDTNER/GEIBEL 2008; SEEBER 2008; WINTHER 2008).

Speziell im Kfz-Bereich erwiesen sich die Bloomsche Taxonomie und die Wissensvernetztheit als Prädiktoren der Aufgabenschwierigkeiten (GSCHWENDTNER 2008; NICKOLAUS/GSCHWENDTNER/GEIBEL 2008).

Andere, zunächst als relevant erachtete Schwierigkeitsparameter, wie die Hinweisgüte des Tabellenbuches¹ zur Bewältigung der Aufgaben oder der Vertrautheitsgrad der Aufgabenzuschnitte aus der Sekundarstufe I oder der Grundstufe, wurden bei Aufrechterhaltung des Signifikanzniveaus von 5% nicht in die Modellbildung aufgenommen (GSCHWENDTNER 2008; NICKOLAUS/GSCHWENDTNER/GEIBEL 2008). In der Domäne der Elektroniker wird zwar eine ähnlich hohe Varianzaufklärung des Schwierigkeitsgrades erzielt, als mächtigster Prädiktor erweist sich mit 39,6% allerdings die Hinweisgüte im Tabellenbuch; an zweiter Stelle wird auch in diesem Fall die Wissensvernetztheit in die Modellbildung integriert. Das starke Gewicht der Hinweisgüte des Tabellenbuchs überrascht einerseits wegen der fehlenden Integration dieses Parameters bei den Kfz-Mechanikern und andererseits wegen der beobachteten geringen faktischen Nutzung desselben bei der Aufgabenbearbeitung. Erklärbar ist das Phänomen gegebenenfalls durch den Tatbestand, dass einerseits deutlich ausgeprägte Kovarianzen zwischen der Hinweisgüte des Tabellenbuches und den anderen Schwierigkeitsparametern existieren und andererseits die Hinweisgüte auch Ausdruck für den Stellenwert der Inhalte sein dürfte. Bei den Überlegungen zu den Schwierigkeitsparametern der Aufgaben zur Problemlösefähigkeit ist einerseits zu beachten,

¹ Tabellenbücher finden in gewerblich-technischen Domänen sowohl in der Praxis als auch in Prüfungssituationen Verwendung.

dass Angehörige dieser Berufe im Sinne DÖRNERs (1982, 1987) in aller Regel mit Interpolationsproblemen bzw. in Abhängigkeit von der subjektiven Verfügbarkeit relevanten Wissens mit Synthesebarrieren konfrontiert sind² und andererseits im beruflichen Bereich die Bewältigung fachspezifischer Probleme an die Verfügbarkeit deklarativen und prozeduralen Fachwissens gebunden ist (MANDL/FRIEDRICH/HORN 1986; KLIEME u.a. 2001; NICKOLAUS/GSCHWENDTNER/GEIBEL 2008). Wesentliche Unterschiede alltagstypischer Problemstellungen in der Facharbeiterpraxis zu den in der bisherigen Problemlöseforschung dominierenden Problemzuschnitten bestehen darin, dass in der Regel geschlossene Systeme vorliegen, deren Funktionalität bekannt ist und z.B. durch Experten Analyseverfahren benannt werden können, die prinzipiell zum Erfolg führen. Damit stellt sich die Frage nach der Relation von Wissen und Kompetenz in spezifischer Weise. Für das eigene Untersuchungsanliegen werden primär analytische Problemtypen zugrunde gelegt, deren Bearbeitung iterative, messwertgestützte Soll-Ist-Vergleiche erfordert. Mit diesem Zuschnitt wird zugleich gewährleistet, dass die fachspezifischen beruflichen Anforderungen von Kfz-Mechatronikern angemessen fokussiert werden, die vor allem in den Bereichen der Wartung, Reparatur und Montage/Installation angesiedelt sind³ (BECKER 2005; HÄGELE 2002; SPÖTTL 2005). So kommen beispielsweise im Kfz-Bereich dem Standardservice mit ca. 40 % und Diagnose- und Reparaturarbeiten mit insgesamt etwa 50 % die größten Arbeitsanteile zu (SPÖTTL 2005, S. 71).

Als relativ anspruchsvolle Aufgaben erweisen sich vor allem Fehlerdiagnosen. Die Fehlerdiagnosefähigkeit setzt meist ein „Verstehen“ des technischen Systems voraus. In der Literatur finden sich z. T. pointierte Thesen, die zur Rolle des Wissens um die Funktion der Systemelemente und deren Zusammenspiel (innere Funktionalität) und die äußere Funktion des Gesamtsystems vertreten werden (BROWN/DE KLEERK 1981; GREENO/BERGER 1987; NEBER 2000). Dies ist vor allem vor dem Hintergrund verständlich, dass durch die zunehmende Integration von Systemen partiell lediglich fehlerhafte Systemkomponenten diagnostiziert und ausgetauscht werden müssen, ohne dass die innere Funktion der Systemelemente verstanden sein muss. In der Kfz-Technik ist der Einsatz von Diagnoseinstrumenten üblich, deren Einsatz in vielen Fällen eine

2 Vahling (1995) vertritt zwar die These, dass es sich bei der Störungssuche in technischen Systemen um dialektische Barrieren handle, da die Kenntnis über den Zielzustand gering sei, dem steht allerdings entgegen, dass die Zielzustände des funktionierenden Systems durchaus bekannt sind (s.u.).

3 Aufgabenzuschnitte wie sie im Kontext einer gestaltungsorientierten Didaktik präferiert werden, sind u. E. durch den Einbezug ökonomischer, ökologischer, arbeitsorganisatorischer und sozialer Belange nicht in Einklang mit psychometrischen Kriterien zu bringen. Ferner scheinen solche Aufgabenzuschnitte kaum mit den realen Anforderungen kompatibel zu sein.

eigenständige Durchdringung der technischen Funktionalitäten erübrigt, jedoch häufig auch zu keinen klaren Fehlerdiagnosen führt, so dass dennoch eine eigene systematische Fehleranalyse notwendig wird (RAUNER/SCHREIER/SPÖTTL 2002). Unseres Erachtens kann die Frage, ob ein detailliertes Wissen über die innere Funktionalität eines Systems für die Diagnose notwendig ist, nicht generell beantwortet werden, sondern ist von der Systemcharakteristik abhängig (GSCHWENDTNER/GEIBEL/NICKOLAUS 2007). Je nach Aufbau, Funktionsweise und Komplexität des elektrotechnischen/elektromechanischen Systems und der Fehlercharakteristik, stellen sich bei der Fehlerdiagnose je eigene Anforderungen. Je nach Fehlercharakteristik reicht bereits Erfahrungswissen hin, um aus einer äußeren Fehlfunktion des Systems auf die Fehlerursache schließen oder eine einschlägige Hypothese generieren zu können. In anderen Fällen sind systematische Fehlereingrenzungen notwendig, in welchen gegebenenfalls eigenständige Modellierungsleistungen zu erbringen sind und ein mehr oder weniger breites, flexibel einsetzbares Set übergreifender Strategien und Heuristiken zur Problemlösung notwendig wird. In eigenen Vorstudien ergaben Analysen von Problemstellungen und zugehörigen Itemschwierigkeiten, dass primär folgenden Schwierigkeitsindikatoren Bedeutung zukommt:

- 1) Der Systemkomplexität;
- 2) Der Notwendigkeit, eigenständig Modellbildungen zur Funktionsweise von Funktionselementen oder zum Gesamtsystem vornehmen zu müssen
- 3) Der Transparenz der Fehlersituation;
- 4) Der Anzahl der Messmöglichkeiten und den damit verbundenen Anforderungen, eine angemessene Messstrategie zu entwickeln;
- 5) Der Hinweisgüte der Fehlerbeschreibung, die mehr oder weniger hilfreich ist, Hypothesen zu den Fehlerursachen und zu potentiell erfolgreichen Prüfverfahren zu entwickeln.

Aus der Fähigkeitsperspektive erweisen sich u. E. folgende Kompetenzaspekte als bedeutsam:

- 1) Ein Wissen über das Gesamtsystem und die Funktionsweise von Systemkomponenten und deren Zusammenspiel;
- 2) Die Fähigkeit, selbständig Modelle zu Funktionsweise von Funktionselementen bzw. dem Gesamtsystem zu entwickeln oder entsprechendes Wissen zu aktualisieren;
- 3) Die Fähigkeit, aus externen Fehlfunktionen auf interne Zustände des Systems zu schließen;
- 4) Wissen zu Messverfahren;
- 5) Die Fähigkeit, eine adäquate Messstrategie zu entwickeln und schließlich die Fähigkeit, Messdaten angemessen zu interpretieren (GSCHWENDTNER/ GEIBEL/ NICKOLAUS 2007).

Funktionszusammenhänge können mehr oder weniger komplex und transparent sein. Besonders hohe Anforderungen liegen dann vor, wenn eigene Modellierungsleistungen zu erbringen sind. Das notwendige Wissen über Wirkungen/Effekte von Systemelementen und deren Topologie kann mehr oder weniger umfangreich und vernetzt sein. Eine regelgeleitete systematische messtechnische Eingrenzung setzt nicht nur messtechnische Kenntnisse und die Verfügbarkeit messtechnischer Eingrenzungsstrategien voraus, sondern ein systemspezifisches Wissen über funktionale Zusammenhänge⁴. Als relevante Prädiktoren der Fehleranalysefähigkeit erweisen sich das fachspezifische Wissen (als stärkster Prädiktor), die Motivation und partiell der IQ⁵.

Vor diesem Hintergrund scheint es zweckmäßig, neben Aufgabenzuschnitten, die eng an reale, problemhaltige Aufgaben angelehnt sind, auch fachspezifisches Wissen zu erfassen.

⁴ Auch VOLLMEYER/FUNKE (1999) vermuten im Anschluss an die Sichtung einschlägiger Befunde Abhängigkeiten zwischen Wissen über Zusammenhänge innerhalb des Systems und der Steuerleistung bei Problemlöseprozessen innerhalb des Systems.

⁵ Die z. T. im Reflex auf Lohhausen durchgeführten Studien zum Einfluss des IQ auf die Problemlösefähigkeit zeigen, dass der IQ in Abhängigkeit von der Fehlercharakteristik bzw. dem Anspruchsniveau prädiktive Kraft erhält (vgl. Diagnostica Bd. 37, H4, 1991 und H4 der Psychologische Rundschau 1999). Strohschneiders Ergebnisse deuten darauf hin, dass der Konkretisierungsgrad der Probleme wesentlichen Einfluss auf den Zusammenhang von Problemlöseleistung und Intelligenz hat (Strohschneider 1991).

II Ergebnisse

1 Testverfahren

Generiert wurden zwei Tests, (1) ein Fachwissenstest und (2) ein Test zur Erfassung der fachspezifischen Problemlösefähigkeit.

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, Werkstattmeister, Berufliche Schulen, Fachleiter, Ausbildungsmeister und Auszubildende). Alle Inhalte der Erhebungsinstrumente sind sowohl im dritten als auch vierten Ausbildungsjahr abgesichert. Die Inhalte beziehen sich auf zwei große Fahrzeugsysteme, die Beleuchtungsanlage und das Motormanagement.

1.1 Authentische Fehlerfälle im realen Fahrzeug und in der Computersimulation

a) Testkonstruktion

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. Realisiert wurden sowohl Simulationen des Kraftfahrzeuges selbst als auch eines Expertensystems. Die ursprünglich geplante Simulation von zwei unterschiedlichen Kraftfahrzeugtypen (Diesel, Benziner) wurde zugunsten einer umfangreicheren Simulation eines Typs (Diesel) aufgegeben, da sich im Entwicklungsprozess erhebliche Schwierigkeiten abzeichneten, für die Testphase genügend baugleiche Kraftfahrzeuge – ohne erhebliche Mehrkosten – bereitzustellen.

Entwickelt wurden Fehler am Injektor, am Ladedruckventil, an der Lichtanlage (2 Fehlerfälle), am Kraftstofftemperatursensor und an verschiedenen Sicherungen, wobei die Fehlerzuschnitte eine mehr oder weniger ausgeprägte Unterstützung durch das Expertensystem ermöglichen.

Die Pilotierung der Aufgaben, im Rahmen derer mit Auszubildenden Interviews zur Bearbeitung durchgeführt und deren Herangehensweisen 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 betrug je Fehlerfall 30 Minuten. Damit ergab sich eine Gesamttestzeit von 4 Stunden je Proband. Aus den Pilotierungen und Durchführungen in der Hauptstudie lässt sich folgern, 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 anzugeben, warum es nicht auch ein anderer Fehler sein könnte (schlussfolgerndes Denken, Messwertinterpretationen, Kontrollmessungen). 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.

Der **Test zur Erfassung der fachspezifischen Problemlösefähigkeit** umfasst acht Problemfälle, deren Generierung so erfolgte, dass

- 1) realitätsgerechte Fehler abgebildet,
- 2) die Fehler ohne dauerhaften Schaden am Kfz realisiert,
- 3) das Schwierigkeitsspektrum möglichst umfassend abgedeckt wurde,
- 4) die als relevant vermuteten Schwierigkeitsparameter eine hinreichende Varianz aufwiesen und
- 5) eine hohe Qualität bei der Simulation der Fehlerfälle gewährleistet werden konnte.

Wünschenswert ist zur Einlösung des vierten Kriteriums allerdings eine substantielle Erweiterung der Fehlerfälle. Generell ergeben sich in solch anwendungsbezogenen Anforderungssituationen Probleme, eine systematische Variation der Schwierigkeitsparameter zu realisieren.

Bei allen Testaufgaben ist gewährleistet, dass die zentralen Arbeitsprozessschritte in der Fehlerdiagnose zu vollziehen sind. Da in den Werkstätten die Auftragsannahme durch den Meister erfolgt, beginnt der Arbeitsauftrag auch im Test mit dem Lesen und der Interpretation des Arbeits-

auftrages, woran sich die Arbeitsplanung und die eigentliche Diagnose anschließen. Zu realisieren war auch eine Arbeitsdokumentation und eine Begründung für die erstellte Diagnose.

b) Realisierung der Fehlerfälle in der Simulation

Im Folgenden ist mit einigen kommentierten Ausschnitten aus der Simulation angedeutet, wie die Fehlerfälle in der Computersimulation realisiert wurden und welche Möglichkeiten sie bietet. Neben einfachen Sichtprüfungen sind auch Prüfungen elektrotechnischer Komponenten möglich, wofür die erforderlichen Messgeräte wie Multimeter, Strommesszange und Oszilloskop ebenfalls simuliert wurden. Hinterlegt wurden ca. 1500 Messwerte an zahlreichen Messpunkten.

Da an modernen, hochkomplexen Kraftfahrzeugen zur Fehlerdiagnose auch in den Werkstätten Expertensysteme herangezogen werden, ergab sich auch für die Simulation die Notwendigkeit, ein Expertensystem zu implementieren. 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 in Gänze 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. Abbildung 2 zeigt die beiden Programmebenen, das Kraftfahrzeug und das Expertensystem (ESI[tronic]), zwischen welchen jederzeit gewechselt werden kann.



Abbildung 2: Einstiegsfenster der Diagnose-Simulation mit den beiden Arbeitsbereichen ESI[tronic] und Fahrzeug

Abbildung 3 gibt einen Einblick in den Motorraum und die bereit gestellten Tools, wie Messgeräte (Oszilloskop, Multimeter, Strommesszange), Adapter, den Buchsenkasten (Repräsentant des Steuergeräts), den Sicherungskasten und einen Ausschnitt des Cockpits. Einzelne Bereiche, wie z.B. der Luftfilter, können auch einer Sichtprüfung unterzogen werden.



Abbildung 4: Signalmessung am Drehzahlgeber mit dem Oszilloskop

In gleicher Weise sind Messungen an vielen weiteren elektrotechnischen Komponenten einschließlich der Beleuchtungsanlage möglich (vgl. z.B. Abbildung 5, in dem eine Spannungsmessung zwischen Batterie und Buchsenkasten vorgenommen wird).

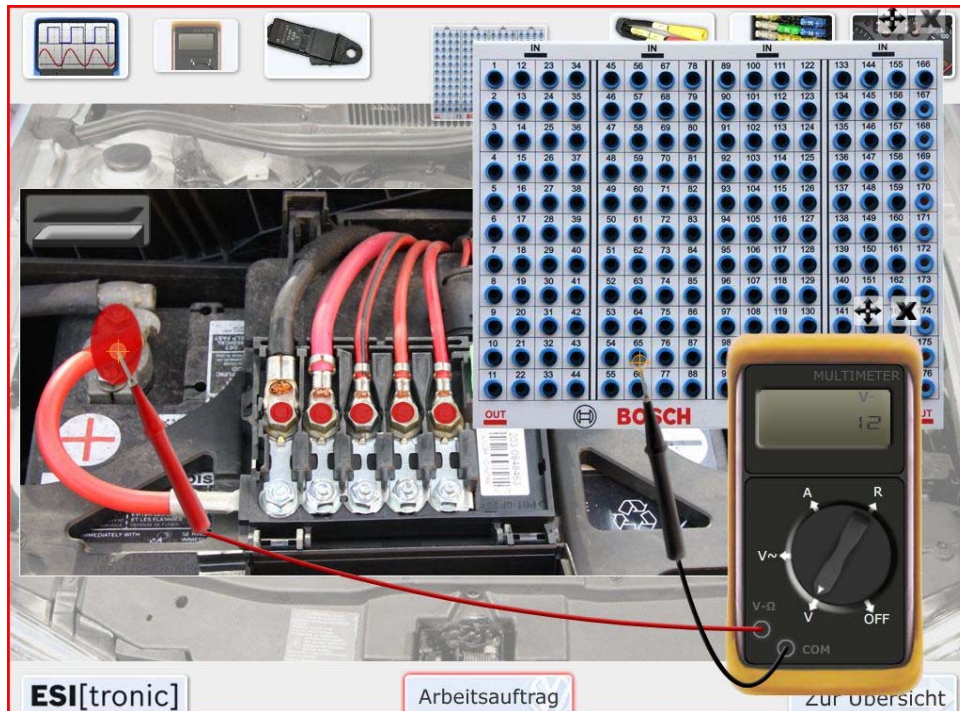


Abbildung 5: Spannungsmessung zwischen Buchsenkasten und Fahrzeugbatterie

Die zweite Simulationsebene, das heißt das simulierte Expertensystem wird für die Auszubildenden notwendig, sobald sie eine unterstützte oder geleitete Fehlerdiagnose durchführen wollen. Zur Aktivierung des Systems ist es notwendig, zunächst die relevanten Informationen aus dem Arbeitsauftrag zu entnehmen, wie Schlüsselnummer und Motortyp, mit welchen aus dem breiten Spektrum des Systems in der Praxis jene Variante aktiviert wird, die zur Analyse des jeweiligen Kraftfahrzeugs geeignet ist. Beispielhaft könnte in diesem Arbeitsauftrag angegeben sein: „Fahrzeug wurde vom ADAC angeliefert: ADAC Servicetechniker berichtet, dass der Wagen nicht mehr anspringt. Anlasser dreht aber noch durch“.

Mit dieser Information können bereits erste Hypothesen zur Ursache generiert werden. An dieser Stelle bietet sich der Rückgriff auf das Expertensystem zu einer gestützten Analyse und unter Aktivierung des eigenen Fachwissens an.

Das Expertensystem stellt dem Auszubildenden auf unterschiedlichen Ebenen Informationen zur Verfügung, wie z.B. Informationen zur Einbaulage von Komponenten, Schaltplänen und insbesondere Informationen zu möglichen Ursachen des Fehlerfalles und Hinweise, wie ein Teil dieser möglichen Ursachen messtechnisch verifiziert bzw. falsifiziert werden kann.

Das folgende Schaubild zeigt eine Widerstandsmessung am Drehzahlgeber, der als eine mögliche Ursache des Defekts aus dem Fehlerspeicher des Expertensystems ausgelesen wurde und zu dessen Funktionsprüfung Referenzdaten, wie das akzeptable Widerstandsspektrum des Innenwiderstandes, zu erwartende Kennlinienverläufe etc. abgreifbar sind.

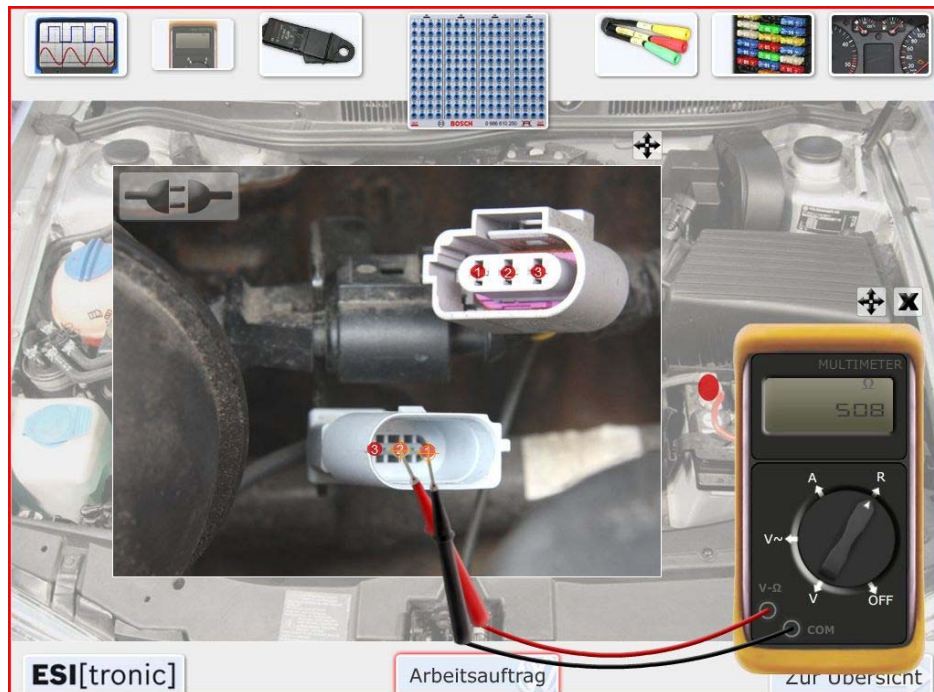


Abbildung 6: Innenwiderstandsmessung des Drehzahlgebers

Der hier angezeigte Widerstandswert liegt im akzeptablen Spektrum, 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 ist. So wäre beispielsweise die Auswechslung des Steuergerätes sehr kostenträchtig. Die zwei zuerst genannten Fehlersachen würden vermutlich weniger kostenträchtig, machen allerdings eine systematisch 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, da das Steuergerät des Kfz nicht mit Spannung versorgt ist, so dass sofort eigene Fehlersuchstrategien entwickelt werden müssen.

1.2 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 - der Test für die meisten Probanden günstigerweise ein Power- und kein Speedtest ist. Der Wissenstest ist analog zu den realen Arbeitsprozessen bei der Fehleranalyse in der Werkstatt konzipiert, d.h. die Items erfassen mehrere Facetten der Fehlersuche (Strategieschritte, Interpretationen von Messwerten, etc.) in diesen Systemen. Der Anspruch ist, den gesamten Test oder eine Itemauswahl als Ankeritems und damit als Vergleichsmedium zu benutzen (s. o.). Beispielhaft sind hier zwei Aufgaben angefügt, für deren Bearbeitung z. T. weitere Materialien (Schaltpläne) zur Verfügung standen.

Beispiel 1

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			

Beispiel 2

Im Rahmen einer TÜV-Vorbereitung überprüfen Sie die Beleuchtungsanlage eines Fahrzeugs und stellen lediglich ein schwach leuchtendes linkes Standlicht (E1) fest. Dies hat mit großer Wahrscheinlichkeit folgende Ursache:

- 1) Eine zu schwach leuchtende Schlussleuchte E2.
- 2) Ein Schalter S2 mit einem hohen Übergangswiderstand.
- 3) Ein defekter Glühdraht von E1.
- 4) Eine korrodierte Masseleitung 201.
- 5) Eine defekte Sicherung F5.

1.3 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 (s.o.) hoch. Damit sollten auch Teillösungen mit in das Modell aufgenommen werden, d.h. Lösungen, die im Ansatz zwar richtig, letztlich aber nicht zu einer präzisen Identifikation des Defektes geführt haben. Die Beurteilung der Angemessenheit solcher *partial-credit items* sollte anhand psychometrischer Kriterien erfolgen (s. u.). 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.

1.4 Anforderungen der Aufgabenstellungen und notwendige Fähigkeiten zu deren Bewältigung

Um das Verfahren der qualitativen Anforderungsanalyse der Aufgaben und der zu deren Bewältigung notwendigen Fähigkeiten transparent zu machen, wird im Folgenden für eine Aufgabe dieses Verfahren in den Grundzügen demonstriert. Als konkretes Beispiel wird ein Fehlerfall zum Motormanagement herangezogen. Die Analyse erfolgt bezogen auf die Simulationsvariante der Aufgabe in drei Schritten: (1) Ausdifferenzierung von Teilaufgaben, (2) Identifikation der damit verbundenen Anforderungen, (3) Zuordnung kognitiver Anforderungen, wobei weitere Ausdifferenzierungen auf allen Ebenen möglich sind. Dabei könnte beispielsweise in Feinanalysen auf das von MINNAMEIER (2005) vorgeschlagene Verfahren zurückgegriffen werden, das allerdings in hohem Grade aufwändig würde. Ausgangspunkt der Aufgabe ist der nachstehende Arbeitsauftrag:

„Fahrzeug wurde vom ADAC angeliefert. ADAC-Servicetechniker berichtet, dass der Wagen nicht mehr anspringt. Anlasser dreht aber noch durch. „Motorstörung Werkstatt“ wird angezeigt und Motorkontrolllampe leuchtet.“

Voraussetzung für die Bearbeitung der Aufgabe ist die Vertrautheit mit der Handhabung des Simulationstools, die durch eine gründliche Einführung (s. u.) gewährleistet wird.

Tabelle 1: Exemplarische Gegenüberstellung von tätigkeitsbezogenen und kognitiven Anforderungen

Teilaufgabe 1: Anforderungen	
<i>Arbeitsschritte:</i> Informationsaufnahme aus dem Arbeitsauftrag, relevante Informationen identifizieren und entnehmen	<i>Kognitive Anforderungen:</i> Lesen, Kenntnis relevanter Fachbegriffe; Wissen, welche Informationen im Weiteren benötigt werden bzw. relevant werden könnten
Teilaufgabe 2: Anforderungen	
<i>Arbeitsschritte:</i> Entscheiden, welche Schritte im Anschluss an den Arbeitsauftrag Ziel führend sein können -> Entscheidung, das Expertensystem zur Fehleranalyse heranzuziehen, da einfachere Fehlerursachen vermutlich bereits durch den ADAC-Service-Techniker überprüft wurden	<i>Kognitive Anforderungen:</i> Hypothesenbildung zu möglichen Ursachen im Anschluss an die im Arbeitsauftrag enthaltenen Informationen und/oder Wissen, wie mit dem Expertensystem Hinweise zu den Fehlerursachen zu erlangen sind

Teilaufgabe 3: Anforderungen	
<i>Arbeitsschritte:</i> Expertensystem nutzen	<i>Kognitive Anforderungen:</i> Wissen zu den Nutzungsvoraussetzungen (eingeschaltete Zündung, Anschluss); Wissen zu den Nutzungsmöglichkeiten des Expertensystems (Unterstützungsmodi, Fehlerauslesen, bereitgestellte Informationen wie Schaltpläne, Einbaulage von Komponenten etc.); Wissen, welche Informationen nötig sind, um das Expertensystem adäquat zu nutzen (Schlüsselnummer, Motortyp)
Teilaufgabe 4: Anforderungen	
<i>Arbeitsschritte:</i> Fehler auslesen, Entscheidung, ob gestützte Eigendiagnose oder geführte Fehlersuche	<i>Kognitive Anforderungen:</i> Deklaratives Wissen, welche Daten dazu eingegeben werden müssen und welche Systemeinheit relevant ist; Prozedurales Wissen, zur Aktivierung des Systems (Eingabe von Startdaten, Auswahl rel. Systemeinheit, Start der Fehlerdiagnose); Wissen, welche Unterstützung die beiden Diagnosevarianten bieten; Hypothese, welcher Weg schneller zum Erfolg führt
Teilaufgabe 5 (nach Entscheidung für Eigendiagnose): Anforderungen	
<i>Arbeitsschritte:</i> Interpretation der Fehlerhinweise und vorgeschlagene Prüfstrategie umsetzen	<i>Kognitive Anforderungen:</i> Kenntnisse der Fachbegriffe, Verständnis der Referenzdaten (z. B. Widerstandskennwerte, Kennlinienverläufe), wozu fachspezifische Lesekompetenz nötig ist; Hinweise auf Prüfungsmöglichkeiten umsetzen (Wissen zu Messverfahren, Einbaulage der Komponenten), Messdaten interpretieren unter Rückgriff auf Fachwissen und die Hinweise im Expertensystem

Teilaufgabe 6: Anforderungen	
<i>Arbeitsschritte:</i> Nach Identifikation erwartungskonformer Messdaten und globalen Hinweisen auf sechs weitere Fehlermöglichkeiten eigene Analysestrategie entwickeln	<i>Kognitive Anforderungen:</i> Kenntnisse der Fachbegriffe, die in den Hinweisen enthalten sind; Wissen, wie die weiteren Fehlermöglichkeiten getestet werden können; Entscheidung, welche Fehlermöglichkeiten zuerst überprüft werden sollen unter Berücksichtigung des Aufwandes und ökonomischer Implikationen
Teilaufgabe 7: Anforderungen	
<i>Arbeitsschritte:</i> Entwickelte Analysestrategie umsetzen	<i>Kognitive Anforderungen:</i> Wissen zu adäquaten Messverfahren, Einbaulage der Komponenten und funktionaler Abfolge der Analyseschritte; Fähigkeit, fehlendes Wissen (mit Hilfe des Expertensystems) zu erschließen; Schaltpläne lesen, darauf aufbauend Messschritte festlegen; Wissen, welche Messwerte erwartungskonform sind und wie erwartungswidrige Messwerte zu interpretieren sind
Teilaufgabe 8: Anforderungen	
<i>Arbeitsschritte:</i> Begründung der Fehlerzuschreibung	<i>Kognitive Anforderungen:</i> Wissen zur Interpretation der Messdaten; Fähigkeit, die Daten vor dem Hintergrund relevanten Fachwissens in einer schlüssigen Argumentation zu verknüpfen

Deutlich wird in dieser immer noch relativ groben Analyse des Fehlerfalles, dass berufliches Handeln (in diesem Tätigkeitsfeld) in hohem Grade wissensgebunden ist, und letztlich alle kognitiven Ebenen nach Bloom bzw. Anderson/Krathwohl (2001) umfasst. Vor allem bei Fehlschlägen des eigenen Vorgehens werden auch metakognitive Fähigkeiten für die Steuerung des eigenen Analyseverhaltens notwendig. Dies gilt prinzipiell für alle Diagnoseaufgaben, die in diesem Test realisiert wurden. Gleichwohl unterscheiden sie sich im Anspruch erheblich, da sowohl beim Einsatz von Messverfahren, der Interpretation von Daten, der Umsetzung und Entwicklung eigener Strategien, dem „Lesen“ von Schaltplänen etc. unterschiedliche Niveaus realisiert sind.

2 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 Verfahren an: (1) ein auf Basis klassischer Testtheorie angelegter Vergleich der simulierten ($S_1 \dots S_8$) und realen ($R_1 \dots R_8$) Aufgaben und (2) eine Skalierung der Aufgaben auf der Basis der Item-Response-Theorie.

2.1 Prüfschritte zur Beantwortung der Forschungsfrage

Die Forschungsfrage ist zu beantworten, indem wir in einem ersten Schritt die Legitimität prüfen, von einem der Lösung realer und simulierter Aufgaben gemeinsam zugrunde liegenden Fähigkeitsbündel auszugehen, das als eine 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 empirisch zutrifft **und** dass die Gruppen hinsichtlich des Zielmerkmals gleichverteilt sind (also keine systematischen Fähigkeitsverzerrungen aufweisen), 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. Daraus ergeben sich drei ineinander verschränkt zu sehende Prüfungen: (1) Die Dimensionalitätsprüfung mittels latenter Korrelationen zwischen der Lösung realer und simulierter Aufgaben und der Prüfung der Itemfitwerte in einzelnen Gruppenskalierungen. Die Ergebnisse können als starke Indizien für eine Dimensionalitätsentscheidung angesehen werden. (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*.

2.2 Beantwortung der Forschungsfrage

Zu (1): In der Gruppe 1 verzeichnen wir sehr hohe latente Korrelationen zwischen den Itempaketen aus der Realität (R_1, R_3, R_5, R_7) und Simulation (S_2, S_4, S_6, S_8) in Höhe von $r=.94$. Die gleiche Höhe (.94) erhalten wir in der Gruppe 2 zwischen den Itempaketen aus der Realität ($R_2,$

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 (siehe Abbildung 7). Kein Item hat einen signifikant schlechten Fit (d.h. T-Wert von $MNSQ > 2$). Die MNSQ-Werte liegen durchgängig zwischen 0.79 und 1.14. Einzige Ausnahme, und nur durch die klassische punktbiseriale Item-Skala-Korrelation (Trennschärfe) auffällig, ist die niedrige Trennschärfe (.18) des Items G2: R2 Das bedeutet, dass sich die Items aus der Realität gut zu jenen verhalten, die in der Simulation gelöst wurden. Und dies in beiden Gruppen. Zusammen genommen können diese Befunde als starke Indizien für eine Dimensionalitätsentscheidung gesehen werden, dass nämlich zur Lösung der realen und simulierten Aufgaben gleiche Fähigkeitsbündel benötigt werden.

Item	Itemparameter	Schätzfehler	Weighted Fit (MNSQ, (T-Wert))	Itemtrennschärfe (Punktbiseriale Korrelation)
G1: R1	1.255	0.160	1.02 (T = 0.2)	0.68
G2: S1	1.026	0.151	0.99 (T = 0.0)	0.64
G2: R2	-1.237	0.208	1.12 (T = 1.1)	0.18
G1: S2	-1.612	0.217	1.09 (T = 0.8)	0.40
G1: R3	-1.096	0.201	1.01 (T = 0.1)	0.47
G2: S3	-2.067	0.249	0.90 (T = -0.5)	0.41
G2: R4	1.536	0.285	1.07 (T = 0.4)	0.32
G1: S4	0.263	0.241	1.14 (T = 1.3)	0.43
G1: R5	-2.279	0.252	1.05 (T = 0.4)	0.36
G2: S5	-2.478	0.280	0.97 (T = -0.1)	0.38
G2: R6	1.888	0.247	1.00 (T = 0.0)	0.51
G1: S6	1.430	0.271	0.86 (T = -1.0)	0.65
G1: R7	1.231	0.157	0.79 (T = -1.4)	0.75
G2: S7	0.712		1.00 (T = 0.0)	0.72
G2: R8	0.621	0.139	1.11 (T = 0.9)	0.68
G1: S8	0.807		1.03 (T = 0.3)	0.76

Abbildung 7: Itemwerte der Fehlerfälle gruppenweise skaliert (Gruppe 1, Gruppe 2); (Fehlerfälle G2: S7 und G1: S8 wurden zur Summennormierung *constraint*, womit keine Standardschätzfehler vorliegen)

Zu (2): Unser weiter unten entfalteter multimethodischer Zugang scheint insofern angezeigt, als die Untersuchung einen „blinden Fleck“ aufweist⁶, der von den drei Ansätzen 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 ANOVA baut auf der gleichen Verteilung der beiden Versuchsgruppen auf, die mit dem der Fehleranalysefähigkeit hoch korrelierten Wissenstest (~.8) ermittelt wurde. 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* kontrolliert etwaige Gruppendifferenzen in dem Ankertestmerkmal (hier Wissenstest) zur Skalierung der Fehlerfälle.

Um etwaige Gruppendifferenzen in der abhängigen Variable (Fehleranalysefähigkeit) zu kontrollieren und damit die Fehleranalysen der einzelnen Fehlerfälle bezüglich Schwierigkeitsdifferenzen vergleichbar 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

⁶ Der „blinde Fleck“ ist selbst nicht in dieser Untersuchung geboren, sondern ein Desiderat 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.

nur bedingt geeignet scheint⁷. Das Problem einer Verankerung mit diesen Items kann sein, dass die zu skalierenden Fehlerfälle durch die Wissenstestitems fixiert und deshalb „inhaltsuntreu“ verzerrt geschätzt würden. 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.

Nun zur Prüfung der beiden Untersuchungsgruppen (Gruppe 1 und Gruppe 2) auf gleiche Verteilung. Wir strebten an, diese durch eine randomisierte Zuordnung der Probanden auf eine der beiden Versuchssettings und (bezogen auf die Problemlöseleistungen) durch die randomisiert eingespeiste Bearbeitung der Fehlerfälle je Proband (Vermeidung von systematischen Verzerrungen durch Platzierungseffekte) zu realisieren. Als Kriterium für eine gleiche Verteilung der Gruppen wählen wir die Fähigkeiten im Wissenstest (zur Auswahlentscheidung siehe oben). Dazu musste zunächst einmal der Wissenstest psychometrisch beurteilt werden: Die Wissenstestitems (skaliert mit $N = 274$) erwiesen sich auf Anhieb als psychometrisch gut (siehe Abbildung 8). Kein Item hat einen signifikant schlechten Fit (d.h. T-Wert von $MNSQ > 2$). Die MNSQ-Werte liegen fast durchgängig (Ausnahme Item 6) zwischen 0.90 und 1.10. Die Reliabilität ist mit .645 (Verhältnis von EAP zu PV) bzw. .67 (Cronbach's 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. Ein Blick auf die Verteilung der latenten Itemschwierigkeiten zeigt 3 größere Lücken in einem Bereich hoher Personenparameterdichte, die für den Reliabilitätswert verantwortlich gemacht werden können: Zwischen Item 5 (1.86) und Item 16 (1.04), zwischen Item 10 (0.78) und Item 13 (-0.02) und zwischen Item 6 (-0.07) und Item 14 (-0.52). Notwendig dafür ist eine substantielle Erweiterung des Instrumentariums, die im Rahmen der Aufgabenentwicklung für ein VET-LSA zu leisten ist.

⁷ Selbst bei vorheriger regressionsanalytischer (mit schrittweiser Integration) Ermittlung günstigster Itempakete (sechs Wissensitems) erhöhte sich die Korrelation zwischen Gruppe 1 und Wissenstest lediglich auf .81.

Item	Itemparameter	Schätzfehler	Weighted Fit (MNSQ, (T-Wert))	Itemschwierigkeit (Lösungshäufigkeit)	Itemrennschärfe (Punktbiseriale Korrelation)
1	-1.18	0.167	1.12 (T = 1.1)	82.78 %	0.35
2	-1.85	0.201	1.06 (T = 0.4)	89.78 %	0.34
3	-0.79	0.152	1.05 (T = 0.6)	77.37 %	0.39
4	-1.05	0.161	0.98 (T = -0.1)	81.02 %	0.47
5	1.86	0.147	0.95 (T = -0.7)	27.01 %	0.49
6	-0.07	0.137	0.88 (T = -2.1)	64.96 %	0.58
7	0.86	0.132	1.03 (T = 0.7)	45.99 %	0.48
8	-2.92	0.299	1.01 (T = 0.1)	95.99 %	0.25
9	2.71	0.177	1.07 (T = 0.6)	14.96 %	0.28
10	0.78	0.132	0.93 (T = -1.4)	47.81 %	0.56
11	-1.06	0.162	1.06 (T = 0.6)	81.18 %	0.31
12	-1.33	0.173	1.07 (T = 0.7)	84.56 %	0.35
13	-0.02	0.137	1.05 (T = 0.9)	63.97 %	0.46
14	-0.52	0.146	1.07 (T = 1.0)	73.16 %	0.40
15	3.53	0.229	0.96 (T = -0.2)	7.72 %	0.31
16	1.04		0.91 (T = -1.8)	42.28 %	0.52

Abbildung 8: Itemwerte des Wissenstests (Item 16 wurde zur Summennormierung constraint, womit keine Standardschätzfehler vorliegen)

Gruppe 1 hat einen Summenscore⁸ 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 Wissen als „gleiche“ Gruppen aufgefasst werden und somit die Itemschwierigkeiten gestützt auf die Indizien zur Eindimensionalität (s. o.) direkt verglichen werden. Die möglichen Prüfverfahren sind ANOVA, DIF-Analyse und *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 multimethodischen Verfahren zur Anwendung kommen.

Zu (3):

Auswertung mittels Lösungshäufigkeiten (bessere Anschaulichkeit) und ANOVA

Im folgenden Schaubild (Abbildung 9) sind aufgabenspezifisch die Anteile der Auszubildenden gegenübergestellt, die keine zutreffende Diagnose stellen konnten⁹. Die Verteilung der Diagno-

⁸ Die weiteren Analysen basieren auf dem Summenscore. Diese Darstellungsform wird an manchen Stellen auf Grund einfacherer Verständlichkeit und damit Kommunizierbarkeit notwendig, verändert jedoch nichts an den Aussagen.

seleistungen 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.

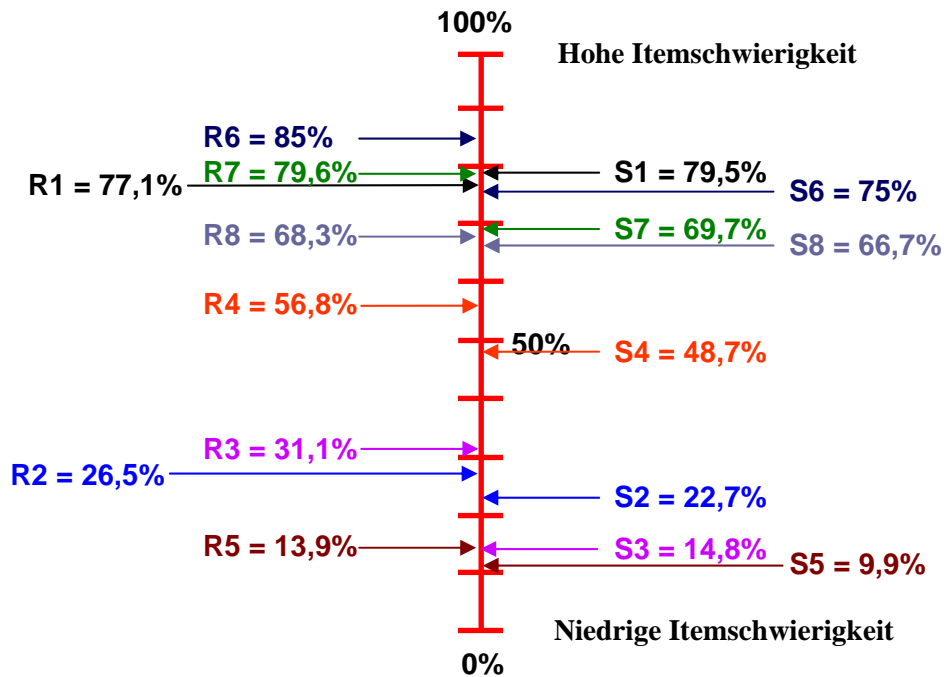


Abbildung 9: Lösungsquoten (falsche Lösung) aller 8 Fehlerfälle in beiden Zwillingversionen (links reale Fehlerfälle und rechts computersimulierte Fehlerfälle)

Inwiefern diese Unterschiede tatsächlich signifikant sind, zeigt die Auswertung mittels ANOVA (siehe Abbildung 10).

⁹ 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.

Item	Mittelwertsdifferenz	Signifikanz
Fehlerfall 1	Zwischen den Settings	.898
Fehlerfall 2	Zwischen den Settings	.494
Fehlerfall 3	Zwischen den Settings	.002**
Fehlerfall 4	Zwischen den Settings	.017*
Fehlerfall 5	Zwischen den Settings	.340
Fehlerfall 6	Zwischen den Settings	.082
Fehlerfall 7	Zwischen den Settings	.112
Fehlerfall 8	Zwischen den Settings	.890

Abbildung 10: ANOVA: Einfaktorielle Varianzanalyse auf dem Faktor (Setting: Realität, Simulation)

Somit sind die Unterschiede zwischen den Settings (Realität und Simulation) lediglich bei zwei Fehlerfällen signifikant: Bei Fehlerfall 3 (sign.= 0.002) und bei Fehlerfall 4 (sign.= 0.017). Bei dem Fehlerfall 6 sind die Unterschiede kurz davor, die Signifikanzgrenze von .05 zu erreichen (0.082). Insgesamt lässt sich mit der ANOVA zusammenfassen, dass es bei 6 (5) Fehlerfällen gelungen ist, innerhalb einer Fähigkeitsdimension von Realität und Simulation gleich schwierige Items zu erzeugen. Bei zwei Items ist dies nicht gelungen.

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, die einzelnen Items richtig zu 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, was einen Vorteil zum ersten Prüfverfahren darstellt. Bei signifikanten Schwierigkeitsverzerrungen (größer als der zweifache Standardschätzfehler) dient die Varianz der Stichprobe als Maß der Bedeutsamkeit der Verzerrung (ähnlich der Effektstärke). Das Problem an signifikanten und bedeutsamen Item*Setting Interaktionen ist, dass für diese Unterschiede dann nicht mehr nur die modellierte Fähigkeit alleine verantwortlich gemacht werden kann und somit das spezifiziertere Raschmodell ungültig ist. Praktisch gewendet, verzerren solche Items die Personenparameter in eine Richtung und führen somit zur systematischen Unter- bzw. Überschätzung der Stichprobenfähigkeiten.

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.

Immerhin sechs Items (Fehlerfälle 1, 2, 5, 6, 7, 8) weisen keine signifikante Schwierigkeitsverzerrung auf! Bei den Items 3 und 4 (siehe Abb. 5, G1: R3 und G1: S4) zeigen sich jedoch signifikante Schwierigkeitsverzerrungen, zu deren Ursachen wir begründete Vermutungen anstellen können (s.u.). Fehlerfall 6 hat einen (bei dieser Stichprobenszusammensetzung und vor allem Stichprobengröße!) nichtsignifikanten DIF.

Item	Logitdifferenz	Schätzfehler	Verzerrung
G1: R1	-0.001	0.105	nicht signifikant
G1: S2	-0.167	0.147	nicht signifikant
G1: R3	0.499	0.157	signifikant
G1: S4	-0.312	0.117	signifikant
G1: R5	0.149	0.189	nicht signifikant
G1: S6	-0.326	0.179	nicht signifikant
G1: R7	0.153	0.100	nicht signifikant
G1: S8	0.005		nicht signifikant

Abbildung 11: DIF-Analyse der Fehlerfälle (Fehlerfall S8 wurde zur Summennormierung constraint, womit kein Standardschätzfehler vorliegt)

Die wechselnden Vorzeichen in Abbildung 11 unter „Logitdifferenz“ zeigen, dass, wie sich das im obigen Kapitel anhand der Lösungsquoten bereits gezeigt hatte, nahezu durchgängig 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 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. 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 (nicht nur auf der Darstellungsebene) zu einer (empirischen) 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).

Bezogen auf die Stepparameter wissen wir aus einer hierarchischen Modelltestungsstrategie, dass ein polytomes Modell, das die Stepparameter berücksichtigt (*partial credit*) besser auf die Daten passt als ein polytomes Modell, das dies nicht tut. Das heißt, dass die Stepitems in den beiden Versuchsanordnungen unterschiedlich funktionieren, was letztlich ein Stichprobeneffekt sein kann, da die Trennung zwischen „sicher richtig“ und „sicher falsch“ nur wenige Probanden auf der Zwischenstufe verbleiben ließ, was vor allem die Messsicherheit stark beeinflusst. Dem ist in der Pilotierung des Tests für die Hauptstudie nachzugehen.

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

Die letzte Prüfmethode führt zu einer vergleichbaren Einschätzung wie die beiden Vorangegangenen. 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“ signifikante Differenz bei Fehlerfall 6, die sich im ersten Analyseverfahren gezeigt hat, jedoch signifikant.

2.3 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 bedeutsamen 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.

2.4 Weitere Befunde jenseits der gestellten Forschungsfrage

Die Unterschiede in den Testleistungen zwischen vierjährigen Handwerksklassen, vierjährigen Industrieklassen und dreijährigen Berufskollegklassen sind in den Bereichen Wissen und Fehleranalysefähigkeit hochsignifikant zu Gunsten des dreijährigen Berufskollegs ausgeprägt. An zweiter Stelle stehen die vierjährigen Industrieklassen und an dritter Stelle die vierjährigen Handwerksklassen. Die Unterschiede erreichen in etwa die Größenordnung einer Standardabweichung zwischen den dreijährigen Berufskollegklassen und den Handwerksklassen sowie ca. eine halbe Standardabweichung zwischen Berufskolleg und den Industrieklassen. Dieser Befund ist selbstverständlich konfundiert durch Selektionsmaßnahmen der Industrie: der Auswahl einer kognitiv leistungstärkeren Schülerklientel.

Dass jedoch die Handwerksklassen, die über die zurückliegenden Ausbildungsjahre in höherem Grad in die praktische Arbeit eingebunden waren und später genau diese Diagnosearbeit, mit der wir sie in der Fehleranalyse konfrontiert haben, sicher ausführen müssen, in dieser Facette auch an letzter Position stehen, zeigt Probleme in der Erreichung wünschenswerter Kompetenzausprägungen auf.

3 Verwertbarkeit der Ergebnisse

3.1 Die Nutzung der Ergebnisse für diagnostische Zwecke

Die Ergebnisse dokumentieren, dass im Bereich der Diagnoseaufgaben von Kfz-Mechatronikern Simulationen geeignet sind, die Performanz verlässlich abzuschätzen. Damit sind zumindest in diesem Bereich, der einen Kernbereich der Fachkompetenz von Kfz-Mechatronikern darstellt, die Voraussetzungen gegeben, auch in einem *large-scale assessment* die Fachkompetenz arbeitsprozessbezogen, objektiv, reliabel und valide zu erfassen.

Die Generalisierung der Befundlage auf alle Tätigkeitssegmente ist beim gegenwärtigen Forschungsstand nicht möglich. Eine parallel durchgeführte Studie zur Validität von Simulationen im elektrotechnischen Bereich (Fehlerdiagnosen) bestätigt allerdings das in dieser Studie gewonnene Ergebnis (WIESNER 2009). Zu Serviceaufgaben und bei Anforderungen im mechani-

schen Bereich lassen sich u. E. mit Hilfe von Videoaufnahmen relativ gut Anforderungssituationen simulieren, die sich für die Erfassung arbeitsprozessbezogener Fähigkeiten (z. B. Beurteilung des Arbeitsverfahrens, der Arbeitsausführung) eignen. Genutzt wird dieses Verfahren bisher beispielsweise zur Erfassung didaktischer Kompetenzen von Lehrkräften (OSER u. a. 2009). Inwieweit auf der Basis von solchen Bewertungen auf die Ausprägung der motorischen Fähigkeiten selbst geschlossen werden kann, wäre zu untersuchen.

Im Kontext von anschließenden Forschungsprojekten zur Kompetenzmodellierung und Entwicklung ist vorgesehen, das generierte Diagnoseinstrument weiter zu entwickeln. Das gilt einerseits für das realisierte Fehlerspektrum und andererseits für Optimierungen der Programmierung, die sich in spezifischen, selten auftretenden Anwendungssituationen noch als instabil erweist.

3.2 Die Nutzung der Ergebnisse für didaktische Bereiche

Das Lernen in realen komplexen Systemen kann aufgrund von negativen Implikationen, wie mangelnder Überschaubarkeit, kognitiver Überlastung der Lernenden, Sicherheitsproblemen und hohen Kosten bei Fehlhandlungen sowohl in pädagogischer als auch in ökonomischer Hinsicht erhebliche Probleme bereiten. Vor diesem Hintergrund wurden spätestens mit der Einführung von Juniorenfirmen beginnend „simulierte“, in ihrer Komplexität häufig auch reduzierte Lernumwelten geschaffen. Durch die technologischen Fortschritte sind inzwischen Simulationen in vielfältigen Feldern nutzbar. Die im Rahmen des Projektes entstandene Simulation ist nach den Einschätzungen von Auszubildenden und Lehrenden in hohem Grade geeignet, den Erwerb diagnostischer Kompetenz zu fördern. Bei den Kooperationspartnern besteht starkes Interesse daran, mit den Simulationen künftig didaktisch zu arbeiten. Die Freigabe dazu sollte u. E. jedoch erst dann erfolgen, wenn keine Kollisionen mehr mit der diagnostischen Verwertungsperspektive zu erwarten sind.

3.3 Während der Durchführung des Vorhabens bekannt gewordene Fortschritte auf dem Gebiet bei anderen Stellen

Parallel zur Durchführung dieser Validierungsstudie liegt am BWT der Universität Stuttgart eine weitere, jedoch im Umfang bescheidener angelegte Studie mit gleicher Zielrichtung in der elektrotechnischen Ausbildung vor. Gegenstand dieser Untersuchung war der Vergleich der Problemlöseleistungen von 52 Elektronikern für Energie- und Gebäudetechnik. Auch hier ging es um die Erfassung der Fehlerdiagnosefähigkeit. Als technisches System wurde eine Grundschaltung aus dem Bereich der Gebäudeinstallation gewählt (Wechselschaltung). Auch in dieser Untersuchung,

deren Ergebnisse noch nicht publiziert wurden, bestätigt sich die Validität der Simulationsaufgaben.

Im kaufmännischen Bereich wurde parallel zu der durchgeführten Studie von ACHTENHAGEN/WINTHER eine weitere Validierungsstudie durchgeführt, deren Ergebnisse dem BMBF vorliegen. Im Bereich der Lehrerbildung wird gegenwärtig häufiger auf Videovignetten zurückgegriffen, die allerdings einem anderen Simulationstypus zuzuordnen sind (vgl. z. B. OSER u.a. 2009). Darüber hinaus wurden uns keine weiteren einschlägigen Arbeiten bekannt.

4 Geplante Veröffentlichung

Vorgesehen ist die Veröffentlichung der Ergebnisse in der Zeitschrift für Berufs- und Wirtschaftspädagogik (voraussichtlich Heft 4, 2009) sowie in den internationalen Zeitschriften *Empirical Research in Vocational Education and Training* und *Vocations and Learning*. Vorteilhaft wäre u. E. die Übersetzung des Abschlussberichts ins Englische und dessen Streuung durch das BMBF.

III Anlage zum Schlussbericht

1 Beitrag der Studie zu den förderpolitischen Zielen

Förderpolitisches Ziel war die Klärung der Validität simulierter Anforderungskontexte am Beispiel von Kfz-Mechatronikern. Mit dem Ergebnis der Studie kann festgestellt werden, dass sich hinreichend authentisch gestaltete Simulationen tatsächlich dazu eignen, Performanz verlässlich abzuschätzen. Vor diesem Hintergrund steht u. E. ein auch für internationale Vergleichsstudien praktikables und zugleich valides Instrument zur Erfassung von Fachkompetenzen zur Verfügung, das allerdings erweiterungs- und anpassungsbedürftig ist.

2 Wissenschaftliche Ergebnisse

Das zentrale wissenschaftliche Ergebnis besteht in der Validitätsprüfung (vgl. Abschnitt II des Abschlussberichts). Nebenergebnisse sind einerseits der entwickelte Fachwissenstest, der in dieser Form bereits eine gute Annäherung an die tatsächlich erzielten Leistungen im realen Diagnoseprozess darstellt. Des Weiteren wurden wertvolle Erfahrungen gesammelt wie solch hochwertige Simulationen zu entwickeln sind.

3 Weitere Verwendungsmöglichkeiten

Die entstandenen Simulationen können in überarbeiteter Form künftig eventuell für didaktische Zwecke genutzt werden.

4 Wirtschaftliche Erfolgsaussichten

Wirtschaftliche Erfolgsaussichten wurden bisher nicht systematisch geprüft. Bei Verwertung im didaktischen Kontext sind ökonomische Verwertungsmöglichkeiten gegeben.

5 Wissenschaftliche Erfolgsaussichten

Mit dem entwickelten Test wurde Pionierarbeit geleistet, die vermutlich auf die weitere Testentwicklung in anderen Segmenten ausstrahlen wird. In einem laufenden DFG-Projekt und einem Kooperationsprojekt mit dem Bundesinstitut für Berufsbildung sind Anschlussarbeiten vorgesehen.

6 Anschlussfähigkeit/Weiterentwicklung

Von Seiten der PAL und des DIHK wurden Kontakte zum AN aufgenommen, um auszuloten, welche Entwicklungsmöglichkeiten sich aus der Innovation für die Prüfungsgestaltungen ergeben. Eine Voranfrage bei der DLR wegen der Weiterförderung erbrachte ein positives Signal. Angedacht ist u. a. eine Validitätsprüfung der gängigen Prüfungen.

7 Arbeiten, die zu keiner Lösung führten

Fehlanzeige

8 Kosten und Zeitplanung

Sowohl die Kosten als auch die Zeitplanung bedurfte partieller Anpassungen. Durchführbar war das Projekt nur durch die Einbringung von Eigenmitteln und dem großzügigen Entgegenkommen zahlreicher Kooperationspartner. Die Bereitstellung verzögerte sich aufgrund des hohen Arbeitsaufwandes den das Projekt notwendig machte um ca. 6 Wochen. Zur Kostenplanung siehe auch die Kostenabrechnung.

Literatur

Achtenhagen, F./Winther, E. (2009): Konstruktvalidität von Simulationsaufgaben: Computergestützte Messung berufsfachlicher Kompetenz – am Beispiel der Ausbildung von Industriekaufleuten. Bericht an das Bundesministerium für Bildung und Forschung.

Anderson, L. W./Krathwohl, D. R. (2001): A Taxonomy for Learning, Teaching and Assessing. New York u.a.: Longman.

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.

Bloom, B. S. (Hrsg.) (1973): Taxonomie von Lernzielen im kognitiven Bereich. 3. Aufl. Weinheim-Basel: Beltz.

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

Brown, J. S./de Kleerk, J. (1981): Towards a theory of qualitative reasoning about mechanisms and its role in troubleshooting. In: Rasmussen, J./Reese, W. B. (eds.): Human detection and diagnosis of system failures. New York, 317-335.

Dörner, D (1982): Lernen des Wissens- und Kompetenzerwerbs. In: Treiber, B./ Weinert, F.E. (Hrsg.): Lehr-Lernforschung. Ein Überblick in Einzeldarstellungen. München u.a.: Urban und Schwarzenberger, S. 134-148.

Dörner, D (1987): Problemlösen als Informationsverarbeitung. Stuttgart u. a., 3. Aufl.

Geißel, B. (2008): Ein Kompetenzmodell für die elektrotechnische Grundbildung: Kriteriumsorientierte Interpretation von Leistungsdaten. In: Nickolaus, R./ Schanz, H. (Hrsg.): Didaktik gewerblich-technischer Berufsbildung. Baltmannsweiler: Schneider, 121-141.

Greeno et al. (1984): Conceptual Competence and Children's Counting. *Cognitive Psychology*, 16, 94-143.

Greeno, J.G./ Berger, D. (1987): A model of functional knowledge and insight. Berkley.

Gschwendtner, T. (2008): Raschbasierte Modellierung berufsfachlicher Kompetenzen in der Grundbildung von KraftfahrzeugmechatronikerInnen. In: Breuer, K./ Deißinger, T./ Münk, D. (Hrsg.): Tagungsband zur Herbsttagung der Selektion für Berufs- und Wirtschaftspädagogik (vorläufiger Titel) (zur Veröffentlichung eingereicht).

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>)

Klieme, E. u. a. (2001): TIMSS – Impulse für Schule und Unterricht: Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente. Hrsg. Bundesministerium für Bildung und Forschung (BMBF).

Lehmann, R./ Seeber, S. (2007): Untersuchungen von Leistungen, Motivation und Einstellungen der Schülerinnen und Schüler in den Abschlussklassen der Berufsschlussklassen der Berufsschulen (ULME III). Behörde für Bildung und Sport Hamburg

Mandl, H./ Friedrich, H./ Horn (1986): Psychologie des Wissenserwerbs. In: Weidenmann, B./ Krapp, A. (Hrsg.): Pädagogische Psychologie. Ein Lehrbuch. München/Weinheim: Urban und Schwarzenberg, S. 146-218.

Minnameier, G. (2005): Wissen und inferentielles Denken. Peter Lang.

Neber, H. (2000): Nutzbares Wissen durch konditionalisierte und funktionalisierte technische Erklärungen. Rezeptives Lernen oder Entdecken durch Generieren? In: Zeitschrift für pädagogische Psychologie 14, 2/3, S. 124-136; Online:

<http://psycontent.metopress.com/content5h36260736k3541/fulltext.html> (06.07.2007).

Nickolaus, R. (2008): Vorstellung zur Modellierung beruflicher Handlungskompetenz und erste Versuche zur empirischen Prüfung. In: Nickolaus, R./ Schanz, H. (Hrsg.): Didaktik der gewerblich-technischen Berufsbildung. Baltmannsweiler: Schneider, 87-101.

Nickolaus, R. (2008): Modellierung zur beruflichen Fachkompetenz und ihre empirische Prüfung. In: Zeitschrift für Berufs- und Wirtschaftspädagogik (ZBW), 104. Bd., H.1, S. 1-6.

Nickolaus, R./ Gschwendtner, T./ Geißel, B. (2008): Entwicklung und Modellierung beruflicher Fachkompetenz in der gewerblich-technischen Grundbildung. In: Zeitschrift für Berufs- und Wirtschaftspädagogik (ZBW), 104. Bd., H.1, S. 48-73.

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.

Oser, F. u.a. (2006): Schlussbericht der ersten Phase – Leading House „Quality of Vocational Education and Training“.

Oser, F. u.a. (2009): Zwischenberichte – Leading House „Qualität der beruflichen Bildung“ – Professional Minds.

Ott, B. (2008): Eigenverantwortliches und arbeitsprozessorientiertes Lernen als technikdidaktische Kategorie. In: Nickolaus, R./Schanz, H. (Hrsg.): Didaktik der gewerblich-technischen Berufsausbildung – Diskussion Berufsbildung, Bd. 9, S. 10-25.

Rauner, F./ Schreier, N./ Spöttl, G. (Hrsg.) (2002): Die Zukunft computergestützter Kfz-Diagnose: rechnergeführte Handlangerarbeit oder qualifizierte Facharbeit? Bielefeld: Bertelsmann.

Seeber, S. (2008): Ansätze zur Modellierung beruflicher Fachkompetenz in kaufmännischen Ausbildungsberufen. In: Zeitschrift für Berufs- und Wirtschaftspädagogik (ZBW), 104. Bd., H.1, S. 74-97.

Spöttl, G. (2005): Fahrzeug-Kommunikationstechnik – forciert sie die Generalisierung und/ oder Spezialisierung des Kfz-Mechatronikers? In: Pangalos, J./ Spöttl, G./ Knutzen, S./ Howe, F. (Hrsg.): Informatisierung von Arbeit, Technik und Bildung. Münster: LIT, S. 71.

Strohschneider, S. (1991): Problemlösen und Intelligenz: Über Effekte der Konkretisierung komplexer Probleme. In: Diagnostica - Zeitschrift für Psychologische Diagnostik und differentielle Psychologie. Band 37, 1991. S. 353–371.

Vahling, L. (1995): Determinanten von Problemlösefähigkeit für die Störungssuche. In: Die Berufsbildende Schule. 47.Jg., H.2: Teil 1: S. 99-103; H.3: Teil 2: S. 140-145.

Vollmeyer, R./Funke, J. (1999): Personen- und Aufgabenmerkmale beim komplexen Problemlösen. In: Themenheft Komplexes Problemlösen. Psychologische Rundschau, Heft 4 1999, S. 213–219.

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.

Wiesner, K. (2009): Simulationen zur Erfassung der Fehleranalysefähigkeit in der elektrotechnischen Ausbildung. Universität Stuttgart (Diplomarbeit, in Vorbereitung).

Winther, E. (2008). Vocational Competence – Constructing Measures in the Field of Business and Administration. Paper presented at the Berkeley Evaluation and Assessment Research Center, University of California, Berkeley. November, 17th, 2008.