Sandra Bley / Michaela Wiethe-Körprich / Susanne Weber

Formen kognitiver Belastung bei der Bewältigung technologiebasierter authentischer Testaufgaben – eine Validierungsstudie zur Abbildung von beruflicher Kompetenz[1]

Kurzfassung: Kompetenzen valide abzubilden bedeutet im Sinne Messicks (1989; 1995), integrative empirische Evidenzen für die Angemessenheit der Leistungsmessung zu erbringen. Eine Quelle fehlerhafter Fähigkeitseinschätzungen entsteht durch Aktivierung unbeabsichtigter kognitiver Kapazitäten (Belastungen). Unter Berücksichtigung der Cognitive Load Theory (Sweller 2004; van Merriënboer/Kirschner 2013) wurden zur Messung von Intrapreneurship-Kompetenz 18 authentische Testaufgaben technologiebasiert konstruiert. Dabei wurde versucht, unerwünschte (extraneous cognitive load: = ECL) Belastungen möglichst vollständig in der Testplattform zu reduzieren und konstrukt-relevante (intrinsic cognitive load: = ICL) Belastungen in den Testaufgaben – durch Kombination theoriebasierter ICL-Kriterien – systematisch in ihrer Höhe zu variieren. Die vorliegende Studie prüft mittels einer kombinierten Analyse aus verbalen Protokollen (Studien Lauten Denkens) und tatsächlich erbrachten Leistungen von 26 Auszubildenden, inwiefern ICL und ECL in der intendierten Weise beim Lösen der authentischen technologiebasierten Testaufgaben (zum Intrapreneurship) vorkommen. Die Validierung erfolgt im Mixed Method Design: Über varianzanalytische und grafische Tests werden aufgabenbasiert Abweichungen zu intendierten ICL und ECL identifiziert; mittels qualitativer Analysen werden Erklärungsmuster für die Abweichungen herausgearbeitet. Die Studie zeigt bei einer Vielzahl der Aufgaben erwartungskonforme Ergebnisse. Identifizierte Abweichungen und zugehörige Erklärungsmuster für ICL und ECL liefern konkrete Überarbeitungsvorschläge für das vorliegende Testinstrument und darüber hinaus wegweisende Handlungsempfehlungen für die Konstruktion technologiebasierter authentischer Testaufgaben.
Stichworte: Kompetenzmessung, berufliche Kompetenzen, Validität, Cognitive Load Theory, Lautes Denken, Mixed Method

Abstract: According to Messick (1989; 1995) a valid assessment of competencies is given if integrative empirical evidence for adequate proficiency estimations exists. A source of an incorrect estimation of skills emerges by activating unintended cognitive capacities (loads). 18 authentic test tasks were constructed technology-based in order to assess intrapreneurship competencies on the basis of the Cognitive Load Theory (Sweller 2004; van Merriënboer/Kirschner 2013). In this design process, we tried to minimize undesired (extraneous cognitive load: = ECL) loads in the platform and to systematically vary construct-relevant (intrinsic cognitive load: = ICL) loads in the test tasks by linking theory-based ICL-criteria. By combining the analysis of verbal protocols and the measured performance of 26 apprentices, this study examines to what extent ICL and ECL occur as intended while solving the test tasks. The validation follows a Mixed Method Design: Deviations from intended ICL and ECL are identified task-based by conducting variance-analytical and graphical tests; qualitative analyses enable explanation patterns for the deviations. With respect to a large number of tasks the study shows results which are conform to the expectations. Few deviations and the corresponding explanation patterns for ICL and ECL deliver precise proposals for revisions on the existing test instrument. Furthermore, they provide pioneering recommendations for the construction of technology-based authentic test tasks.
Key words: assessment of competencies, competencies in vocational education and training, validity, Cognitive Load Theory, Think Aloud, Mixed Method

1 Das Projekt wird im Rahmen der ASCOT-Initiative (www.ascot-vet.net) durch das Bundesministerium für Bildung und Forschung gefördert (AZ: 01DB1118).

1. Einführung

„Validity is […] the most fundamental consideration in developing and evaluating tests“ (American Educational Research Association/American Psychological Association/National Council on Measurement in Education 2014, S. 9). Qualitätsbewertungen bisheriger Assessment-Entwicklungen fokussierten jedoch eher auf Reliabilitäts-, denn auf Validitäts-Aspekte der Testinstrumente (Pellegrino/DiBello/Brophy 2014). Der Einfluss und die damit verbundenen Konsequenzen von (high-stake) Vergleichsmessungen erfordern es, verschiedene Aspekte von Validität (u. a. inhaltliche/ kognitive/ strukturelle/ prädiktive) im Sinne Messicks (1989; 1995) zur Bewertung der Testqualität zu bestimmen und integrativ im Sinne empirischer Evidenzen für die Zuverlässigkeit der Leistungsmessung zu interpretieren (Pellegrino/DiBello/Brophy 2014; Leuders 2014). Mit anderen Worten: Wir müssen Klarheit darüber besitzen, WAS wir messen.

Authentische Aufgaben zur Messung beruflicher Kompetenzen

Lehr-lerntheoretische Diskussionen in der Berufs- und Wirtschaftspädagogik betonen, dass die Entwicklung und valide Messung beruflicher Kompetenzen nur dann gelingen kann, wenn „authentische“ berufstypische und handlungsbasierte Lern- und Testaufgaben eingesetzt werden, die eine Abbildung der Befähigung zu selbstständigen und selbstverantworteten Handlungen erlauben (Baethge u. a. 2006; Achtenhagen/Winther 2009). Damit benötigen papier- und bleistiftbasierte Lern- und Testaufgaben im schulischen Alltag und insbesondere in beruflichen Abschlussprüfungen (AKA-Aufgaben) zur Inszenierung der Realität umfangreiche Beschreibungen. So umfassten beispielsweise die Beschreibungen zu Geschäftsprozess-Situationen, die Grundlage für die Aufgabenbewältigung waren, in den AKA-Prüfungsaufgaben 2008 für Industriekaufleute insgesamt 20 Druckseiten (Achtenhagen/Winther 2009, S. 10). Diese Tatsache legt die Schlussfolgerung nahe, dass Prüfungsformate mit derartigem Umfang und sprachlicher Komplexität die eigentlich zu messende „kaufmännische Kompetenz zur Bewältigung typischer Geschäftsprozesse“ nicht zuverlässig messen können, da erhöhte Konzentrationsleistungen zur Situationserfassung notwendig werden. Diese können leicht zu kognitiver Überforderung und damit zum Scheitern an den Aufgaben führen.

Die Forschungsinitiative „Technologiebasierte Kompetenzmessung in der Berufsbildung“ (ASCOT) des Bundesministeriums für Bildung und Forschung (BMBF) hat sich daher zum Ziel gesetzt, valide computerbasierte Testinstrumente zur Abbildung beruflicher Kompetenzen zu entwickeln. Im Teilprojekt „Modellierung und Messung von Intrapreneurship-Kompetenz“ der Forschergruppe um Weber (Institut für Wirtschaftspädagogik, Fakultät für Betriebswirtschaft, Ludwig-Maximilians-Universität München) – auf das hier Bezug genommen wird – wurden 18 authentische Testaufgaben mit insgesamt 51 Items zur Messung der beruflichen Kompetenz „Intrapreneurship“ technologiebasiert konstruiert und als Teilbereich in die Unternehmenssimulation ALUSIM, wie diese von Achtenhagen und Winther (Achtenhagen/Winther 2009; Winther 2010) konstruiert und im ASCOT-Projekt parallel weiterentwickelt wurde, eingebettet (Weber u. a. 2014). Dabei werden in verschiedenen video-basierten Szenarien authentische intrapreneurship-bezogene Arbeitsanforderungen simuliert (z. B. ein IP-Projekt zum Aufbau eines neuen Vertriebsweges mittels Online-Shop;

Generierung von neuen Maßnahmen zur Rekrutierung von Mitarbeitern). Dabei werden die Auszubildenden/Probanden in Rollen hineinversetzt und erhalten mittels simulierter E-Mail einen Arbeitsauftrag, in den die Testaufgaben integriert sind. In dieser Studie beziehen wir uns auf die erste Version der Testaufgaben, die anhand der hier gewonnenen Ergebnisse überarbeitet wurden. Die Zielgruppe des Assessments sind Auszubildende zum/zur Industriekaufmann/Industriekauffrau am Ende ihrer beruflichen zwei- bis dreijährigen Erstausbildung.

Potentiale und Grenzen von Technologie bei authentischen Testaufgaben

Unter der Voraussetzung einer adäquaten technischen Ausstattung des Testortes bieten technologiebasierte Testumgebungen für die Datenerhebung und -auswertung eine Reihe von Vorteilen (Hartig/Klieme 2007; Jude/Wirth 2007). So ermöglicht der Einsatz computerbasierter Verfahren (z. B. in Form von Unternehmenssimulationen) eine realitätsbezogene fachdidaktische Aufbereitung der Aufgaben, die vergleichbar im Papier-Bleistift-Format nicht realisierbar wäre. Dazu gehört z. B. die Implementierung von ERP-gestützten (ERP: = Enterprise Resource Planning) Systemen sowie Endbenutzerwerkzeugen (wie z. B. Tabellenkalkulations-, Textverarbeitungs- oder Kommunikationsprogramme), die den kaufmännischen Berufsalltag dominieren (Jassmeier 2005; Ruf 2006). Neue Medien enthalten jedoch nicht per se ein Potenzial zu Lern-/Test-Innovationen. Ihre Wirksamkeit entfaltet sich erst in der adäquaten fachdidaktischen Aufbereitung einer authentischen Ziel- und Inhaltsdimension unter Berücksichtigung lern- und entwicklungstheoretischer Erkenntnisse, wie z. B. aus der Kognitionspsychologie zum Einsatz von Multimediaanwendungen (Mayer 2005) oder der Theorie der kognitiven Belastung (CLT: = Cognitive Load Theory; z. B. van Merriënboer/Sweller 2005; Blömeke 2003). Eine Reihe von Studien mit instruktionalem Fokus zeigt, dass die Reduzierung der kognitiven Belastung durch das Lernmaterial zu einer Steigerung der Lerneffizienz führt (Clark/Nguyen/Sweller 2005). Aus einer Assessmentperspektive verstärken sich die Herausforderungen an eine technologiebasierte fachdidaktische Aufgaben- und Plattformgestaltung; denn irrelevante/ungewollte Faktoren der Testumgebung bzw. der Testaufgaben wie aufmerksamkeitsablenkende multimediale Elemente, überhöhte Elementinteraktivität oder inhaltliche Redundanzen können neben der Fähigkeit eines Probanden ebenfalls Einfluss auf die Testergebnisse haben (Shaftel u. a. 2006; Cawthon u. a. 2012; Martiniello 2008; Haladyna/Downing/Rodriguez 2002; Haladyna/Downing 2004). Diese – bezogen auf die zu messende berufliche Kompetenz (als latentes Konstrukt) – ungewollt verursachten kognitiven Belastungen werden nach der CLT als extrinsischer kognitiver Load (ECL) bezeichnet und vermindern die Validität der Testergebnisse. Zuverlässig messende Testaufgaben/-arrangements hingegen ermöglichen Testergebnisse der Probanden, die sich – unter der Voraussetzung eines adäquaten Messmodels – allein auf Basis der Ausprägungsvariation der Messmodellparameter erklären lassen. Beispielsweise löst nach der Item-Response-Theory ein Proband mit 50 %iger Wahrscheinlichkeit ein Item, dessen Schwierigkeitsparameter gleich seinem Personen-Fähigkeitsparameter ist. Entsprechend wird das Testergebnis systematisch über die Schwierigkeit der Items beeinflusst. Schwierigkeitsbestimmende Merkmale sind domänenspezifisch. In der Domäne des Intrapreneurship können bspw. Eigenschaften wie eine Variation der ‚Elementinteraktivität' oder der ‚inhaltlichen Komplexität' die Schwierigkeiten charakterisieren. Die Bestimmung

und Einordnung der Formen kognitiven Loads (ECL vs. ICL) kann dabei in unterschiedlichen Fachdidaktiken unterschiedlich interpretiert werden. Während in der Studie von Gillmor, Poggio und Embretson (2014) zur Messung mathematischer Fähigkeiten die Kontextualisierung von Rechenaufgaben oder die Verwendung realer Zahlen als konstrukt-irrelevante Belastungen (ECL) interpretiert werden, sind – in Bezug auf die o. g. Überlegungen zur Authentizität – aus wirtschafts- und berufspädagogischer Perspektive derartige Belastungen als konstrukt-relevant und somit intrinsisch zu bewerten.

2. Ziel dieser Studie

Ziel der Studie ist es, mit Hilfe der CLT als Heuristik den kognitiven Load des von uns im Rahmen des ASCOT-Teilprojektes konstruierten Testinstruments, der authentischen technologiebasierten Testaufgaben zur Messung von Intrapreneurship, einzuschätzen und damit eine empirische Evidenz für die Zuverlässigkeit der Leistungsmessung i. S. d. Messick-Konzepts zu erhalten. Weitere Evidenzen wie zur Inhaltsvalidität finden sich bei Weber und Kollegen (2014; im Druck) und zur kognitiven Validität bei Bley (im Erscheinen). Dafür werden zunächst Kriterien extrinsisch und intrinsisch kognitiven Loads (ECL_{theo}, ICL_{theo}) aus der Kognitionspsychologie sowie der Fachdidaktik herausgearbeitet. Anschließend wird skizziert, welche Strategien zur Reduzierung des ECL im IP-Testarrangement der ALUSIM-Plattform unternommen wurden und wie die einzelnen IP-Testaufgaben theoretisch hinsichtlich der erarbeiteten ICL-Kriterien zu bewerten sind. Zur Prüfung der tatsächlich aufgetretenen kognitiven Belastungen werden Protokolle aus Studien lauten Denkens (LD) mit 26 Auszubildenden herangezogen sowie deren Leistungen beim Lösen der Aufgaben erfasst. Durch lautes Denken beim Lösen der Aufgaben können kognitive Lösungsprozesse sowie auftretende Belastungen sichtbar gemacht werden (Ericsson/Simon 1996; Pressley/Afflerbach 1995). Die verbalen Protokolle werden vollständig in Satz-für-Satz-Analysen deduktiv nach den Kategorien extrinsischer (ECL), intrinsischer (ICL) und kein Load (kL) kodiert (Chi 1997). In komparativen Analysen wird untersucht, ob sich die Höhe der theoretisch intendierten Loads empirisch über die Häufigkeiten der ICL/ECL-relevanten Nennungen in den verbalen Protokollen zum lauten Denken empirisch bestätigen lässt. Zusätzlich erfolgt die Prüfung der intendierten Loads anhand der tatsächlich erbrachten Leistungen der Probanden. Von den Erwartungen abweichende Aufgaben werden qualitativ inhaltlich auf mögliche Erklärungsmuster untersucht. Dieses Vorgehen hilft uns, die folgenden zwei zentralen Forschungsfragen zu beantworten:

FF 1: Lassen sich für die Formen kognitiven Loads die theoretisch intendierten Umfänge (ECL: möglichst vollständig reduziert, ICL: aufgabenspezifisch variierend) empirisch über die Häufigkeiten an ICL- und ECL-Nennungen in den verbalen Protokollen sowie über die tatsächlich erbrachten Leistungen in den jeweiligen Aufgaben bestätigen?
FF 1.1: Lässt sich die theoretisch intendierte vollständige Reduzierung von *ECL* empirisch über die Häufigkeiten an ECL bestätigen?
FF 1.2: Lassen sich die theoretisch intendierten unterschiedlichen Umfänge an *ICL* pro Aufgabe empirisch über die Häufigkeiten an ICL sowie über die tatsächlich erbrachten Leistungen in den Aufgaben bestätigen?

FF 2: Welche Erklärungsmuster lassen sich mittels vertiefender qualitativer Analysen für die unter FF 1 identifizierten Abweichungen finden?
FF 2.1: Welche *ECL*-relevanten Erklärungsmuster lassen sich finden?
FF 2.2: Welche *ICL*-relevanten Erklärungsmuster lassen sich finden?

3. Die Cognitive Load Theory als Heuristik zur Validierung kognitiver Belastungen bei der Konstruktion von Testaufgaben

3.1 Grundlagen der Cognitive Load Theory

Die Cognitive Load Theory basiert auf folgenden grundsätzlich anerkannten Annahmen, die sowohl für die Anwendung in Lern- als auch Testaufgaben Relevanz haben: Das Arbeitsgedächtnis ist im Gegensatz zum Langzeitgedächtnis in seiner Kapazität begrenzt und Informationen werden in Form von Schemata repräsentiert. Es wird angenommen, dass Schemata in abstraktem Format vorliegen und ggf. durch Übung automatisiert werden können (Ausbildung von Routinen). Jede Informationsverarbeitung erfordert kognitive Kapazitäten (cognitive load) des Arbeitsgedächtnisses (Sweller 2009; van Merriënboer/ Sweller 2005). Danach unterscheidet die derzeitige Forschung zur CLT drei Typen kognitiver Belastung: intrinsisch kognitive Belastung (ICL: = intrinsic cognitive load), extrinsisch kognitive Belastung (ECL: = extraneous cognitive load) und germane cognitive load (GCL). Der *ICL* hat seine Quelle in der inhaltlichen Komplexität der Domäne (intendierte Lernziele) und korrespondierend in der Aufbereitung des Testmaterials (z. B. über die intendierten Aufgabenschwierigkeiten hinweg). Die Höhe des ICL für einen Probanden hängt maßgeblich von seinem Vorwissen ab. Somit kann der ICL einer Aufgabe bei dem einen Probanden zu Überforderung führen, während er für einen anderen Probanden nur eine geringe oder keine Belastung darstellt. Wenn der ICL dem intendierten Lernziel entspricht, ist diese Überforderung bzw. das damit verbundene Scheitern an der Aufgabe messtheoretisch erwünscht, weil sich damit Rückschlüsse auf die Leistungsfähigkeit der einzelnen Probanden ziehen lassen. *ECL* hingegen ist unerwünscht und wird durch schlecht konstruierte Test-Arrangements verursacht. Der Proband muss zur Aufgabenbewältigung kognitive Ressourcen aktivieren, die unabhängig vom intendierten Lern-/ Aufgabenziel zusätzlich aufzuwenden sind. Hierzu zählen eine hohe Elementinteraktivität, die Auswahl aus überflüssigen oder die Suche nach versteckten Informationen – insofern nicht explizit als Lern-/ Aufgabenziel erwünscht – als auch eine wenig intuitive Benutzerführung des online-Lern-/ Testtools. Das Scheitern an einer Aufgabe aufgrund von ECL führt zu invaliden Rückschlüssen auf die latente Personenfähigkeit. Der *GCL* ist für den Aufbau von Schemata sowie deren Automatisierung verantwortlich und damit aus instruktionaler Perspektive sehr bedeutsam. Da aus der Assessmentperspektive die valide Leistungsfeststellung und weniger der Prozess des Wissenszuwachses im Fokus steht, wird der GCL im Folgenden nicht weiter betrachtet.

3.2 Cognitive Load bei Testaufgaben – Zum Stand der Forschung

Während die CLT in den Bereichen der Instruktionspsychologie und der Didaktik weit verbreitet und anerkannt ist, sind Studien zur kognitiven Belastung bei Test-

Arrangements/-Aufgaben bisher nur sehr selten. Die Forschergruppe um Kettler (Kettler u. a. 2011) modifizierte kalibrierte MC-Items mittels CL-Kriterien für den Einsatz bei Personengruppen mit Behinderung. Sie konnte zeigen, dass Items mit reduziertem Load zu höheren Leistungen führen. Miller (2011) wies nach, dass ästhetisch verbesserte Lernumgebungen die kognitive Belastung senken können und dass eine höhere Zufriedenheit der Testpersonen zugleich in besseren Leistungen resultiert. Gillmor, Poggio und Embretson (2014) untersuchten experimentell die Verbesserung der Validität von Mathematik-MC-Test-Items durch eine CLT-basierte Modifikation. Effektive Strategien zur Reduzierung des Loads (wie z. B. die Vermeidung redundanter Informationen oder die Reduzierung sprachlicher Komplexität) konnten identifiziert werden. Danach zeigen Studien, dass Test- ebenso wie Lernaufgaben von der Berücksichtigung der CLT profitieren. Bisherige Studien konzentrierten sich darauf, existierende Testaufgaben ex-post mittels Strategien zur Vermeidung von ECL zu modifizieren. Diese Studien beziehen sich inhaltlich primär auf den Bereich der Mathematik und des Lesens sowie auf einfache Multiple-Choice-Formate.

In der vorliegenden Studie gehen wir über diese bisherigen Ansätze insofern hinaus, als dass bereits bei der Aufgabenkonstruktion ECL und ICL berücksichtigt und dieser Load anschließend überprüft wird. Die zu prüfenden Aufgaben (hier in der ersten Version des IP-Testinstruments) besitzen aufgrund ihres authentischen Charakters überdies sehr variable Aufgabenformate (z. B. in Form von Handlungsprodukt-Aufgaben, Vervollständigungsaufgaben oder Auswahlaufgaben).

3.3 ECL-reduzierende Maßnahmen in der ALUSIM-Testplattform

ECL ist für das Lernen, aber auch für die valide Leistungsmessung ungünstig und sollte somit weitestgehend vermieden werden. In der Literatur diskutierte Strategien und Kriterien zur Reduktion sind: *(1) die Kontiguität, (2) die Signalisierung, (3) die Vermeidung von Redundanzen, (4) die Modalität, (5) die Reduzierung sprachlicher Komplexität, (6) die ästhetische/professionelle Darbietung sowie (7) die Sequenzierung* (Mayer 2005; Kettler u. a. 2011; Clark/Nguyen/Sweller 2005; Miller 2011; Gillmor/Poggio/Embretson 2014; van Merriënboer/Kirschner 2013; Sweller 2004). Die fachdidaktische Interpretation dieser Kriterien wird nachfolgend diskutiert.

Die *Kontiguität (1)* bezieht sich auf die maximale Verarbeitungskapazität und kann verbessert werden, indem Aufgabenformate und -inhalte systematisch, übersichtlich und organisiert dargeboten werden. Kaufmännische (IP-Projekt-)Arbeit ist durch eine simultane Verarbeitung von zum Teil mehreren Dokumenten unter Einsatz von zum Teil verschiedenen Endbenutzerwerkzeugen und darüber hinaus ggf. unter Hinzunahme von Hilfswerkzeugen gekennzeichnet. Daher bietet die ALUSIM-Benutzeroberfläche u. a. zwei Arbeitsbereiche sowie weitere im Vordergrund zu öffnende Hilfsmittel an. Zur Aufmerksamkeitsfokussierung sowie einer Erhöhung der Verarbeitungskapazität werden wichtige Elemente (z. B. Schlüsselwörter) speziell hervorgehoben (*Signalisierung (2)*) und (falls nicht durch den ICL intendiert) für die Aufgabenlösung unwichtige Elemente (wie z. B. der Geschäftsbericht des Unternehmens) weggelassen (*Vermeidung von Redundanzen (3)*). Die *Modalität (4)* spricht die Informationsaufnahme über unterschiedliche Verarbeitungskanäle (visuell und auditiv) an. Indem die Inszenierung der Realität (hier in Form einer „Cover Story“) über Videos anstatt über zu lesenden Text erfolgt und realitätsnah nachgebildete Endbenutzerprogramme anstelle von Arbeitsblättern mit abstrakten

Nachbildungen echter Anwenderprogramme zur Aufgabenlösung verwendet werden, wird eine erhebliche Entlastung der kognitiven Kapazität der Probanden ermöglicht. Die Aufgabenformulierung sowie die zu verwendenden Texte und Arbeitsmaterialien wurden explizit auf den Adressatenkreis abgestimmt (Reduzierung der *sprachlichen Komplexität (5)*). Die *ästhetische/professionelle Aufbereitung (6)* wurde in Analogie zum real existierenden Modellunternehmen vorgenommen (z. B. Integration einer E-Mail-Signatur). Durch *Sequenzierung (7)* wird ermöglicht, dass komplexe Sachverhalte entsprechend dem individuellen Lösungsverhalten strukturiert bzw. wiederholt werden können (z. B. indem Videos/E-Mails bei Bedarf wiederholt angesehen/gelesen werden können). Ziel dieser Aktivitäten war es, den ECL möglichst vollständig zu reduzieren. Ein Überblick über alle ECL-reduzierenden Aktivitäten für das IP-Testinstrumentarium (authentischen IP-Aufgaben) im Teilbereich der ALUSIM-Plattform sortiert nach den unterschiedlichen ECL-Kriterien ist in Tab. 1 dargestellt.

Tab. 1: Aktivitäten zur Reduzierung des ECL im IP-Testinstrumentarium der ALUSIM-Testplattform

ECL-Kriterien	**Aktivitäten zur Reduzierung des ECL**
Kontiguität	• zwei Arbeitsbereiche ermöglichen die simultane Bearbeitung von Dokumenten (oberes Fenster) und die Nutzung von Arbeitsprogrammen (unteres Fenster) • bei einer Vielzahl von Dokumenten sind diese durch eine übersichtliche Reiterstruktur sortiert • Fenster der zwei Arbeitsbereiche sind in ihrer Größe veränderbar • weitere Hilfswerkzeuge wie Notizen, Taschenrechner und Kalender öffnen sich im Vordergrund des unteren Fensters • Notizen werden innerhalb einer Aufgabe dauerhaft gespeichert • übersichtliche, realitätsnahe Anordnung und Nachbildung der Arbeitsprogramme und Zusatzfunktionen sowohl in der Plattform als auch in den Arbeitsprogrammen selber • wiederkehrende Routinen in der Aufgabenbearbeitung: Aufgabeneingang per E-Mail oder Video und Aufgabenblatt, Beendigung über „Aufgabe beenden“-Button
Signalisierung	• Arbeitsprogramme sind durch typische/bekannte Symbole und Bezeichnungen (z. B.: „Text“) gekennzeichnet, die Bezeichnung erscheint automatisch, sobald der Cursor auf dem Symbol steht • explizite Angabe des Pfades zu Vorlagen und Dokumenten im Arbeitsauftrag • farbige Unterstützung komplexer Tabellenkalkulationen • Schlüsselwörter werden hervorgehoben
Vermeidung von Redundanzen	• Vermeidung redundanter Informationen (Ausnahme: die Selektion von Informationen ist explizites Lernziel – siehe bspw. ICL: Aufgabe 3)
Modalität	• Videos erzählen die „Cover Story“: Projektvideos stellen Teammitglieder, Stand im Projektablauf und anstehende Aufgaben vor • abwechslungsreiche Darbietungen unterstützen die Informationsaufnahme und -verarbeitung über auditiven und visuellen Kanal

ECL-Kriterien	Aktivitäten zur Reduzierung des ECL
Reduzierung sprachlicher Komplexität	• Tutorial führt in die Unternehmenssimulation ein (Unternehmensgeschichte und -daten, Rolle des Probanden in der Unternehmenssimulation, Navigation und Funktionen der Plattform) • sowohl in den Arbeitsaufträgen (E-Mails) als auch in den zur Lösung notwendigen Arbeitsmaterialien: keine unnötigen Verneinungen, präzise Sprache, Vermeidung von verschachtelten Sätzen, konsistente Wortwahl
ästhetische/ professionelle Aufbereitung	• Nachbildung eines echten Unternehmens mit authentischem Datenkranz • professionelles E-Mail-Layout, Kalkulationen, Dokumente (Quellen, Protokoll, GANTT-Plan usw.), Videos mit echten Schauspielern und authentischen Materialien (z. B. Zeitungsmeldung)
Strukturierung/ unterstützende Information durch Sequenzierung	• Videos können so oft wie gewünscht angeschaut werden • Mitglieder der Projektgruppe mit Namen und Funktionsbereichen sind als Übersicht jederzeit aufrufbar • strukturierter Ausweis (meist durch Aufzählungszeichen), wenn die Aufgabe mehrere Teilaufgaben umfasst und/oder auf mehrere Vorlagen und Dokumente zurückgegriffen werden muss

3.4 Intendierter ICL in den Aufgaben zur IP-Kompetenz

Der ICL hat seine Quelle in der inhaltlichen Komplexität der Domäne/Aufgabe. Aus Assessmentperspektive gilt es, diesen systematisch über die verschiedenen Aufgaben hinweg zu variieren, um die trennscharfe Abbildung unterschiedlicher Leistungsniveaus der Probanden zu ermöglichen. Für die vorliegende Studie wurden unter Rückgriff auf die Literatur zu kognitiven Belastungen bei Lernaufgaben sowie zur Fachdidaktik folgende domänenspezifische Kriterien intrinsischen kognitiven Loads definiert: *(1) intendierte Informations-Redundanzen, (2) Elementinteraktivität (Wechselhäufigkeiten), (3) Unterstützungsumfang (completion (Vervollständigungsaufgabe, z. B. Lückentext) vs. conventional task (vollkommen offenes Aufgabenformat)), (4) Komplexität der Aufgabenstruktur (unabhängige vs. komplexe Aufgabe), (5) Vertrautheitsgrad der auszuführenden Handlung und (6) Tiefgang und Verknüpfung der auszuführenden kognitiven Prozesse* (Sweller/van Merriënboer/Paas 1998; van Merriënboer/Kirschner 2013; Sweller/Chandler 1994; Winther/Achtenhagen 2009; Seeber 2008; Nickolaus/Gschwendtner/Geissel 2008). Die fachdidaktische (Achtenhagen/Winther 2009) Interpretation sowie Beispiele für die einzelnen ICL-Kriterien werden nachfolgend beschrieben.

Berufliche Alltagssituationen zeichnen sich häufig durch schlecht strukturierte Probleme aus. Das bedeutet, aus der Vielzahl gegebener Informationen müssen notwendige von *redundanten Informationen (ad 1)* unterschieden werden. Dies ist in Aufgabe 3 unseres Itemsatzes, welche die Aufbereitung eines GANTT-Plans erfordert, beispielsweise derart gelöst, dass das für die Aufgabenlösung notwendige Protokoll aus dem letzten Projektmeeting neben den relevanten zeitlichen Informationen zu Arbeitspaketen u. a. auch redundante Informationen zu Teilschritten innerhalb der Arbeitspakete oder zu Verantwortungsbereichen durch Mitarbeiter enthält. Das Kriterium

wird mit „1“ (vorhanden) bewertet, wenn zwei oder mehr redundante Informationen in dieser Aufgabe enthalten sind, ansonsten mit „0“. Der Aspekt der *Elementinteraktivität (ad 2)* beschäftigt sich mit der Frage, wie viele Dokumente gleichzeitig verarbeitet werden müssen und inwiefern diese räumlich nah zu betrachten sind. Wie im Kap. 3.1 beschrieben, umfasst die Plattform zwei Arbeitsbereiche. Somit kann ein Dokument (oberes Fenster) gleichzeitig mit einem Endverarbeitungsprogramm (unteres Fenster) aufgerufen werden. Falls mehr als ein Dokument bzw. mehr als ein Endverarbeitungsprogramm zur Aufgabenlösung erforderlich sind, ist ein Wechsel innerhalb eines Arbeitsbereiches notwendig. Mit „1“ wird dieses Kriterium bewertet, wenn zwei oder mehr Wechsel notwendig werden, ansonsten mit „0“. Ein weiterer Aspekt betrifft den *Unterstützungsumfang (ad 3)*, den eine Aufgabe leistet. In Anlehnung an van Merriënboer und Kirschner (2013) werden die beiden Aufgabenformate „completion task“ und „conventional task“ in der IP-Testumgebung eingesetzt. Hierbei handelt es sich um die zwei Formate mit der geringsten gegebenen Unterstützung (2013). In Aufgaben im Format einer „completion task“ (mit „0“ bewertet) werden Teillösungen z. B. in Form von Entscheidungsvorlagen vorgegeben, welche der Proband beurteilen, verarbeiten bzw. vervollständigen muss. Dieses ist z. B. in Aufgabe 11 der Fall, wobei für eine Teaminteraktionssituation eine Beurteilung aus gegebenen Handlungsalternativen vorgenommen werden muss. Bei Aufgaben im Format einer „conventional task“ (mit „1“ bewertet) ist eine Unterstützungsleistung kaum bis gar nicht gegeben. Hierzu zählen u. a. Aufgaben, die begründete Entscheidungen (z. B. Aufgabe 13) oder auch ein Brainstorming (z. B. Aufgabe 17) verlangen. Die *Komplexität der Aufgabenstruktur (ad 4)* bezieht sich nur auf Aufgaben mit zwei oder mehr Teilaufgaben (=Items). Dies betrifft alle Aufgaben außer 1, 2, 10 und 11. Hierbei wird unterschieden, ob der Lösungsprozess der Teilaufgaben aufeinander aufbauend (= aufbauend: mit „1“ bewertet) oder unabhängig voneinander (= unabhängig: mit „0“ bewertet) ist. Aufgabe 13 umfasst zwei Aufgabenteile. Im ersten Teil (Zuordnungsaufgabe) geht es darum, Maßnahmen und Folgen einer betriebswirtschaftlichen Entscheidung miteinander in Verbindung zu bringen. Im zweiten Teil erfolgt die begründete Auswahl für die unternehmensseitig beste Maßnahmen-Folgen-Entscheidung. Messtheoretisch notwendige Aufgabenunabhängigkeit wird trotz der inneren Aufgabenkomplexität gewahrt, indem Fehler bzw. Unvermögen im ersten Aufgabenteil zur Berücksichtigung von Folgefehlern bzw. zu fehlenden Werten im zweiten Aufgabenteil führen. Das heißt, eine richtige Lösung im Teilschritt 2 bei falscher Lösung im Teilschritt 1 wird durch die Berücksichtigung von Folgefehlern gewährleistet. Das Kriterium der *Vertrautheit mit der auszuführenden Handlung (ad 5)* rekurriert auf den Unterschied zwischen schulähnlichen Aufgabenformaten, wie sie üblicherweise in Leistungsfeststellungen verwendet werden (gilt als eher berufsschulisch vertraut: = 0) und realitätsnahen beruflichen Handlungssituationen (gilt als eher berufsschulisch unvertraut: = 1), wie sie an kaufmännischen Arbeitsplätzen vorkommen. Auch wenn das Assessment einen hohen Grad an Realitätsnähe anstrebt, so bleibt – in der Testsituation – aus Sicht der Auszubildenden trotzdem eine höhere Assoziation zu berufsschulischen Leistungsfeststellungsverfahren als zum betrieblichen Alltag. Dies ist nicht zuletzt bedingt durch den Ort der Testadministration: die Berufsschule. Das Vervollständigen eines GANTT-Plans unter Zuhilfenahme eines Tabellenkalkulationsprogramms (Aufgabe 3) ist eher eine berufsnahe Handlung, wie sie bisher selten in bekannten Schulleistungsformaten vorkommt. Die Bewertung und Auswahl der besten Handlungsalternative für gute Teamarbeit (Aufgabe 11) hingegen ähnelt dem in schulischen Leistungstests häufiger verwendeten Format von best-choice-Verfahren. Während

die vorangegangenen Kriterien primär struktureller Natur sind, fokussiert das letzte Kriterium *Tiefgang und Verknüpfung (ad 6)* auf die inhaltliche Komplexität und Tiefe der Aufgabe. Um der Inhaltlichkeit bei der Bestimmung des Gesamtindizes ICL eine adäquate Gewichtung zu geben, wird das Kriterium „Tiefgang und Verknüpfung" mit „2" (statt mit „1") für hohen Tiefgang und „0" für geringen Tiefgang bewertet. Die unternehmensseitig beste Maßnahmen-Folgen-Entscheidung (Aufgabe 13) impliziert beispielsweise einen hohen Grad an Tiefgang und Verknüpfung (und wurde damit mit „2" bewertet), da sie ein Verständnis für die Gestaltungslogik der sich marktbedingt veränderten Kosten-/Gewinnkalkulationen und der damit verbundenen Maßnahmen und Folgen sowie eine Interpretationsleistung bezüglich der sich ergebenden Ergebnisse unter Anwendung betriebswirtschaftlicher Routinen/Theorien erfordert. In Tab. 2 sind alle 18 Aufgaben hinsichtlich der eingeführten ICL-Kriterien klassifiziert. Die letzte Zeile gibt einen prozentualen Gesamtsummenwert des ICL einer jeden Aufgabe an. Dieser ergibt sich als Summe der ICL-Kriterien-Werte zuzüglich eines Basis-ICL von „1" relativiert an dem Maximalwert von „8" (Ausnahmen bilden die Aufgaben 1, 2, 10, 11, für die das Kriterium Aufgabenkomplexität nicht berücksichtigt werden kann: Maximalwert = 7). Der Idee des ICL folgend wird für Aufgaben mit hohem intrinsisch kognitiven Load mehr kognitive Kapazität erforderlich, was wiederum zur Folge hat, dass Personen mit geringerer Leistungsfähigkeit häufiger an diesen Aufgaben scheitern. Die vorgestellten theoretischen Überlegungen ergeben, dass der Aufgabenpool 6 einfache (0–33 %), 8 mittelschwere (34–66 %) und 4 schwere Aufgaben (67–100 %) zur Messung von IP-Kompetenz beinhaltet.

Tab. 2: Kriterien-basiert intendierter ICL der IP-Aufgaben

	IP-Aufgaben																	
ICL-Kriterien	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
Redundanz[1]	0	0	1	1	0	0	0	0	1	0	0	0	1	0	0	1	0	0
Element-interaktivität[2]	0	0	1	1	0	0	1	0	1	0	0	1	1	0	0	1	0	0
Unterstützungs-umfang[3]	0	0	0	0	0	0	1	0	1	0	0	0	1	0	1	1	1	1
Aufgaben-komplexität[4]	–[5]	–[5]	0	0	0	0	1	0	1	–[5]	–[5]	0	1	1	0	0	0	0
Vertrautheit der auszuführenden Handlung[6]	0	0	1	0	0	0	1	0	1	0	0	0	1	1	1	1	1	0
Tiefgang, Verknüpfung[7]	0	0	2	0	2	0	2	2	2	0	0	0	2	2	2	0	2	2
Intendierter ICL_theo,rel[8]	**14**	**14**	**75**	**38**	**38**	**13**	**88**	**38**	**100**	**14**	**14**	**25**	**100**	**63**	**63**	**63**	**63**	**50**

Anmerkungen: [1] zwei oder mehr überflüssige Informationen = 1, sonst 0; [2] zwei oder mehr Wechsel notwendig = 1, sonst 0; [3] conventional task = 1, completion task = 0; [4] aufbauend = 1, unabhängig = 0; [5] fehlender Wert, da Aufgabe nur eine Teilaufgabe umfasst; [6] eher schulisch vertraut = 1, eher schulisch unvertraut = 0; [7] hoher Tiefgang = 2, sonst 0; [8]der intendierte ICL ergibt sich aus der Summe der ICL-Kriterien erhöht um einen Basis-ICL von 1 relativiert an der Anzahl der Merkmale (in %; maximale Summe 8 = 100 %).

Um zu prüfen, ob die intendierten ECL- und ICL-Umfänge tatsächlich vorliegen, stellt sich nun die Frage nach einem geeigneten Evaluationskonzept.

3.5 Lautes Denkens als Methode zur Überprüfung des CL

Brünken, Plass und Leutner (2003) klassifizieren die unterschiedlichen Methoden zur Messung von CL nach zwei Dimensionen: der Objektivität (objektive vs. subjektive Maße) und dem kausalen Zusammenhang (indirekte vs. direkte Messverfahren). Subjektive Maße sind dabei i. d. R. Selbsteinschätzungen entweder bezogen auf die investierte Anstrengung (indirekte Messung) oder bezogen auf die Schwierigkeit des Aufgabenmaterials bzw. auf den eigenen Stresslevel (direkte Messung). Objektive Messverfahren sind vom Subjekt nicht beeinflussbar. Dabei werden ebenfalls direkte (u. a. die Messung von Hirnaktivitäten oder sogenannte Dual-task-Messungen) von indirekten Maßen (u. a. Verhaltens-, Eye-tracking- oder Leistungsmessungen) unterschieden. Die vorgestellten Studien zu CL bei Testaufgaben greifen auf indirekt objektive Maße zurück, indem sie in experimentellen Designs modifizierte (ECL-reduzierte) vs. nicht-modifizierte Items einsetzen und die erzielten Leistungen der Probanden vergleichen. Idee der vorliegenden Studie ist es, bereits in der Phase der Aufgabenkonstruktion ECL- und ICL-Kriterien zu berücksichtigen. Zur Evaluation des intendierten ICL und ECL sollen daher gleichzeitig zur Aufgabenbearbeitung stattfindende audioaufgezeichnete und protokollierte Studien lauten Denkens in Kombination mit Leistungsdaten der Probanden herangezogen werden (Kalyuga/Plass 2009). Simultane Studien lauten Denkens (concurrent think alouds) ermöglichen die bis dato bestmögliche Erfassung der während der Aufgabenlösung stattfindenden kognitiven Prozesse (Ericsson/Simon 1996). Obwohl Studien lauten Denkens hypothetisch kognitiven Load erhöhen können, zeigen erste Ergebnisse, dass diese Methode den kognitiven Prozess nicht zu behindern scheint (Leow/Morgan-Short 2004). Nach der Klassifikation von Brünken et al. (2003) lässt sich die Methode des lauten Denkens zur Erfassung kognitiver Belastungen als ein objektives direktes Messverfahren einstufen. Damit wird in dieser Studie eine Kombination aus objektiv direktem (Methode des Lauten Denkens) und objektiv indirektem (Leistungsmessung) Verfahren gewählt.

4. Methodik

Instrument

Intrapreneurship (IP) ist eine fortgeschrittene kaufmännische Kompetenz, die die Planung und Umsetzung von Projekten sowie die Generierung neuer Projektideen umfasst. Inhaltlich basiert die Konstruktion der 18 IP-Testaufgaben auf einer umfassenden Domänenanalyse (Weber u. a. 2014). Neben den in dieser Studie diskutierten Kriterien zu kognitiven Belastungen wurden außerdem fachdidaktische sowie messtheoretische Anforderungen bei der Entwicklung der Testumgebung/-aufgaben berücksichtigt. Alle Aufgaben sind in die – einer realen Unternehmung nachempfundenen – Unternehmenssimulation der ALUSIM GmbH eingebettet. Die Probanden versetzen sich zum Lösen der Aufgaben in die Rolle des ALUSIM-Auszubildenden Leon Stein. Dieser wirkt in zwei Projektgruppen mit und übernimmt dabei verschiedene Arbeitsaufträge, die ihn wie in der Realität per E-Mail erreichen: z. B. die ‚Beurteilung guter/schlechter Teamarbeit' (Aufgabe 11) oder die ‚Generierung eines GANTT-Plans auf Basis des Teammeeting-Protokolls' (Aufgabe 3). Die Aufgaben besitzen aufgrund ihrer hohen Realitätsnähe sehr unterschiedliche

Aufgabenformate mit variierender inhaltlicher Komplexität (Tiefgang/Verknüpfung) und struktureller Komplexität (Aufgabenformat, Unterstützungsumfang, Vertrautheitsgrad der auszuführenden Handlung) (siehe auch Tab. 2) und einer unterschiedlich großen Anzahl an Items pro Aufgabe (zwischen 1 und 8; insgesamt 51 Items). Über Videos, Podcasts und E-Mails werden die Probanden über Hintergrund und Fortgang beider Projektarbeiten informiert. Basis der Erhebung bilden die zum Zeitpunkt der Erhebung (Frühjahr 2013) vorliegenden Testaufgaben im ersten Entwurf der technologiebasierten ALUSIM-IP-Testplattform.

Stichprobe und Design

Für die Validierungsprüfung dienen Transkripte zu Studien Lauten Denkens von Auszubildenden (N = 26) am Ende ihrer beruflichen Erstausbildung zum Industriekaufmann/zur Industriekauffrau (Bley im Erscheinen). Die freiwillig teilnehmenden Auszubildenden (12 Männer und 14 Frauen) waren dazu aufgefordert, während eines Zeitraumes von je ca. zwei Zeitstunden alle Aufgaben der ALUSIM-IP-Testplattform am Computer zu lösen. Sie sind zum Erhebungszeitpunkt in kleinen (n = 1), mittleren (n = 4), großen (n = 17) und sehr großen (n = 4) Unternehmen aus München bzw. dem Münchner Umland beschäftigt und im Durchschnitt 21 Jahre alt (SD = 1,89). Die Hälfte der Auszubildenden besitzt die Hochschulzugangsberechtigung, die andere einen Abschluss der mittleren Reife. Das Teilnehmer-Recruiting erfolgte per Werbung und Selbstselektion; alle Teilnehmer erhielten für ihre Kooperation eine Aufwandsentschädigung von 40€.

Durchführung

Jede der 2-stündigen Studien wurde in separaten Räumlichkeiten des Instituts für Wirtschaftspädagogik durchgeführt. Die Arbeitsplätze waren mit Computern, auf denen die technologiebasierte Simulation ALUSIM lief, ausgestattet. Die Durchführung der Sitzungen wurde von geschulten Test-Administratoren unterstützt, die außerhalb des Sichtfeldes der Teilnehmer saßen und den Lautes-Denken-Prozess nur nach einer Stille von mehr als 10 Sekunden unterbrachen, um den Probanden an das Weitersprechen zu erinnern (Pressley/Afflerbach 1995, S. 125f.; Ericsson/Simon 1996, S. 82f.). Zur Erzielung eines hohen Grades an Reliabilität starteten die Test-Administratoren mit einer standardisierten Instruktion (Ericsson/Simon 1996, S. 80). Da „laut zu denken" für die Teilnehmenden eine kaum bekannte Technik ist (Ericsson/Simon 1996, S. 78), übten die Administratoren daher mit den Teilnehmern durch Vormachen und Nachmachen laut zu denken, während sie logische Puzzles lösten. Die Studien des lauten Denkens wurden mittels Tonbandgerät aufgezeichnet und in Anlehnung an Dresing & Pehl (2011) transkribiert. Die Antworten zu jeder Aufgabe wurden über die ALUSIM-Plattform automatisch erfasst.

Quantifizierung verbaler Daten

Für die Auswertung der verbalen Daten folgen wir dem Ansatz von Chi (1997): Zunächst muss entschieden werden, ob *eine Reduzierung bzw. die Ziehung einer Stichprobe (1)* aus der Gesamtheit des vorliegenden Materials erfolgen soll. Die 26 Transkripte umfassen insgesamt 220 Seiten (Arial, einzeilig, Schriftgröße 11). Mit der Absicht, auch quantitative Analysen durchzuführen, wird das vollständige Protokollmaterial kodiert. Da viele Verfahren der induktiven Statistik metrisches Datenniveau

voraussetzen, werden Sätze (N = 8841) anstelle von Abschnitten oder Sinneinheiten als *Segmentierungseinheit (2)* gewählt. Der Kodierprozess ist dreistufig angelegt: Erstens erfolgt eine Satz-für-Satz basierte Entscheidung, ob der jeweilige Satz auf eine kognitive Belastung schließen lässt oder ob es sich um einen „normalen" Lösungsprozess (kein Load: = kL) handelt. Zweitens werden belastungsrelevante Sätze nach extrinsischer (ECL) bzw. intrinsischer (ICL) kognitiver Belastung unterschieden. Diese Datenbasis bildet die Grundlage für die quantitativen Analysen. Für die zweite Forschungsfrage – die qualitative Analyse – erfolgt eine dritte Kodierstufe (siehe Kap. 5.2, Forschungsfrage 2, Tab. 5 und 7). In Bezug auf die ersten beiden Kodierstufen sind das *Kodierschema (3)* sowie eine *Operationalisierung in Form von Evidenzen (4)* in Tab. 3 abgebildet. Die Kodierung wurde von zwei unabhängigen Kodierern, die mit der kognitiven Belastungstheorie gleichermaßen wie mit den Aufgaben zum IP vertraut waren, durchgeführt. Ein Minimum von drei Durchgängen durch die Transkripte unter Nutzung der Software NVivo 9 (Bazeley/Richards 2000) erzielte ein stabiles Kodierschema basierend auf akzeptablen Zuordnungen (Cohens κ = .923; Wirtz/Caspar 2007, S. 55). Die Schritte *(5) bis (7)* im Ansatz von Chi beinhalten die *Datenauswertung*, die *Darstellung der Ergebnismuster* sowie deren *Interpretation*, wie sie im Ergebniskapitel diskutiert werden.

Tab. 3: Kodierschema

Kategorie	Kodierregel	Kodierbeispiele
ECL	die Aufgaben-gestaltung/-darbietung betreffende Belastungen	A5, LD18: „Mist, das geht nicht zu markieren. Wie funktioniert denn das nur?"
ICL	den Inhalt/die Lösung der Aufgabe betreffende Belastungen	A9, LD08: „Häh? Nein, ich muss es doch abziehen, minus 16800, ach jetzt, das stimmt gar nicht. Was muss ich denn da noch abziehen? Minus 1200 vielleicht? Hmm blöd, blöd, blöd. Break-Even-Point, wie rechnet man das jetzt gleich nochmal? Ach wie geht das jetzt?"
Keine Load (kL)	Lösungsprozess ohne erkennbare Belastung	A6, LD02: „Dann geh ich jetzt mal in die Analyse rein. Da kann man weiterscrollen: Absatzkontingent für Händler Phonestore beträgt pro Quartal (.) die komplette Absatzmenge beträgt (.) das stimmt nicht, das ist der Buyweb."

Anmerkungen: A5 = Aufgabe 5; LD18 = Lautes-Denken-Protokoll des Auszubildenden mit der Nummer 18; (.) = kurze Pause

Neben den Variablen des empirisch erhobenen ECL und ICL werden für die Validierung des theoretischen ICL zusätzlich die tatsächlich erbrachten Leistungen der Probanden in jeder Aufgabe herangezogen.

Analysen

Die Validierung der technologiebasierten Testaufgaben erfolgt im Mixed Method Design. Auf Basis der quantifizierten verbalen Daten werden zunächst quantitative (Friedman-Test, Vorzeichen-Rang-Tests von Wilcoxon) und deskriptiv-grafische Analysen zur Identifikation von Erwartungsabweichungen vorgenommen (FF 1). Alle Analysen werden mit SPSS Statistics 19 durchgeführt. In Forschungsfrage 2

werden mittels deduktiver qualitativer Analysen Erklärungsmuster für diese Abweichungen identifiziert.

5. Ergebnisse

Tab. 4 beschreibt die zur quantitativen Analyse relevanten Variablen: theoretischer Load (*ICL_theo,rel)*, empirischer Load (*ECL_emp*, *ICL_emp,* bzw. *ECL_emp,rel*; *ICL_emp,rel*) und Leistung (*Leistung_rel*). Die Beschreibung erfolgt zunächst ungeachtet der aufgabenspezifischen Unterschiede auf Basis von 457 Fällen (26 Probanden bei je 18 Aufgaben; 11 fehlende Werte, da zwei Probanden die Plattform zeitlich bedingt nicht beendeten).

Die Variable *ICL_theo,rel* beschreibt den theoretisch angenommenen intrinsischen Load der Aufgaben, wie er in Tab. 2 bestimmt wurde. Dieser ist am geringsten für Aufgabe 6 ($ICL\_theo,rel_6 = 0,13$) und am höchsten für die Aufgaben 9 und 13 ($ICL\_theo,rel_{9,13} = 1,00$). Aufgrund des mittleren Wertes von 0,48 und der Standardabweichung von 0,29 besitzen die Aufgaben eine angemessen variierende (theoretisch intendierte) Schwierigkeitsverteilung.

Tab. 4: Deskription der Variablen

	Minimum	**Maximum**	**MW**	**SD**	**Modus**	**Median**	**Schiefe** (StdF = 0,11)	**Kurtosis** (StdF = 0,23)
ICL_theo,rel	0,13	1	0,48	0,29	0,14	0,38	0,35	-1,09
ECL_emp	0	37	1,75	4,21	0	0	4,78	30,09
ECL_emp,rel	0	0,97	0,06	0,12	0	0	2,71	11,19
ICL_emp	0	50	2,02	4,26	0	0	4,81	38,89
ICL_emp,rel	0	1	0,08	0,14	0	0	2,58	9,63
Leistung_rel	0	1	0,58	0,45	1	0,75	-0,36	-1,69

Anmerkungen: MW = Mittelwert, SD = Standardabweichung, StdF = Standardfehler; Die asymptotische Signifikanz (2-seitig) des Kolmogorov-Smirnov-Tests ergibt für alle Variablen p-Werte < 0,01, sodass für keine der Variablen Normalverteilung angenommen werden kann.

Die folgenden Zeilen beschreiben die empirisch erhobenen Werte des kognitiven Loads (*ECL* und *ICL*). Dabei sind zunächst die absoluten (*ECL_emp*, *ICL_emp*) und anschließend die relativierten Werte (*ECL_emp,rel*; *ICL_emp,rel*) aufgeführt. Die Relativierung erscheint notwendig, da sich die Aufgaben hinsichtlich ihrer Aufgabenkomplexität stark unterscheiden und somit die absoluten Werte zwischen den Aufgaben nicht vergleichbar sind. Die Verteilung der absoluten sowie der relativen empirischen Werte zum intrinsischen und extrinsischen Load ist stark linkssteil. Dies ist im Wesentlichen damit zu begründen, dass eine Vielzahl von Personen hinsichtlich einer Vielzahl von Aufgaben keinen ECL und/oder ICL nennt. Die Anzahl der Ausprägung „0“ beträgt bei ECL n = 309 Fälle (= 67,6 %) und bei ICL n = 271 Fälle (= 59,3 %). Diese Verteilung schlägt sich entsprechend auch im Median und Modus nieder, die für alle vier Variablen Null sind. Maximal werden in einem Fall (Proband 13 in Aufgabe 9) 37 ECL-relevante Sätze genannt. Relativiert ergibt dieser Wert

97 % (ECL_emp,rel$_{13,9}$). Dieser Wert ist gleichzeitig Maximum von *ECL_emp,rel*. Die maximale Anzahl an ICL-relevanten Sätzen, die ein Proband in einer Aufgabe nennt, ist 50 (Proband 6 in Aufgabe 3), was relativiert einem Wert von 67 % (ICL_emp,rel$_{6,3}$) entspricht. Das Maximum von ICL_emp,rel liegt bei 1 (entspricht 100 %). Dieser relative Wert entspricht absolut 10 ICL-relevanten Sätzen (Proband 7 in Aufgabe 1). Somit ergeben sich für beide Formen der kognitiven Belastung sehr geringe Mittelwerte von 1,75 ECL-Sätzen (entspricht einem Durchschnitt von 6,22 % aller Sätze pro Aufgabe) und 2,02 ICL-Sätzen (entspricht einem Durchschnitt von 7,92 % aller Sätze pro Aufgabe). Daraus lässt sich schließen, dass die Probanden während des gesamten Lösungsprozesses nur durchschnittlich 14 % belastungsrelevante Sätze nannten.

Die letzte Variable beschreibt die durchschnittliche Leistung der Probanden. Sie wird ebenfalls relativ angegeben, da die Aufgaben ungleich viele Items besitzen (*Leistung_rel*). Die Variable nimmt damit Werte zwischen 0 und 1 (Min/Max) an, hat ihren Median bei 0,75 und den Modus bei 1. Mittelwert und Standardabweichung betragen 0,58 und 0,45. Die Verteilung der Variable *Leistung_rel* ist leicht rechtssteil (Schiefe < 0).

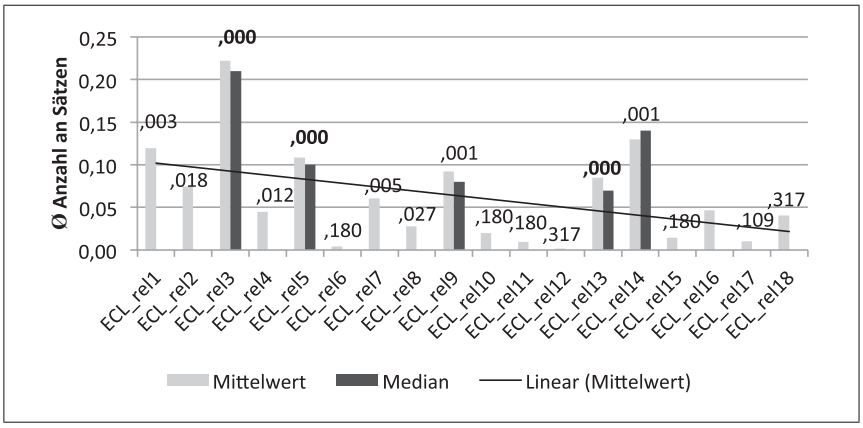
5.1 Ergebnisse der quantitativen Analyse

Forschungsfrage 1.1: Lässt sich die theoretisch intendierte vollständige Reduzierung von ECL empirisch über die Häufigkeiten an ECL bestätigen?

Der theoretisch angenommene ECL beträgt für alle Aufgaben Null. Dieses Ziel konnte, wie die deskriptiven Analysen bereits vermuten lassen, nicht vollständig erreicht werden. Es stellt sich nun die Frage, ob es aufgabenspezifische Unterschiede in der Höhe der Abweichung gibt. Übergreifend wird dies mit dem Friedman-Test (nicht-parametrische einfaktorielle Varianzanalyse mit Messwiederholung) geprüft (Bühner/Ziegler 2008, S. 466 ff.). Die Methodenwahl ist darin begründet, dass eine nicht-normalverteilte abhängige Variable *(ECL_emp,rel)* geprüft wird, die aufgrund der Personen*Aufgaben-Struktur eine abhängige Datencharakteristik besitzt. Der Friedman-Test testet die Nullhypothese, dass sich die 18 Aufgaben in der Höhe ihrer relativen ECL-Werte nicht signifikant voneinander unterscheiden. Die mittleren Ränge des *ECL_emp,rel* variieren zwischen 7,07 (für Aufgabe 17) und 16,52 (für Aufgabe 3). Im Ergebnis heißt das, dass sich die Aufgaben signifikant in der Höhe ihrer ECL-Äußerungen voneinander unterscheiden ($X^2 = 134{,}693$, df = 17, asymptotische Signifikanz: p = 0,000, Power = 1) (Faul u. a. 2009). Es gibt demnach Aufgaben, die – wie gewünscht – wenig extrinsischen Load zeigen und welche, die noch zu viel extrinsischen Load zeigen. Abbildung 1 veranschaulicht für jede Aufgabe Median und Mittelwert. Da in einer Vielzahl von Fällen kein ECL geäußert wird, ist bei 13 der 18 Aufgaben der Median Null. Das bedeutet, dass bei diesen Aufgaben mehr als die Hälfte der Probanden keine extrinsische Belastung äußert. Dennoch wird für alle Aufgaben zusätzlich post-hoc geprüft, ob die Mediane der Probanden gleich Null ausfallen. Dafür werden 18 Vorzeichen-Rang-Tests von Wilcoxon (nicht-parametrischer Test für die Prüfung abhängiger Stichproben) genutzt (Bühner/Ziegler 2008, S. 267 ff.). Das Signifikanzniveau (von 0,05) wird nach Bonferroni-Korrektur (0,05/18) auf den Wert alpha = 0,00056 gesetzt (Bühner/Ziegler 2008, S. 547). Dementsprechend sind alle p-Werte, die kleiner sind als 0,000, als

signifikant zu bewerten und sagen aus, dass für diese Aufgaben eine bedeutende Höhe an ECL-Äußerungen vorliegt. Die empirischen p-Werte einer jeden Aufgabe sind in Abbildung 1 abgetragen. Drei der fünf Aufgaben mit einem Median größer als Null sind signifikant (Aufgabe 3, 5 und 13) und sollten somit auch qualitativ betrachtet werden. Trotz des nicht-parametrischen Datencharakters interessiert uns zusätzlich die Verteilung der Mittewerte über die Aufgaben, da ansonsten große Schwierigkeiten einzelner Personen nicht sichtbar werden. Die Trendlinie der Mittelwerte zeigt deutlich, dass besonders am Anfang durchschnittlich mehr ECL-Sätze geäußert werden als im Fortgang der Plattform.

Für die Aufgaben 3, 5 und 13 sowie die anfänglichen Aufgaben 1 und 2 sollen in der FF 2 (qualitative Analyse) Erklärungsmuster herausgearbeitet werden.



Legende: Für jede Aufgabe ist das asymptotische Signifikanzniveau des Vorzeichen-Rang-Tests von Wilcoxon abgetragen. Fett gedruckte Werte sind unter Berücksichtigung der alpha-Fehler-Korrektur signifikant.

Abb. 1: Mittelwerte und Mediane des empirischen Loads (ECL_emp,rel) pro Aufgabe

Forschungsfrage 1.2: Lassen sich die theoretisch intendierten unterschiedlichen Umfänge an ICL pro Aufgabe empirisch über die Häufigkeiten an ICL sowie über die tatsächlich erbrachten Leistungen in den Aufgaben bestätigen?

Durch unterschiedliche Kombinationen der ICL-Kriterien in den einzelnen Aufgaben ergeben sich in der Summe entsprechend unterschiedliche Gesamt-ICL-Werte pro Aufgabe. Es wird angenommen, dass (1) ein hoher theoretischer ICL-Wert (*ICL_theo,rel*) eine schwierige Aufgabe darstellt, die beim Lösen wiederum durchschnittlich relativ viele intrinsisch-kognitive Belastungsäußerungen (*ICL_emp,rel*) auslöst. Des Weiteren wird (2) angenommen, dass ein hoher theoretischer ICL-Wert eher zum Scheitern an der Aufgabe (*Leistung_rel*) führt, im Gegensatz zu einer Aufgabe mit einem geringen theoretischen ICL-Wert (vgl. Abb. 2). Diese beiden Betrachtungsweisen sollen zur Validierung des theoretisch intendierten intrinsisch-kognitiven Loads herangezogen werden.

Die aufgabenübergreifende Betrachtung mittels Spearman-Rangkorrelation (Bühner/Ziegler 2008, S. 612 ff.) bestätigt beide Vermutungen: (1) die Korrelation

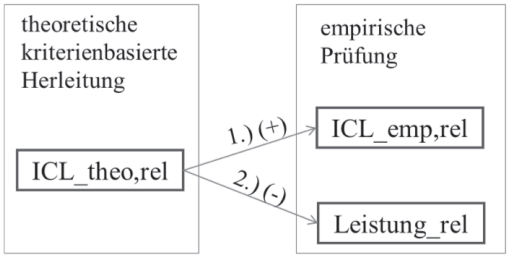


Abb. 2: Struktur zur Validierung des ICL

zwischen *ICL_theo,rel* und *ICL_emp,rel* beträgt 0,273, p = 0,000 und (2) die Korrelation zwischen *ICL_theo,rel* und *Leistung_rel* beträgt -0,233, p = 0,000. Beide Korrelationen sind damit auf dem 1 % Niveau signifikant und besitzen einen mittleren Effekt (Bühner/Ziegler 2008, S. 603). Entsprechend der Betrachtung des ECL interessiert uns auch für den ICL die Validierung auf Aufgabenniveau. Das heißt, ist unsere Erwartung über die Höhe des ICL für alle Aufgaben gleichermaßen gültig bzw. welche Aufgaben zeigen deutliche Abweichungen?

Die Validierung mittels Varianzanalysen ist hierbei nicht möglich, da keine vergleichbaren Daten vorliegen. Die drei Variablen besitzen unterschiedliche Metriken und stark voneinander abweichende Verteilungen. Eine Standardisierung mittels z-Transformation setzt normalverteilte Daten voraus (welche nicht gegeben sind, siehe Tab. 4). Die Stanine-Normalisierung hingegen ist zwar eine nicht-lineare Transformation, sie bewirkt im vorliegenden Fall jedoch einen zu starken Informationsverlust, was zu einer Verzerrung der Ergebnisse führt[2] (Bühner/Ziegler 2008, S. 56 ff.). Grund hierfür könnte die Überrepräsentation von Nullwerten in den Daten sein. Daher erfolgt die aufgabenbasierte komparative Analyse rein deskriptiv/grafisch.

In Analogie zum ECL werden in Abbildung 3 Median und Mittelwert einer jeden Aufgabe für die empirischen ICL-Werte sowie die Mittelwert-Trendlinie über die Aufgaben hinweg abgebildet. Hierbei zeigt sich, dass 6 der 18 Aufgaben einen Median größer Null besitzen. Die drei Aufgaben 3, 9 und 13 sind theoretisch als schwere

2 Ein Beispiel: Die Aufgabe 1 ist theoretisch intendiert eine eher einfache Aufgabe (14 % Schwierigkeit). Betrachtet man die belastungs-relevanten Äußerungen, zeigt sich, dass 23 der 26 Probanden keinen ICL-relevanten Satz äußern. Für die anderen drei Probanden liegt der relative ICL bei 0,3, 0,7 bzw. 1,0. Im Mittel zeigen also – wie erwartet – nur wenige Probanden Schwierigkeiten mit der Aufgabe.
Die Prüfung der Daten bei Verwendung der Stanine-transformierten Daten ergibt im Vorzeichen-Rang-Test von Wilcoxon ein Signifikanzniveau von 0,000: Pro Variable werden Rangprozentwerte (P) für jeden Datenpunkt gebildet: theoretischer ICL: P-ICL_theo,rel$_1$ = 17,07 (unabhängig vom Probanden); empirischer ICL: der minimale Wert von 0 ergibt einen P-ICL_emp,rel$_1$ = 29,76; der maximale Wert von 1 ergibt einen P-ICL_emp,rel$_1$ = 100. Der Vorzeichen-Rang-Test bildet für jeden Probanden die Differenz aus den beiden transformierten Variablen und bestimmt darüber negative (Differenz negativ) und positive (Differenz positiv) Ränge. Anschließend werden Rangsummen für negative und positive Ränge gebildet. Je größer die Differenzen der Rangsummen, desto eher wird der Test signifikant. Die Differenzbildung im vorliegenden Beispiel erbringt für jeden Probanden einen negativen Wert resp. negativen Rang (max. 17,07-100<0 und min. 17,07-29,76<0). Die Rangsumme der positiven Ränge ist Null. Die Rangsummendifferenz erzielt damit ihren maximal erzielbaren Wert und wird signifikant (Bühner/Ziegler 2008, S. 267 ff.).

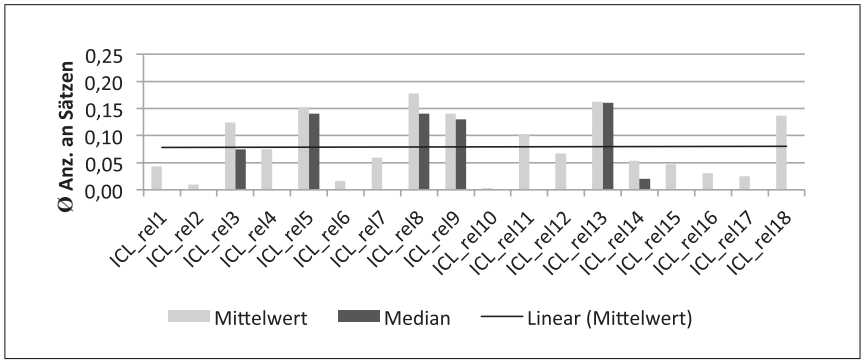


Abb. 3: Mittelwerte und Mediane des empirischen Loads (ICL_emp,rel) pro Aufgabe

Aufgaben klassifiziert, die drei Aufgaben 5, 8 und 14 als mittelschwer, sodass ein Median größer Null erwartungskonform ist. Die Mittelwerte geben zusätzlich Information über die empirisch gezeigte Schwierigkeit jeder Aufgabe. Hierbei fällt auf, dass Aufgabe 8 den höchsten Mittelwert erreicht, was nicht erwartungskonform ist.

Abbildung 4 zeigt die Mittelwerte der Lösungshäufigkeiten pro Aufgabe und deren Trendlinie. Die Lösungshäufigkeiten sind dabei in umgekehrter Reihenfolge abgetragen. Damit ist die Grafik wie folgt zu lesen: Mehr als 95 % der Probanden konnten die Aufgabe 6 lösen, sie ist damit – wie erwartet – die leichteste Aufgabe. Schwere Aufgaben sind nach dieser Grafik beispielsweise Aufgabe 3, 7 und 9.

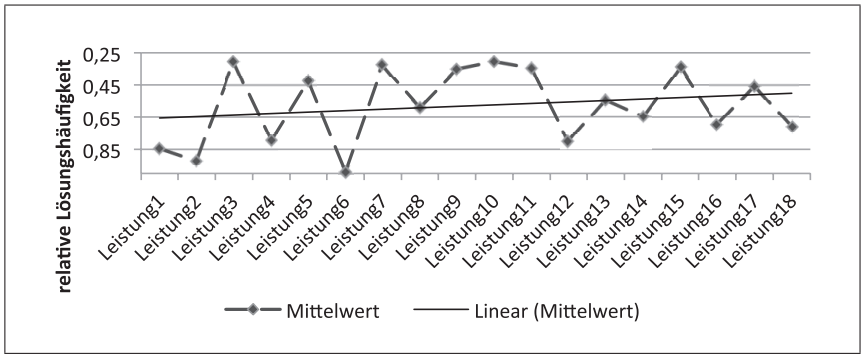


Abb. 4: Mittelwerte der Leistung (Leistung_rel) pro Aufgabe

Der theoretisch intendierte Schwierigkeitsverlauf dürfte über den Aufgabenfortschritt hinweg leicht steigend sein (vgl. *ICL_theo,rel* in Tab. 2). Die Trendlinie in Abb. 4 zeigt einen erwartungsgemäß steigenden Verlauf. Die Trendlinie der intrinsischen Belastungsäußerungen (Abb. 3) ist jedoch relativ konstant. Das könnte heißen, dass die Belastungsäußerungen (*ICL_emp,rel*) der vorderen Aufgaben höher sind als erwartet oder die der hinteren Aufgaben niedriger als erwartet. Im Folgenden wird gezeigt, wie sich – grafisch veranschaulicht – der *ICL_theo,rel* einer jeden Aufgabe zu den empirischen Befunden verhält (Abb. 5). In Abb. 5a sind die Mittelwerte der

empirisch erhobenen Belastungen dem relativen theoretischen ICL gegenübergestellt. Dabei zeigt sich, dass für Aufgabe 5, 8 und 11 der empirisch geäußerte Load den theoretisch klassifizierten augenscheinlich *deutlich* übersteigt. Die Aufgaben 7, 15, 16 und 17 zeigen hingegen *deutlich* geringere durchschnittlich geäußerte empirische Loads als ursprünglich erwartet.

Als weitere Validierungsmöglichkeit werden die tatsächlichen Leistungen der Probanden herangezogen (Abb. 5b). Nach dieser Grafik wiederholt sich der Effekt der Überschätzung für Aufgabe 5 und 11. Aufgabe 8 ist sowohl theoretisch als auch mit Blick auf die Leistungen mittelschwer. Eine große Abweichung (Überschätzung) ergibt sich zusätzlich für Aufgabe 10. Deutlich unterschätzt wurde nach diesem Vergleich die Aufgabe 13.

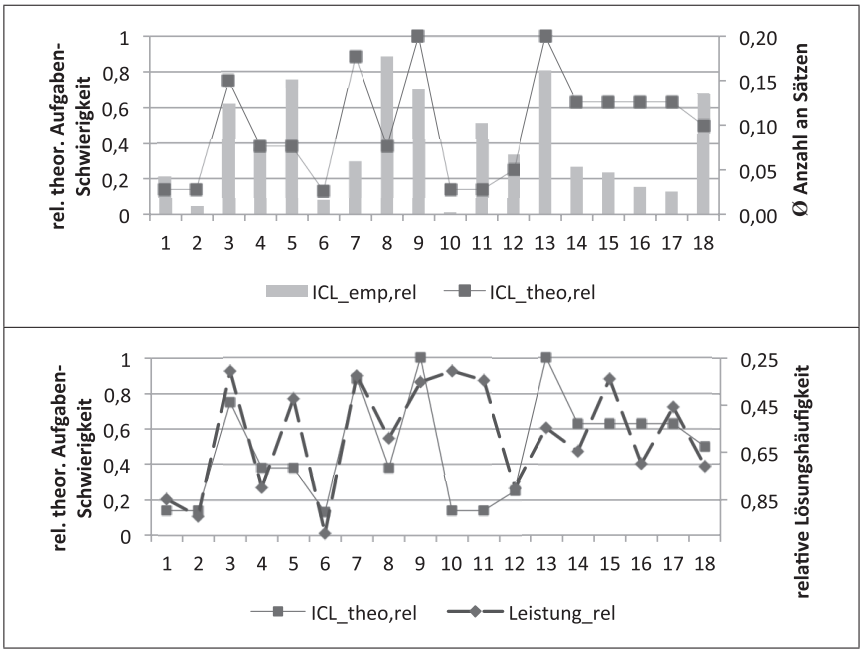


Abb. 5: grafisch komparative Analyse des theoretischen ICL sowie (5a) des empirischen Mittelwerts des ICL bzw. (5b) des Mittelwerts der Leistungen

Die Ergebnisse der Forschungsfrage 1.2 lassen sich wie folgt zusammenfassen: Die Aufgaben unterscheiden sich nach augenscheinlichen Analysen in ihrer Validität. Acht Aufgaben zeigen keine deutlichen Abweichungen. Zwei Aufgaben (Aufgabe 5 und 11) zeigen in beiden Validierungsprüfungen Auffälligkeiten, indem sie scheinbar für die Probanden deutlich schwerer sind als dies theoretisch erwartet wurde.

5.2 Ergebnisse der qualitativen Analyse

Forschungsfrage 2: Welche Erklärungsmuster lassen sich mittels vertiefender qualitativer Analysen für die unter FF 1 identifizierten Abweichungen finden?

Mittels qualitativer Analysen sollen Merkmale und Kriterien identifiziert werden, die zu den unter Forschungsfrage 1 aufgedeckten Abweichungen geführt haben. Dafür werden in einer dritten Stufe des Kodierprozesses die belastungsrelevanten ECL- und ICL-Äußerungen zusätzlich detaillierter nach den theoretisch bestimmten Kategorien (Tab. 1 und 2) kodiert. Für nicht zuordnungsfähige Sätze werden neue Kategorien generiert. Kodierregeln und Beispiele sind in den Tabellen 5 und 7 zusammengefasst.

Tab. 5: Kodierregeln und Kodierbeispiele für ECL-Kategorien

ECL-Kategorie	**Kodierregel** Probleme mit …	**Kodierbeispiel**
Kontiguität: Navigation (1)	… Funktionalitäten der Plattform	A1, LD16: „Wie soll ich die Aufgabe beenden? und jetzt?“
Kontiguität: technisches Problem (1)	… Features, die nicht oder anders als intendiert funktionieren	A14, LD23: „Okay. Einfach durchstreichen oder was? Äh, kann ich den Stift nicht auswählen?“
Signalisierung (2)	… farblichen Markierungen	A2, LD14: „Hat das jetzt eine Bedeutung, warum das so rot hinterlegt ist?“
Aufgabenanforderung unklar/ unpassend (NEU)	… nicht intendierter Aufgaben-Anforderung	A7, LD16: „Wieso kann ich da nicht mehr weiterschreiben?“
zeitliche Orientierung (NEU)	… fehlender zeitlicher Orientierungsmöglichkeit	A7, LD 24: „Wo seh ich eigentlich, wie weit ich schon bin, gibt’s da überhaupt was?“
Lautes-Denken-Methode (NEU)[1]	… Aufgabenbearbeitung unter Beobachtung	A3, LD 12: „Ich find das relativ schwierig, wenn mir immer jemand über die Schulter schaut.“

Anmerkungen: [1] = ECL-Kategorie nur in 0,02 % aller Sätze kodiert; A1, LD16 steht für Aufgabe 1 und Lautes-Denken-Proband Nummer 16; für folgende Kategorien gab es keine Kodierungen: Vermeidung von Redundanzen; Modalität; Reduzierung sprachlicher Komplexität; ästhetische/professionelle Aufbereitung; Strukturierung/unterstützende Information durch Sequenzierung

Forschungsfrage 2.1: Welche ECL-relevanten Erklärungsmuster lassen sich finden?

Ergebnisse zur Forschungsfrage 1.1 ergaben, dass für die Aufgaben 3, 5 und 13 der Median (für *ECL_emp,rel*) signifikant von Null verschieden ist. Des Weiteren sind – aufgrund ihrer relativ hohen Mittelwerte – die Aufgaben zu Beginn der Plattform (1 und 2) auffällig. Tab. 6 fasst die zentralen Erklärungen zusammen.

Zu Beginn der Plattform (*Aufgabe 1 und 2*) ergeben sich vermehrt Schwierigkeiten mit der Benutzerführung sowie vereinzelten Funktionalitäten (Navigation/ Kontiguität). Die Probanden suchen die notwendigen Programme und Dokumente. Es ist nicht immer klar, wann eine Aufgabe fertig bearbeitet bzw. wie genau diese zu beenden ist. Hinzu kommen missverständliche Aufgabenprogrammierungen,

Tab. 6: Qualitative ECL-Erklärungsmomente für auffällige Aufgaben

auffällige ECL- Aufgaben	Qualitative Erklärungsmomente
1, 2	• Navigation (1) • Signalisierung (2): missverständliche Programmierung
3	• Navigation (1) – große Tabellen auf kleinem Bildschirm • technisches Problem (1): mit dem Tabellenkalkulationsprogramm
5	• technisches Problem (1): mit dem Tabellenkalkulationsprogramm • Signalisierung (2): missverständliche Farbwahl
13	• technisches Problem (1): Aufgabenformat • Aufgabenstellung unpassend: Zeichenbegrenzung im offenen Feld (NEU) • zeitliche Orientierung (NEU)

die zu einer ungewollten Aufmerksamkeitsfokussierung führen: „*Warum ist das rot unterlegt?*“ (A1, LD05). Zur Lösung der *Aufgabe 3* und *Aufgabe 5* ist die Anwendung des Tabellenkalkulationsprogramms notwendig. Da dieses – im ersten Entwurf der technologiebasierten Aufgabenkonstruktion – noch unerwünschte und unerwartete technische Probleme (z. B. fehlerhafte Funktion des Makierbuttons) aufwies, liegen darin auch die zentralen extrinsischen Belastungen hinsichtlich der beiden Aufgaben (technisches Problem/Kontiguität). Des Weiteren rufen die intrinsisch intendierten komplexen (viel Raum einnehmenden) Tabellen extrinsische Belastungen hervor, weil das Scrollen nach oben/unten bzw. rechts/links ungewollt kognitive Kapazitäten beansprucht (Navigation/Kontiguität). Bei *Aufgabe 5* kam hinzu, dass einige Probanden von der genutzten Farbwahl irritiert waren (Signalisierung). *Aufgabe 13* umfasst zwei Aufgabenteile. Im ersten Teil (Zuordnungsaufgabe) geht es darum, Maßnahmen und Folgen einer betriebswirtschaftlichen Entscheidung miteinander in Verbindung zu bringen. Im zweiten Teil erfolgt die begründete Auswahl für die unternehmensseitig beste Maßnahmen-Folgen-Entscheidung. Wiederum sind es technische Schwierigkeiten bezüglich der einwandfreien Anwendung von Zuordnungsoptionen (technisches Problem). Eine ungewollte Hürde stellt die Zeichenbegrenzung (um den Auswertungsumfang zu begrenzen, umfasst das Begründungsfeld max. 160 Zeichen) im offenen Teil der Aufgabe dar. Die präzise Formulierung einer Begründung fällt den Probanden jedoch schwerer als intendiert (unpassende Aufgabenstellung). Aus dem schulischen Kontext sind es Prüflinge gewohnt, eine zeitliche Orientierung zu haben, was im ersten Entwurf der Plattform noch fehlt. Insgesamt lässt sich anmerken, dass den Äußerungen zufolge die Strategien *Vermeidung von Redundanzen; Modalität; Reduzierung sprachlicher Komplexität; ästhetische/professionelle Aufbereitung* sowie *Strukturierung/unterstützende Information durch Sequenzierung* bereits gut umgesetzt wurden, da hierzu keine Äußerungen kodiert wurden.

Forschungsfrage 2.2: Welche ICL-relevanten Erklärungsmuster lassen sich finden?

Auch die ICL-Belastungen wurden hinsichtlich der theoretisch erarbeiteten Kategorien kodiert (Tab. 7). Dabei zeigte sich, dass ein Großteil der Sätze der Kategorie *Tiefgang und inhaltliche Komplexität* zuzuordnen war. Alle anderen Kategorien wurden deutlich seltener genannt; der Kategorie *Unterstützungsumfang* wurde kein

Satz zugeordnet. Daher ist es schwer möglich, die Erklärungsmuster der auffälligen Aufgaben in den Abweichungen der theoretisch intendierten und empirisch identifizierten ICL-Kriterien zu suchen. Es sei jedoch angemerkt, dass von den empirischen Quantitäten der ICL-Kriterien nicht auf deren Bedeutung geschlossen werden sollte. Das heißt, dass der Unterstützungsumfang auch ohne Verbalisierung durch den Probanden Einfluss auf die intrinsisch aktivierte Kapazität haben kann.

Tab. 7: Kodierregeln und Kodierbeispiele für ICL-Kategorien

ICL-Kategorie	**Kodierregel** Probleme mit …	**Kodierbeispiel**
Redundanz	… intendierten überflüssigen Informationen	A9, LD05: „Für was brauch ich denn den Kostenplan? Das sind meine gesamten Kosten am Anfang 200.000. Okay. (.) Hmm.“
Elementinteraktivität	… Seitenwechseln	A3, LD06: „Hä? Ich weiß nicht, aber ich glaub ich brauch diesen blöden Anhang. Ah ja, da steht Dokumente, ach Mist. Nein, da kann ich nicht. Oh Gott was hab ich gemacht? Was ist das hier? (.) Ah, okay. Okay. Gut. (..) Also, da ist das Protokoll vom Onlineshop.“
Aufgabenkomplexität	… der Vernetztheit der Aufgabe	A9, LD03: „Also, es ist auf jeden Fall nicht möglich, die Investitionskosten innerhalb der 2,5 Jahre zu decken, offensichtlicher Weise. So viel weiß ich nach dieser langen Zeit jetzt schon mal. Ich mach das jetzt einfach so weiter.“
Vertrautheit der auszuführenden Handlung	… der Vertrautheit mit dem Aufgabenformat	A5, LD06: „Hmm ach, also muss ich da einfach ein x rein machen oder muss ich da was rein schreiben?“
Tiefgang/ inhaltliche Komplexität	… inhaltlich intendierter Anforderung/ Komplexität (u. a. wiederholtes Lesen der Aufgabenstellung)	A3, LD13: „Dokumente, nein, hmm, okay. Erstellen Sie in der Vorlage unter Tab. Onlineshop aus den Angaben des Protokolls einen Gantt-Plan für das Projekt Onlineshop. (.) Was ist ein Gantt-Plan? Hmm.“

Anmerkungen: A9, LD05 steht für Aufgabe 9 und Lautes-Denken-Proband Nummer 5; (.) = kurze Pause; für folgende Kategorie gab es keine Kodierungen: Unterstützungsumfang

Tab. 8 diskutiert mögliche Erklärungsmuster für Aufgaben, die von den theoretisch intendierten ICL-Erwartungen abweichen. Die theoretische Überschätzung der Schwierigkeit (Aufgabe 7, 13, 14, 15, 16) ist zwar ein wichtiger Hinweis, dürfte aber selten mit ernsthaften Validitätseinbußen einhergehen. Bei den Aufgaben 14, 15 und 16 beispielsweise entspricht die Leistung der Probanden annähernd den intendierten Schwierigkeiten. Die geringe Anzahl an ICL-Sätzen könnte sich daher hypothetisch auf eine abnehmende Bereitschaft laut zu denken oder auf ein weniger aktivierendes Aufgabenformat zurückführen lassen.

Ernstzunehmende Validitätsschwächen weisen diejenigen Aufgaben auf, deren Anspruchsniveau unterschätzt wurde (Aufgabe 5, 8, 10, 11), da dies mit ungewollten oder unterschätzen Schwierigkeiten (z. B. Anregung von bei der Aufgabenkonstruktion nicht in Betracht gezogenen mentalen Modellen, zielgruppeninadäquates

Tab. 8: qualitative ICL-Erklärungsmuster für auffällige Aufgaben

ICL Aufgabe	Auffälligkeit	Qualitative Erklärungsmuster
5	ICL_theo < ICL_emp ICL_theo < Leistung	regt andere mentale Modelle an als intendiert
8	ICL_theo < ICL_emp	Kombination aus inhaltlicher und sprachlicher Komplexität
10	ICL_theo < Leistung	Fehlende Eindeutigkeit der inhaltlichen Distraktoren
11	ICL_theo < ICL_emp ICL_theo < Leistung	Reaktion auf nicht-intendierte Signalwörter Fehlende Eindeutigkeit der inhaltlichen Distraktoren

Anmerkung: Die theoretisch überschätzten Aufgaben 7, 13, 14, 15, 16 sind nicht aufgeführt, da sich hier mangels fehlender ICL-Sätze keine Erklärungsmuster ableiten lassen.

Anspruchsniveau) einhergehen könnte. Die Aufgaben 5 und 11 sind in beiden Validitätsprüfungen auffällig geworden (Abb. 5) und sollen daher für die Diskussion exemplarisch herangezogen werden. In der *Aufgabe 5* geht es um die Bewertung relevanter bzw. irrelevanter Kostenpositionen zur Implementierung eines Onlineshops. Es zeigt sich, dass es den Auszubildenden schwer fällt, sich in das konkrete Vorhaben hineinzuversetzen und zu überlegen, dass bspw. die Lagerhaltung der über den Onlineshop zu verkaufenden Produkte für die Implementierung des Onlineshops keine relevante Kostenposition darstellt: *„Was sind denn überhaupt nicht anfallende Kosten? Warum fallen denn die nicht an? Und woher und wie soll ich denn das jetzt auswählen?“* (LD04). Andere wiederum fragen sich, ob eine genannte Position tatsächlich Kosten verursacht: *„Optimierung der Logistik ist nicht anfallende Kosten, puh. Wieso ist das nicht anfallend?“* (LD05). In *Aufgabe 11* soll unabhängig vom Inhalt der Projektarbeit gutes/schlechtes Teamverhalten bewertet werden. Die qualitative Analyse zeigt jedoch, dass die Probanden Schwierigkeiten haben, losgelöst vom Inhalt die Bewertung vorzunehmen *„Puh, das ist jetzt die Frage, was das für ein Lösungsvorschlag ist, oder?“* (LD01).

6. Diskussion, Limitationen und Ausblick

Anliegen der Studie war es, im Sinne des integrativen Validitätsansatzes von Messick eine weitere empirische Evidenz für die Angemessenheit der Messung von Intrapreneurship-Kompetenz mittels der ALUSIM-Testplattform zu generieren. Während die Studien von Weber und Kollegen (2014) insbesondere die *inhaltliche Repräsentativität* und die von Bley (im Erscheinen) die *kognitive Validität* (Adäquanz der Lösungsprozesse) betrachten, verfolgte die vorliegende Studie das Ziel, die *Validität der technologiebasierten Testplattform/-aufgaben* zu prüfen. Theoretische Grundlage bildeten die Cognitive Load Theory sowie die Fachdidaktik. Zunächst wurden Kriterien kognitiver Belastung (unerwünscht: ECL und erwünscht: ICL) theoretisch hergeleitet sowie die intendierten kognitiven Belastungen auf Basis der einzelnen Aufgaben spezifiziert. Die Prüfung erfolgte im Mixed Method Design auf Basis verbaler Transkripte aus Studien Lauten Denkens. Ziel der Aufgabenkonstruktion war es, unerwünschte kognitive Belastung (ECL) nach Möglichkeit

vollständig zu vermeiden, sodass Testergebnisse ausschließlich auf intendierte Aufgabenschwierigkeiten (ICL) der Aufgaben zurückzuführen sind.

Überarbeitungen und Handlungsempfehlungen zum ECL

Eine vollständige Reduktion des ECL wurde – in der ersten Version der Aufgabenprogrammierung – nicht erreicht. Insbesondere zu Beginn der Plattform und bei vereinzelten Aufgaben zeigten sich vermehrt unerwünscht kognitive Belastungen. Qualitative Analysen erbrachten, dass sich diese Abweichungen zumeist auf technische Fehlfunktionen insbesondere in den Endbenutzerprogrammen sowie Probleme bei der Benutzerführung zurückführen lassen. Weitere – allerdings nur vereinzelt aufgetretene – Belastungen ergaben sich durch irritierende Farbmuster, unpassende Aufgabenanforderungen (Begrenzung von Begründungsfeldern), fehlende zeitliche Orientierungsmöglichkeiten sowie die Erhebungsmethode des lauten Denkens (bei zwei Probanden je einmal genannt). Diese Erkenntnisse veranlassten uns zu folgenden Überarbeitungen der ALUSIM-Plattform: (1) Anpassung des Tutorials zum besseren Verständnis der Funktionalität der Aufgaben und der Benutzerführung innerhalb der Testplattform (z. B. der explizite Hinweis auf den Button „Aufgabe beenden“), (2) Ausräumen technischer Probleme (z. B. das fehlerfreie Rückgängigmachen farblicher Zellmarkierungen in der Tabellenkalkulation), (3) Überarbeitung von darstellungsspezifischen Aufgabenmissverständnissen (z. B. die Erhöhung der Zeichenanzahl in Begründungsfeldern von 160 auf 320 Zeichen), (4) Überarbeitung von missverständlichen Signalisierungen (z. B. intuitivere Farbwahl), (5) die Einführung eines Zeitbalkens zur zeitlichen Orientierung hinsichtlich des Aufgabenfortschritts im Vergleich zur noch verbleibenden Zeit (inkl. eines Pause-/Stopp-Buttons, um Unterbrechungen zu ermöglichen) und (6) Rotation der Aufgabenszenarien, um Ermüdungseffekte kontrollieren zu können.

Überarbeitungen und Handlungsempfehlungen zum ICL

Der ICL variiert auf Basis unterschiedlicher Kombinationen von berücksichtigten ICL-Kriterien in seiner Höhe über die Aufgaben hinweg. Zur Prüfung dessen wird neben dem empirisch erhobenen ICL auch die tatsächlich erbrachte Leistung der Probanden berücksichtigt. Zwei Aufgaben zeigen in beiden Prüfungen eine deutliche Unterschätzung der Aufgabenschwierigkeit (Aufgabe 5 und 11). Auf Basis der qualitativen Analyse kann geschlussfolgert werden, dass in beiden Aufgaben missverständliche Signale „falsche“ mentale Modelle aktivieren. Dies führt zu intrinsisch kognitiven Belastungen, wie sie nicht intendiert waren, und zu insgesamt geringeren Lösungswahrscheinlichkeiten. Diese sowie die anderen abweichenden Aufgaben wurden unter Vorlage der Erkenntnisse erneut mit Vertretern der Praxis diskutiert und überarbeitet.

Limitationen der Studie und Ausblick

Die geringen Ausprägungen an ICL trotz erwartungskonformen Lösungswahrscheinlichkeiten bei den Aufgaben 15, 16 und 17 könnten in der Erhebungsmethode des Lauten Denkens begründet sein. Auch wenn sie die zurzeit beste Methode zur Abbildung des Lösungsprozesses ist, ist sie keinesfalls frei von Fehlern. Die Erhebungsmethode ist zudem verantwortlich für die kleine Stichprobe, welche limitierende Wirkung für die Aussagekraft der quantitativen Analysen ausübt. Des Weiteren muss erwähnt werden, dass die quantifizierten Daten extrem linkssteile

Verteilungen besitzen (große Häufigkeit der Ausprägung „0“) und diese eine quantitativ vergleichende Analyse unmöglich gemacht haben. Der grafische Vergleich konnte lediglich augenscheinliche Abweichungen hervorbringen. Hier fokussierten wir uns auf deutliche Auffälligkeiten. Offen bleibt, ob bei vereinzelten Aufgaben die Existenz von ECL die Lösungswahrscheinlichkeit einer Aufgabe beeinflusst und welche der theoretisch intendierten ICL-Kriterien signifikanten Einfluss auf gewünschte intrinsische bzw. unerwünscht extrinsische kognitive Belastungen nehmen. Die identifizierten Kriterien kognitiven Loads sollten in zukünftigen Studien in experimentellen Designs systematisch manipuliert werden, um belastbare Erkenntnisse für die Konstruktion technologiebasierter Aufgaben zu gewinnen.

Für die Entwicklung technologiebasierter authentischer Testaufgaben zur validen Messung beruflicher Kompetenzen konnte die Studie hilfreiche Erkenntnisse in Form von ICL- und ECL-Kriterien zu Tage fördern. Damit ist die Studie einmal mehr Evidenz dafür, dass Assessments in Zeiten von Accountability neu gedacht werden sollten.

Literatur

Achtenhagen, F./Winther, E. (2009): Konstruktvalidität von Simulationsaufgaben: Computergestützte Messung berufsfachlicher Kompetenz – am Beispiel der Ausbildung von Industriekaufleuten. Seminar für Wirtschaftspädagogik der Georg-August-Universität Göttingen.

American Educational Research Association/American Psychological Association/National Council on Measurement in Education (2014): Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Baethge, M./Achtenhagen, F./Babic, E./Baethge-Kinsky, V./Weber, S. (2006): PISA-VET – A feasibility study. Stuttgart: Steiner.

Bazeley, P./Richards, L. (2000): The NVivo qualitative project book. London, Thousand Oaks, CA: SAGE Publications.

Bley, S. (im Erscheinen): Cognitive processes in use when solving intrapreneurship activities – A formative assessment perspective. Institute for Human Resource Education and Management, Munich.

Blömeke, S. (2003): Lehren und Lernen mit neuen Medien – Forschungsstand und Forschungsperspektiven. In: Unterrichtswissenschaft 31, H. 1, S. 57–82.

Brünken, R./Plass, J./Leutner, D. (2003): Direct measurement of cognitive load in multimedia learning. In: Educational Psychologist 38, H. 1, S. 53–61.

Bühner, M./Ziegler, M. (2008): Statistik für Psychologen und Sozialwissenschaftler. München [u. a.]: Pearson.

Cawthon, S. W./Kaye, A. D./Lockhart, L. S./Beretvas, N. (2012): Effects of linguistic complexity and accommodations on estimates of ability for students with learning disabilities. In: Journal of School Psychology 30, S. 293–316.

Chi, M. T. H. (1997): Quantifying qualitative analyses of verbal data: A practical guide. In: Journal of the Learning Sciences 6, H. 3, S. 271–315.

Clark, R./Nguyen, F./Sweller, J. (2005): Efficiency in learning: Evidence-based guidelines to manage cognitive load. San Francisco: Pfeiffer.

Dresing, T./Pehl, T. (^{2}2011): Praxisbuch Transkription. Regelsysteme, Software und praktische Anleitungen für qualitative ForscherInnen. Marburg: Eigenverlag.

Ericsson, K. A./Simon, H. A. (^{2}1996): Protocol analysis. Verbal reports as data: revised edition. Cambridge, Massachusetts, London: The MIT Press.

Faul, F./Erdfelder, E./Buchner, A./Lang, A.-G. (2009): Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. In: Behavior research methods 41, S. 1149–1160.

Gillmor, S. C./Poggio, J./Embretson, S. (2014): Effects of reducing the cognitive load of mathematics items on student performance. Vortrag auf dem 2014 Annual Meeting der American Educational Research Association (AERA) (03.–07. April 2014 in Philadelphia, USA).

Haladyna, T. M./Downing, S. M. (2004): Construct-irrelevant variance in highstakes testing. In: Educational Measurement: Issues and Practice 23, H. 1, S. 17–27.

Haladyna, T. M./Downing, S. M./Rodriguez, M. C. (2002): A review of multiple-choice item-writing guidelines for classroom assessment. In: Applied Measurement in Education 15, H. 3, S. 309–334.

Hartig, J./Klieme, E. (Hrsg.) (2007): Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik. Bonn, Berlin: Bundesministerium für Bildung und Forschung.

Jassmeier, A. (2005): Zum Verhältnis von Elektronischer Datenverarbeitung und Betriebswirtschaftslehre. Erfahrungen und Perspektiven in den Bereichen Notebook-Einsatz, Nutzung von ERP-Systemen sowie E-Learning. In: Zeitschrift für Berufs- und Wirtschaftspädagogik 101, H. 2, S. 246–271.

Jude, N./Wirth, J. (2007): Neue Chancen bei der technologiebasierten Erfassung von Kompetenzen. In: Hartig, J./Klieme, E. (Hrsg.): Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik. Bonn, Berlin: Bundesministerium für Bildung und Forschung, S. 49–56.

Kalyuga, S./Plass, J. L. (2009): Evaluating and managing cognitive load in games. In: Ferdig, R. E. (Hrsg.): Handbook of research on effective electronic gaming in education. Hershey, PA: Information Science Reference, S. 719–737.

Kettler, R. J./Rodriguez, M. C./Bolt, D. M./Elliott, S. N./Beddow, P. A./Kurz, A. (2011): Modified multiple-choice items for alternate assessments: Reliability, difficulty, and differential boost. In: Applied Measurement in Education 24, S. 210–234.

Leow, R. P./Morgan-Short, K. (2004): To think aloud or not to think aloud. The issue of reactivity in SLA research methodology. In: Studies in Second Language Acquisition 26, S. 35–57.

Leuders, T. (2014): Modellierungen mathematischer Kompetenzen – Kriterien für eine Validitätsprüfung aus fachdidaktischer Sicht. In: Journal für Mathematikdidaktik 35, H. 1, S. 7–48.

Martiniello, M. (2008): Language and the performance of english-language learners in math word problems. In: Harvard Educational Review 48, H. 2, S. 333–368.

Mayer, R. E. (2005): Cognitive theory of multimedia learning. In: Mayer, R. E. (Hrsg.): The Cambridge handbook of multimedia learning. Cambridge, New York: Cambridge University Press, S. 31–48.

Messick, S. ([3]1989): Validity. In: Linn, R. L. (Hrsg.): Educational measurement. New York: American Council on Education [u. a.], S. 13–103.

Messick, S. (1995): Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. In: American Psychologist 50, H. 9, S. 741–749.

Miller, C. (2011): Aesthetics and e-assessment: The interplay of emotional design and learner performance. In: Distance Education 32, H. 3, S. 301–330.

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

Pellegrino, J. W./DiBello, L. V./Brophy, S. P. (2014): The science and design of assessment in engineering education. In: Johri, A./Olds, B. M. (Hrsg.): Cambridge handbook of engineering education research. Cambridge: Cambridge University Press, S. 571–598.

Pressley, M./Afflerbach, P. (1995): Verbal protocols of reading. The nature of constructively responsive reading. Hillsdale, N.J: Erlbaum.

Ruf, M. (2006): Geschäftsprozessorientierung im Unterricht – der Einsatz integrierter Unternehmenssoftware als didaktische Herausforderung für die kaufmännsiche Berufsausbildung. In: Erziehungswissenschaft und Beruf, H. 3, S. 343–355.

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

Shaftel, J./Belton-Kocher, E./Glasnapp, D./Poggio, J. (2006): The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. In: Educational Assessment 11, H. 2, S. 105–112.

Sweller, J. (2004): Instructional design consequences of an analogy between evolution by natural selection and human cognitive architecture. In: Instructional Science 32, S. 9–32.

Sweller, J. (2009): Cognitive bases of human creativity. In: Educational Psychology Review 21, H. 1, S. 11–19.

Sweller, J./Chandler, P. (1994): Why some material is difficult to learn. In: Cognition and Instruction 12, H. 3, S. 185–233.

Sweller, J./van Merriënboer, J. J. G./Paas, F. G. W. C. (1998): Cognitive architecture and instructional design. In: Educational Psychology Review 10, H. 3, S. 251–296.

van Merriënboer, J. J. G./Kirschner, P. A. (^{2}2013): Ten steps to complex learning. A systematic approach to four-component instructional design. New York, London: Routledge.

van Merriënboer, J. J. G./Sweller, J. (2005): Cognitive load theory and complex learning: Recent developments and future directions. In: Educational Psychology Review 17, H. 2, S. 147–177.

Weber, S./Trost, S./Wiethe-Körprich, M./Weiss, C./Achtenhagen, F. (2014): Intrapreneur: An entrepreneur within a company – an approach on modeling and measuring intrapreneurship competence. In: Weber, S./Oser, F. K./Achtenhagen, F./Fretschner, M./Trost, S. (Hrsg.): Becoming an entrepreneur, S. 256–287.

Weber, S./Wiethe-Körprich, M./Bley, S./Weiss, C./Achtenhagen, F. (im Druck): Intrapreneurship-Verhalten an kaufmännischen Arbeitsplätzen – Analysen von Projektberichten. In: Empirische Pädagogik Sonderheft: Ökonomische Kompetenzen in Schule, Ausbildung und Hochschule 28, H. 1.

Winther, E. (2010): Kompetenzmessung in der beruflichen Bildung. Bielefeld: Bertelsmann.

Winther, E./Achtenhagen, F. (2009): Skalen und Stufen kaufmännischer Kompetenz. In: Zeitschrift für Berufs- und Wirtschaftspädagogik 105, H. 4, S. 521–556.

Wirtz, M./Caspar, F. (2007): Beurteilerübereinstimmung und Beurteilerreliabilität. Göttingen [u. a.]: Hogrefe.

Anschrift der Autoren: Dr. Sandra Bley, MBR, Mail: bley@bwl.lmu.de, Michaela Wiethe-Körprich, MBR, Mail: wiethe@bwl.lmu.de, Prof. Dr. Susanne Weber, Mail: susanne.weber@bwl.lmu.de, Ludwig-Maximilians-Universität München, Munich School of Management l Fakultät für Betriebswirtschaft, Institut für Wirtschaftspädagogik, Ludwigstraße 28/RG, D-80539 München
Korrespondenz an: bley@bwl.lmu.de