**Profil 6:**

**Digitale Festschrift**
**für EVELINE WUTTKE**

**Kristina KÖGLER, Andreas RAUSCH & Helmut NIEGEMANN**
(Universität Stuttgart, Universität Mannheim & Universität Frankfurt)

# Interpretierbarkeit von Logdaten in computerbasierten Kompetenztests mit großen Handlungsräumen

Online unter:
https://www.bwpat.de/profil6_wuttke/koegler_etal_profil6.pdf

in

***bwp@* Profil 6** | September 2020

**Berufliches Lehren und Lernen: Grundlagen, Schwerpunkte und Impulse wirtschaftspädagogischer Forschung**

Hrsg. v. **Karin Heinrichs, Kristina Kögler & Christin Siegfried**

www.bwpat.de | ISSN 1618-8543 | *bwp@* 2001–2020

**KRISTINA KÖGLER, ANDREAS RAUSCH & HELMUT NIEGEMANN**
(Universität Stuttgart, Universität Mannheim & Universität Frankfurt am Main)

# Interpretierbarkeit von Logdaten in computerbasierten Kompetenztests mit großen Handlungsräumen

## Abstract

In computerbasierten Kompetenztests bieten Logdaten die Möglichkeit, die Handlungsstrategien der Teilnehmenden im Bearbeitungsprozess zu rekonstruieren und zusätzlich zu den Handlungsprodukten als Basis für die Kompetenzattribution zu dienen. Die Validität der Kompetenzmessung aufgrund von Logdaten hängt dabei von der Eindeutigkeit der Handlungen und der Gültigkeit der daraus gezogenen Schlussfolgerungen ab. Erstere wird in besonderem Maße durch die Gestaltung der Itemformate und Funktionalitäten in der Testumgebung beeinflusst, letztere stehen wiederum in einem Zusammenhang mit der durch die Teilnehmenden wahrgenommenen Benutzerfreundlichkeit. Der Beitrag geht auf analytischem Wege der Frage nach, welchen Einfluss das Design computerbasierter Testumgebungen auf die Interpretierbarkeit von Logdaten und die Benutzerfreundlichkeit hat. Exemplarisch wird für ausgewählte Itemformate mit großem Handlungsraum diskutiert, inwieweit die resultierenden Logdaten valide Information für die Kompetenzmessung bieten. Je größer der Handlungsraum für die Teilnehmenden ist, desto umfangreicher und vielfältiger sind die resultierenden Datenmuster und korrespondierenden Interpretationsspielräume. Mit Blick auf die computerbasierte Messung domänenspezifischer beruflicher Kompetenzen zeichnet sich dabei ein Spannungsfeld aus externer und interner Validität ab: Denn die extern valide Messung beruflicher Handlungskompetenz erfordert einerseits Testitems und Funktionalitäten der Testumgebung, die ein hinreichendes Maß an Komplexität beinhalten, um den Anforderungsgehalt beruflicher Probleme authentisch abzubilden. Andererseits erschweren Umfang und Vielfalt der resultierenden Aktivitätsmuster eine intern valide Interpretation und Kompetenzattribution. Das Design der Itemformate und Funktionalitäten in der Testumgebung spielt daher für die Analyse von Logdaten eine entscheidende Rolle.

***Schlüsselwörter**: Logdatenanalyse, Problemlösen, Kompetenzdiagnostik, Validität, Benutzerfreundlichkeit*

## 1 Einleitung

Computerbasierte Assessments mit komplexen Problemen und großen Handlungsräumen eröffnen Möglichkeiten, auf Basis von Logdaten auch prozessbezogene Kompetenzen wie den Einsatz metakognitiver Strategien zu analysieren. Allerdings sind die Prozessdaten aufgrund der Vielzahl möglicher Lösungswege der Teilnehmenden nicht immer eindeutig interpretierbar. Die gültige Interpretation der Logdaten ist insbesondere von der Wahl des Itemformats und der Gestaltung der Funktionalitäten in der Testumgebung beeinflusst. Der vorliegende Beitrag

thematisiert das resultierende Spannungsverhältnis zwischen interner und externer Validität sowie der Nutzerfreundlichkeit der Testumgebung.

Kompetenzdiagnostik mittels computerbasierter Testumgebungen hat sich im Rahmen von Large-Scale Assessments zum Standard entwickelt, wie bspw. die Erhebungen im Rahmen von PISA (Program for International Student Assessment) oder PIAAC (Program for the International Assessment of Adult Competences) zeigen. In der beruflichen Bildung sind insbesondere im Rahmen der ASCOT-Initiativen des Bundesministeriums für Bildung und Forschung (BMBF) computerbasierte Testumgebungen entwickelt worden (siehe www.ascot-vet.net). Bei der Messung beruflicher Kompetenzen wird versucht, den zunehmend komplexen Arbeitsanforderungen durch authentische und komplexe Testumgebungen gerecht zu werden, um damit die externe (ökologische) Validität sicherzustellen. Die theoretische Modellierung von Kompetenz basiert dabei meist auf der weit verbreiteten Definition von Weinert (2001). Kompetenzen werden als individuelle Disposition beschrieben, die in variablen Anforderungssituationen die erfolgreiche Bewältigung komplexer Probleme ermöglicht und sowohl kognitive, metakognitive wie auch emotional-motivationale Facetten umfasst. Die Attribution von Kompetenz erfolgt im Rahmen der Kompetenzdiagnostik gewöhnlich auf Basis von Handlungsprodukten der Teilnehmenden, also deren Lösungsvorschlägen zu den (mehr oder weniger) authentischen und komplexen Problemstellungen. Aus analytischer Sicht handelt es sich dabei um die Variablenwerte nach Abschluss der Bearbeitung, die im Fall qualitativer Daten (z. B. verfasster Texte) dann meist noch manuell zu kodieren sind. Computerbasierte Testumgebungen können aber i. d. R. auch alle Nutzereingaben (Mausklicks und Tastatureingaben) während der Bearbeitung inklusive eines Zeitstempels speichern. Diese Daten werden als Logdaten (engl. Log Data) bezeichnet, die in Logdateien (engl. Logfiles) für jede teilnehmende Person gespeichert werden. Logdaten ermöglichen die Rekonstruktion aller Nutzereingaben durch eine getestete Person in einer Testumgebung. In jüngerer Zeit werden Logdaten auch in der Berufsbildungsforschung für Prozessanalysen des Nutzerverhaltens genutzt, um unterschiedliche strategische Herangehensweisen zu identifizieren (z. B. Abele 2018; Abele et al. 2017; Rausch et al. 2017). Häufig werden dabei datengetriebene Vorgehensweisen eingesetzt, bei denen die Analyse nicht auf theoretischen Annahmen des Zusammenhangs zwischen Probandenverhalten und Testleistung basiert, sondern ausschließlich mittels komplexer statistischer Verfahren wie Mustersequenzanalysen, künstlichen neuronalen Netzwerken, Markov-Ketten oder n-grams erfolgt (z.B. He/von Davier 2016; Kinnebrew et al. 2014). Aus diesen explorativen Ansätzen resultieren Hypothesen bezüglich erfolgsrelevanter Handlungsschritte der Testteilnehmenden. Modellgetriebene Ansätze basieren dagegen auf einer der Analyse vorausgehenden theoretischen Modellierung der für die Aufgabenlösung relevanten Handlungsschritte und ihrer Sequenzierung im Sinne eines hypothesenprüfenden Vorgehens (Kinnebrew/ Segedy/Biswas 2018). Die spezifischen Herausforderungen der modellgetriebenen Analyse von Logdaten zur Messung metakognitiver Strategien in komplexen Kompetenztests nehmen wir vorliegend näher in den Blick und beleuchten dabei insbesondere die Rolle des Itemdesigns und die funktionale Gestaltung der Testumgebung.

Metakognitive Strategien beziehen sich auf die Planung, Überwachung, Regulation und Bewertung des eigenen Handlungsprozesses (Flavell 1979) und ermöglichen komplexe Hand-

lungssequenzen zum Lösen von Problemen (Betsch/Funke/Plessner 2011; Bransford/Stein 1993; Dörner 2000; Glaser 1994). In ihrem einflussreichen Bericht an das National Research Council fordern Pellegrino, Chudowsky und Glaser (2001), dass metakognitive Fähigkeiten daher auch in Assessments zu berücksichtigen seien. Auch in der deutschsprachigen Berufsbildungsforschung werden metakognitive Facetten oft als Bestandteile beruflicher Problemlösekompetenz modelliert (z.B. Baethge et al. 2006; Rausch/Wuttke 2016). Zur Messung metakognitiver Kompetenzen wurden jedoch lange Zeit nur fragebogenbasierte Selbsteinschätzungsverfahren verwendet, deren Validität diskutabel ist. Statt solcher retrospektiver Verfahren, empfehlen Veenman, Bavelaar, de Wolf und van Haaren (2014) prozessbasierte Verfahren zur Diagnostik von Metakognition, zu denen neben Beobachtung und lautem Denken auch die Analyse von Logdaten gehört.

Soll in einem computerbasierten Test neben der Bewertung der Problemlösungen (Produktperspektive) auch die Anwendung metakognitiver Strategien (Prozessperspektive) diagnostiziert werden, setzt dies voraus, dass komplexe Anforderungen und authentische Werkzeuge möglichst große Handlungsspielräume und damit mehrere unterschiedliche Lösungswege ermöglichen, die den Teilnehmenden entsprechende Handlungsentscheidungen abverlangen (Deutscher/Winther 2019; Frey et al. 2012; Gulikers et al. 2004; Rausch/Kögler 2016). Andernfalls wäre das Zielkonstrukt im Test unterrepräsentiert (Messick 1994, 14). Allerdings führen lange Bearbeitungszeiten, große Handlungsspielräume und ein umfangreiches Werkzeug-Repertoire zu entsprechend individuellen Verhaltensmustern, welche die Interpretierbarkeit der Logdaten erschweren und damit die interne Validität gefährden. Ein möglicher Ausweg ist die punktuelle Reduzierung von Funktionalitäten und Handlungsspielräumen, die jedoch unter Umständen zu Lasten der Benutzerfreundlichkeit der Testumgebung geht. Computerbasierte Testumgebungen sollten einfach und intuitiv zu bedienen sein, nicht vom Testinhalt ablenken und eine große Bandbreite von Strategien ermöglichen (Harms/Adams 2008). Bei eingeschränkter Benutzerfreundlichkeit und zu hoher kognitiver Belastung (Sweller/Ayres/Kalyuga 2011) ist mit konstrukt-irrelevanter Varianz zu rechnen. Erstere bezieht sich etwa auf den Umstand, dass diejenigen, die mit der Bedienung der Testumgebung besser zurechtkommen, höhere Kompetenzwerte erzielen, obwohl die Bedienung der Testumgebung nicht Bestandteil der zu messenden Kompetenz ist; und umgekehrt diejenigen, die eigentlich hohe Kompetenzen aufweisen, diese aufgrund der umständlichen Bedienung nicht umsetzen können. Es ergibt sich somit ein Spannungsverhältnis zwischen Konstruktunterrepräsentation und konstrukt-irrelevanter Varianz (Messick 1994). Im Rahmen des vom BMBF im Rahmen der ersten ASCOT-Initiative geförderten Projekts “Modellierung und Messung domänenspezifischer Problemlösekompetenz bei Industriekaufleuten (DomPL-IK)” war dieses in Abbildung 1 skizzierte Spannungsfeld Anlass umfangreicher Diskussionen im Zuge der Entwicklung der computerbasierten Testumgebung, an der nicht zuletzt auch Eveline Wuttke maßgeblich beteiligt war.

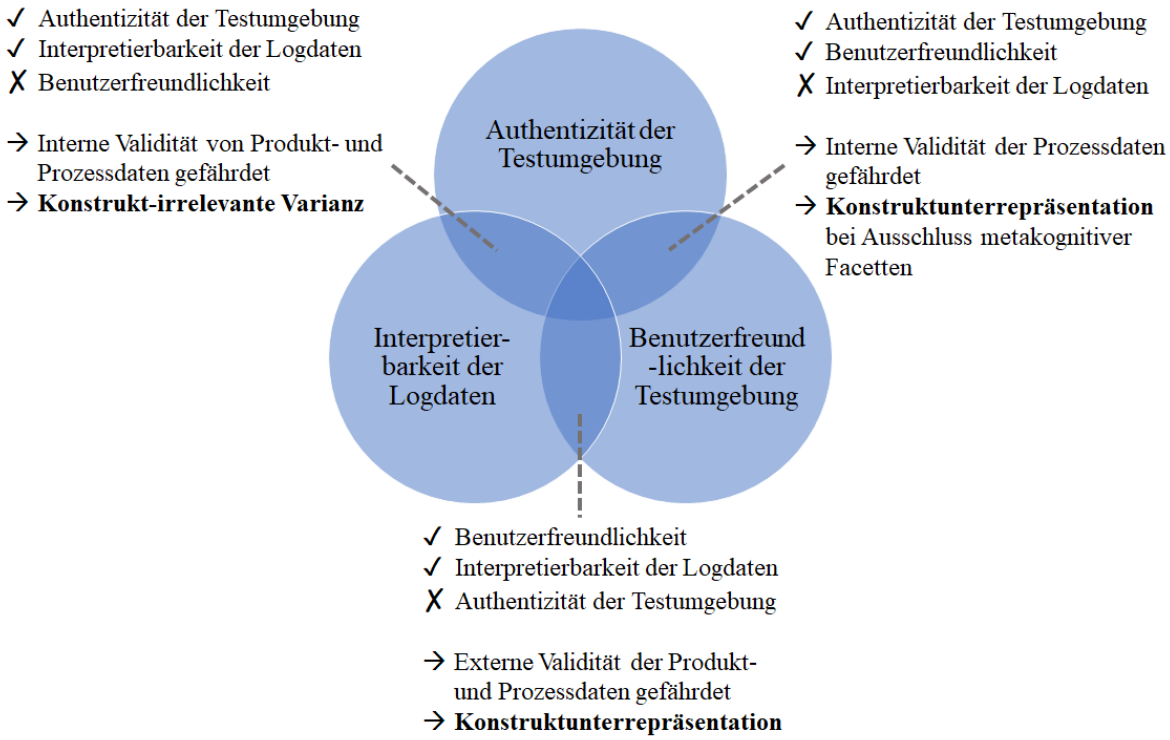


Abbildung 1: Spannungsverhältnis aus Authentizität der Testumgebung, Benutzerfreundlichkeit und Interpretierbarkeit der Logdaten

Die Zielsetzung des vorliegenden Beitrags besteht darin, das skizzierte Spannungsverhältnis anhand ausgewählter Design-Ansätze (Kapitel 2), möglicher Itemformate (Kapitel 3) und Gestaltungsoptionen für Testumgebungen mit offenen Handlungsräumen (Kapitel 4) exemplarisch zu diskutieren.

## 2 Messtheoretische Grundlagen einer prozessnahen Erfassung metakognitiver Kompetenzen

Die Qualität eines Tests wird in erster Linie anhand verschiedener Validitätskonstrukte diskutiert. Validität bezieht sich darauf, was ein Test tatsächlich misst und welche Schlussfolgerungen aus den Testergebnissen gezogen werden können (Cronbach 1971) bzw. wie angemessen diese Schlussfolgerungen angesichts der empirischen Befunde und theoretischen Überlegungen sind (Messick 1989). Das Konstrukt der Validität wurde von verschiedenen Forschenden unterschiedlich ausdifferenziert. Eine relativ grobe, aber hilfreiche Differenzierung ist die Unterscheidung zwischen interner und externer Validität, die auf Campbell (1957) zurückgeht. Interne Validität bezieht sich auf die Gültigkeit der Kausalinterpretationen innerhalb einer Studie und externe Validität fragt, inwiefern die Schlussfolgerungen aus einer Studie auch für vergleichbare Personen, vergleichbare Kontexte, vergleichbare Maßnahmen und vergleichbare Operationalisierungen gelten (Shadish/Cook/Campbell 2002, 38). Bezogen auf eine schriftliche Führerscheinprüfung bedeutet interne Validität zum Beispiel, dass die richtige oder falsche Antwort bei einer Frage allein durch das verfügbare Domänenwissen der Teilnehmenden verursacht wird. Externe Validität bezieht sich in diesem Beispiel auf die Frage, ob Personen mit vielen falschen Antworten inkompetente Verkehrsteilnehmende sind. Obwohl – oder

vielleicht gerade weil – dieses Beispiel uns an der Validität eines solchen Tests zweifeln lässt, ist Messick (1994, 13) zuzustimmen, dass es sich bei Testgütekriterien wie Validität nicht um abstrakte Messprinzipien, sondern um soziale Werte handelt, die auch außerhalb der akademischen Welt überall bedeutsam sind, wo Bewertungen vorgenommen und Entscheidungen getroffen werden. Dass man für den Führerscheinerwerb neben der theoretischen Prüfung auch praktische Fahrstunden und eine praktische Prüfung absolvieren muss, findet daher i. d. R. bevölkerungsweite Zustimmung. Wenn also für Tests Merkmale wie Authentizität und Handlungsbezug gefordert werden, damit man den Testergebnissen und deren Konsequenzen vertraut, dann handelt es sich bei diesen Merkmalen offensichtlich um Validitätskriterien (Messick, 1994, 14).

Für die Entwicklung valider Assessments empfehlen verschiedene Autorinnen und Autoren Modellierungsschritte, die Annahmen über das zu messende Konstrukt, die für die Messung geeigneten Aufgabenstellungen sowie die Verdichtung der Beobachtungen bis hin zu skalierbaren Variablen beinhalten (Arieli-Attali et al. 2019; Pellegrino/DiBello/Goldman 2016). Kane's (2013) Ansatz umfasst beispielsweise vier plausible Schritte: (1) *Scoring* (vom gezeigten Verhalten zu einem beobachteten Wert), (2) *Generalization* (vom beobachteten Wert zum verallgemeinerbaren Wert), (3) *Extrapolation* (vom verallgemeinerten Wert zum Kompetenzniveau), und (4) *Decision* (vom Kompetenzniveau zu einer Konsequenz). Noch umfassender ist der Ansatz des Evidence-Centered Design (ECD) der Arbeitsgruppe um Mislevy, da dieser zusätzlich auch die Entwicklung der Testitems und die Durchführung der Tests in den Blick nimmt (Mislevy et al. 2017, 20):

(1) *Domänenanalyse*: Sammlung aller relevanter Informationen über die betreffende Domäne wie bspw. Konzepte, Terminologien, Werkzeuge, Wissensrepräsentationen, Analysen, Anwendungsfälle und Interaktionsmuster

(2) *Domänenmodellierung*: Formulierung der für das Assessment relevanten Argumente wie bspw. Wissen, Fähigkeiten, Fertigkeiten, typische und variable Aufgabenmerkmale, mögliche Arbeitsprodukte und mögliche Beobachtungs- und Beurteilungsgrundlagen (auf Basis der Domänenanalyse)

(3) *Conceptual Assessment Framework (CAF)*: Strukturierte Darstellung auf Basis von Student model, Evidence model und Task model (s. u.) unter Verwendung von Design-Vorlagen (Principled Assessment Design for Inquiry; PADI), Kodieranleitungen, Messmodellen, Aufgabenspezifikationen, Durchführungsvorschriften usw.

(4) *Implementierung des Assessments*: Einsatzbereite und pilotierte Testmaterialien und Testumgebungen; auf Basis von Pilotstudien kalibrierte Messmodelle

(5) *Einsatz des Assessments*: Durchführung der Tests, Auswertungen, Ergebnisaufbereitung und Interpretation

Das Herzstück des ECD ist das unter (3) verortete *Conceptual Assessment Framework (CAF)*, welches das Zusammenspiel von *Student Model*, *Evidence Model* und *Task Model* aufzeigt. Das *Student Model* beinhaltet das zu messende Konstrukt in all seinen Facetten. Hier geht es also

um die Kompetenzmodellierung, d. h. die Definition der zu messenden Variablen und deren Strukturierung bspw. über mehrere Hierarchieebenen (Mislevy et al. 2003, 8). So könnte etwa Sozialkompetenz auf der oberen Ebene neben Fach- und Methodenkompetenzen eingereiht sein und weitere Subdimensionen wie Empathie, Ausdrucksvermögen und Verhandlungsgeschick aufweisen. Problemlösekompetenz werden üblicherweise die kognitiven Subdimensionen Informieren, Bearbeiten, Entscheiden und Kommunizieren zugeschrieben und daneben metakognitive Subdimensionen wie Planen, Überwachen und Handlungskontrolle sowie ausgehend von der Definition Weinerts (2001) auch emotional-motivationale Subdimensionen wie etwa Selbstkonzept und Interesse (vgl. Rausch & Wuttke 2016). Das *Task Model* beschreibt die Aufgaben, die Aufschluss über die zu messenden Kompetenzen bieten und das *Evidence Model* steht als Verbindungsglied zwischen *Student Model* und *Task Model*. Bei authentischen und damit fast immer komplexen Aufgabenstellungen ist hierbei eine (evtl. nicht nur) kognitive Aufgabenanalyse hilfreich (Schraagen, Chipman und Shalin 2000). Als Werkzeug kann beispielsweise eine Variante von *Business Process Model and Notation* (BPMN 2.0) verwendet werden (Schwantzer/Jannaber 2018), um den Prozess der Aufgabenbearbeitung seitens der Probanden mit dem als Kriterium gesetzten Verlauf zu vergleichen. Das *Evidence Model* besteht seinerseits aus zwei Komponenten: (1) Die *Evaluation Component* enthält ein Regelwerk (*Evidence Rules*) zur Extraktion von Variablen aus dem Verhalten der Testteilnehmenden bei einer bestimmten Aufgabe oder hinsichtlich eines bestimmten Aspekts (= *Task Scoring*). Im Fall eines Multiple-Choice-Tests ist das sehr einfach, denn jede Aufgabe führt i. d. R. zu einer Variablen, deren Wert im einfachsten Fall 0 = nicht korrekt gelöst oder 1 = korrekt gelöst ist. Die Definition von logdatenbasierten Variablen ist dagegen meist schwieriger. Eine einfache Variante bestünde bspw. darin, die Häufigkeit aller Interaktionen mit der Testumgebung in einer Variable Aktivitätsgrad zu speichern. (2) Die *Measurement Component* enthält ein Regelwerk (*Statistical Model*) zur Berechnung der Variablen im *Student Model* auf Basis der gemäß *Evidence Rules* extrahierten Variablenwerte (= *Test Scoring*; Mislevy et al. 2003, 8; Mislevy et al. 2010, 22). Eine einfache Variante ist die Bildung eines Summen-Scores aus allen Variablen eines Multiple-Choice-Tests. Inwiefern die Variable Aktivitätsgrad sinnvolle Interpretationen bzgl. metakognitiver Kompetenzen zulässt, scheint dagegen diskutabel.

Je nach *Student Model* und *Task Model* liegen die Schwierigkeiten der Testentwicklung in unterschiedlichen Bereichen des ECD. Bei kleinschrittigen geschlossenen Aufgabentypen zur Überprüfung von Faktenwissen liegt der Fokus meist auf dem *Statistical Model*, um die vielen Einzelinformationen eines Multiple-Choice-Tests widerspruchsfrei zu Kompetenz-Scores verdichten zu können. Hierzu dienen insbesondere Prüfverfahren auf Basis der Item Response Theory (IRT). Gute Modell-Fit-Indizes, überschneidungsfreie Item Characteristic Curves (ICC) statt Differential Item Functioning (DIF) usw. erlauben Aussagen über die Güte des *Statistical Models*. Die *Evidence Rules* bzw. das Scoring nach Kane (2013), also die Inferenz von der Beobachtung zum beobachteten Wert, stehen bei geschlossenen Testformaten (*Selected-Response Tests*) dagegen kaum im Fokus. Wie oben erläutert, ist hier kaum fraglich, welches Verhalten interessiert (nämlich die Auswahl der Antwort) und wie dieses gescort wird (Punktvergabe bei korrekter Antwort). *Performance Assessments* (*Constructed-Response Tests*; für einen Überblick siehe Lane/Stone 2006) stellen dagegen komplexere Anforderungen, bieten

größere Handlungsspielräume und verlangen während der längeren Bearbeitungszeit die Kombination mehrerer Handlungsschritte und eine entsprechende Selbstregulation, um eine Problemlösung zu erarbeiten. *Performance Assessments* erlauben daher auch die Beurteilung des Lösungsprozesses (Lane/Stone 2006, 387). Einer dieser Prozessaspekte ist die Anwendung metakognitiver Strategien. Den Schritten Kane's folgend (siehe auch Kane/ Crooks/Cohen, 1999) sind bei der logdatenbasierten Diagnostik metakognitiver Kompetenzen folgende Fragen zu beantworten:

(1) *Scoring*: Wie bilde ich auf Basis der Verhaltensdaten Variablen, welche das metakognitiv relevante Verhalten adäquat abbilden?

(2) *Generalization*: Inwiefern ist dieses Verhalten stabil über eine Gesamtheit ähnlicher Aufgabenstellungen (bspw. Cronbachs Alpha) und wie lässt es sich aggregieren (bspw. über einen einfachen Summen-Score oder auf Basis von IRT-Modellen)?

(3) *Extrapolation*: Inwiefern repräsentiert dieser generalisierte Wert das Zielkonstrukt des Kompetenzmodells?

(4) *Decision*: Welche Konsequenzen folgen aus den ermittelten Kompetenz-Scores?

Ein zentrales Attribut von Tests ist die inferenzielle Distanz, d. h. die Länge der oben skizzierten logischen Argumentationskette im Sinne notwendiger argumentativer Schritte von der Beobachtung bis zur Konsequenz (Behrens/DiCerbo/Foltz 2019, 220). Um diese inferenzielle Distanz möglichst gering zu halten, empfehlen Behrens et al. (2019), die Aufgabe, die Bearbeitungsbedingungen und erwarteten Leistungen in der Testsituation möglichst so zu gestalten wie in der Zieldomäne. Dies entspricht auch der Forderung nach Authentizität (s. o.). Dass hohe Authentizität die inferenzielle Distanz automatisch verringert, scheint aber für Prozessmerkmale wie die Messung metakognitiver Kompetenzen nicht in gleicher Weise zuzutreffen, da mentale Prozesse wie Planung, Strategieeinsatz, Überwachung und Handlungskontrolle schon in der Zieldomäne oft kaum eindeutig beobachtbar sind. Die umfangreichen Prozessdaten lassen sehr unterschiedliche Verarbeitungs- und Interpretationsmöglichkeiten zu. So berichten bereits Swanson et al. (1995, 8), dass offene Testformate zwar reichhaltige und aufschlussreiche Verhaltensmuster erzeugen, diese aber kaum vergleichbar und bewertbar seien. Wichtige Rahmenbedingungen der Interpretierbarkeit metakognitiver Kompetenzen auf Basis von Logdaten im Rahmen des *Evidence Models* werden insofern bei der Entwicklung komplexer Assessments im Rahmen des Itemdesigns (*Task Model*) geschaffen. Hier ist einerseits ein möglichst realitätsnaher Anforderungsgehalt der Aufgaben zu gewährleisten, andererseits sind die Interpretationsspielräume der Verhaltensweisen von Teilnehmen im Rahmen der Gestaltung der Testumgebung möglichst so zu begrenzen, dass bei der Logfileanalyse keine konstruktirrelevante Varianz entsteht und die Validität der Kompetenzzuschreibung gewährleistet ist. Nachfolgend werden daher sowohl die zu treffenden Entscheidungen als auch Differenzierungsmöglichkeiten für die Bewertung der Authentizität von Itemformaten beschrieben.

# 3 Handlungsspielraum und Itemkomplexität als zentrale Rahmenbedingungen der logdatenbasierten Analyse metakognitiver Strategien

Der Prozess des Itemdesigns im Rahmen des *Task Models* beinhaltet Entscheidungen bezüglich der Komplexität der zu bearbeitenden Aufgabenstellungen, korrespondierender Handlungs- und Lösungsräume bis hin zu den zur Verfügung stehenden Werkzeugen und Materialien. Dementsprechend identifizieren Behrens, DiCerbo und Ferrara (2012) in ihrem *Four-Spaces-Model* vier relevante Bereiche, die bei der Entwicklung von Assessments im Zuge des Itemdesigns zu berücksichtigen sind und Konsequenzen für die resultierenden Bearbeitungsprozesse durch die Teilnehmenden entfalten. Dabei spannen die Autoren ein Kontinuum von Items mit eng umgrenzten Handlungsmöglichkeiten bis hin zu Items, die viele Freiheitsgrade und einen entsprechend großen Handlungsraum bieten, auf:

- Der sog. *Problem Space* bezieht sich dabei auf die Komplexität des zu lösenden Problems bzw. des in der Aufgabenstellung zu erreichenden Ziels und die dabei zu bewältigenden Arbeitsschritte. Eng umrissene Aufgaben erfordern nur wenige Lösungsschritte, die zudem in ihrer Reihenfolge meist klar definiert sind, während in komplexeren Items mit größeren Problemräumen die notwendigen Bearbeitungsschritte nicht von vornherein klar definiert sind, sondern durch die Teilnehmenden erschlossen werden müssen.

- Im *Tool Space* geht es um die Mittel, Informationen und Materialien, mit denen die Aufgabenbearbeitung vollzogen wird. Je vielfältiger diese sind, desto unterschiedlicher stellen sich auch die möglichen Vorgehensweisen auf dem Weg zur Problemlösung bzw. Aufgabenbearbeitung dar.

- Der *Solution Space* beinhaltet die Anzahl der möglichen Aktivitäten bis zur Aufgaben- bzw. Problemlösung, die in komplexen offenen Assessments naturgemäß höher ist. Der Lösungsraum lässt sich über die Wahl des Lösungsformats und die zur Verfügung stehenden Arbeitsmaterialien gestalten.

- Im *Response Space* geht es schließlich um die Aktivitäten, die zur Erstellung der Lösungsprodukte notwendig sind. Das kann beispielsweise das Auswählen einer Multiple-Choice-Option (Freiheitsgrade in Abhängigkeit der Anzahl der Auswahloptionen) sein oder aber eine ausformulierte Lösungsskizze mit Erläuterungen (Freiheitsgrade in Abhängigkeit der Anzahl der Teilnehmenden).

In Fortführung dieses Gedankens legen DiCerbo und Behrens (2012, s. auch Shute et al. 2016) eine Vier-Stufen-Systematik vor, die technologiebasierte Assessments und inkludierte Itemformate unter anderem in Abhängigkeit der Vielfalt der Handlungsmuster und der Aufgabenkomplexität aufsteigend anordnet und eine Differenzierung unterschiedlicher Itemformate ermöglicht (s. Tabelle 1).

Tabelle 1: Stufen computerbasierter Assessments ausgehend von DiCerbo/Behrens (2012)

| | **Stufe 1** | **Stufe 2** | **Stufe 3** | **Stufe 4** |
|---|---|---|---|---|
| **Assessmenttyp** | Computerbasierter Multiple-Choice-Test | Problemlösetest | Simulierte Handlungsumgebung | Reale Handlungsumgebung |
| **Itemformat** | Diskrete Items | Wohldefinierte dynamische Aufgaben/Probleme | Komplexe Aufgaben/ Probleme | Domänentypische Aufgaben/Probleme |
| **Offenheit des Handlungsraums** | gering | moderat | hoch | sehr hoch bis unbegrenzt |
| **Einbettung** | i.d.R. keine | symbolisch | authentisch | realweltlich |
| **Beispiele** | PISA-Assessments (OECD 2001 et passim) | MicroDyn (Greiff/Funke 2009 et passim) | Computersimulierte Arbeitsproben (Gschwendtner et al. 2009); Assessment domänenspezifischer Problemlösekompetenz (z.B. Wuttke et al. 2015); computerbasierte Lernumgebungen (Kinnebrew/Segedy/ Biswas 2018) | Arbeitsproben im Rahmen von Personalauswahlprozessen (z.B. Schaper 2017) |

Von dieser Systematik ausgehend lassen sich computerbasierte Multiple-Choice-Tests (Stufe 1), die diskrete Items enthalten, einfache Problemlösetests mit wohldefinierten dynamischen Problemen („Puzzleprobleme“, vgl. Jonassen 2000) (Stufe 2) und simulierte Handlungsumgebungen mit komplexeren Items (Stufe 3) bis hin zu Assessments in realen Handlungsumgebungen (Ebene 4) mit entsprechend realweltlichen Aufgabenstellungen unterscheiden. Jenseits von Multiple-Choice-Formaten und Problemlösetests werden die Handlungs- bzw. Ergebnisräume als groß (im Falle simulierter Handlungsumgebungen) bis sehr hoch bzw. nahezu unbegrenzt (im Falle von Arbeitsproben in natürlichen Handlungsumgebungen) beschrieben (Tabelle 1). Ein Merkmal, das typischerweise mit offenen Handlungsräumen einhergeht, ist ferner die Einbettung der Aufgabenstellung in eine konkrete Handlungssituation. Dies ist im Falle simulierter und realer Handlungsumgebungen der Regelfall, wird dagegen in Multiple-Choice-Tests und Problemlösetests meist nicht oder allenfalls symbolisch umgesetzt.

Ausgehend von dieser Systematik sind die Logdaten, die aus computerbasierten Testumgebungen resultieren und aus denen Informationen über die Lösungsstrategien der Teilnehmenden gewonnen werden können, entlang der Größe des Handlungsraums und dem Format der Items unterschiedlich umfangreich und vieldeutig. Die theoriebasierte Modellierung erfolgsrelevanter Vorgehensweisen zur Lösung der Testitems ist dabei umso schwieriger, je authentischer die Testumgebung ist, d.h. je näher sie der tatsächlichen (beruflichen) Anforderungssituation kommt. Im Falle realweltlicher Arbeitsproben, wie sie etwa im Human Resource-Bereich

häufig zur Eignungsprüfung angewandt werden (z. B. Schaper 2005), ist die prozessorientierte Analyse von aufgrund der unbegrenzten Handlungsräume eine kaum lösbare Herausforderung. Hier braucht es umfangreiche Evidenzen aus datengetriebenen explorativen Zugängen und sinnvoll aggregierte sowie eingegrenzte Variablen, die sich auf bestimmte Handlungsausschnitte fokussieren, um überhaupt valide Aussagen zu metakognitiven Strategien treffen zu können. In computerbasierten Lern- oder Testumgebungen, die die Realität mit verhältnismäßig offenen Handlungsräumen simulieren, wie beispielsweise den Mikrowelten „Crystal Island“, „Betty’s Brain“ oder “gStudy” (vgl. Kinnebrew/Segedy/Biswas 2018; Lajoie et al. 2015; Sabourin et al. 2012), aber auch in der im Rahmen der Ascot-Initiative entwickelten Testumgebung zur Messung domänenspezifischer Problemlösekompetenz kaufmännischer Auszubildender (DomPL-IK; Seifried et al. 2020) müssen die vielfältigen möglichen Handlungsmuster kognitiven und metakognitiven Kompetenzfacetten zugeordnet und in eine begründete Reihenfolge gebracht werden, während sich die Handlungsstrategien im Falle weniger komplexer Probleme mit geringeren Handlungsräumen vergleichsweise eindeutiger definieren lassen (vgl. etwa die VOTAT-Strategie; Greiff/Wüstenberg/Avvisati 2015). Allerdings ist in Assessments mit vergleichsweise geringen Handlungsspielräumen und weniger komplexen Aufgaben durch die fehlende Situierung die Authentizität eingeschränkt, so dass die externe Validität der Kompetenzfeststellung unter Umständen infrage steht. Für die Messung beruflicher Kompetenz und insbesondere metakognitiver Strategien sind daher Aufgaben mit möglichst großen Handlungsspielräumen vonnöten, die einerseits aufgrund ihrer Realitätsnähe eine externe Validität der Kompetenzfeststellung ermöglichen und andererseits über gezielte Einschränkungen im Vergleich zur Realität eine intern valide Interpretation der Handlungsmuster erlauben. Die modellgetriebene Definition einer inhaltlich sinnvollen Vorgehensweise bei der Problembearbeitung schließt dabei auch die Nutzung der in der Testumgebung zur Verfügung stehenden Werkzeuge und Funktionalitäten mit ein – ein Umstand, der besonders in komplexen Testumgebungen Relevanz für die Vieldeutigkeit der resultierenden Handlungsmuster hat. Denn in Assessments, die reale Anforderungen simulieren, ist die Gestaltung der Funktionalitäten bestenfalls auch nah an realitätsnahe Funktionsumfänge angelehnt, dies hat wiederum Konsequenzen für die Interpretierbarkeit der Logdaten und nicht zuletzt auch die Wahrnehmung der Benutzerfreundlichkeit der Testumgebung durch die Teilnehmenden. Nachfolgend werden Herausforderungen bei der Logdatenanalyse in komplexen Assessments dargestellt und für ausgewählte metakognitive Kompetenzfacetten illustriert, welchen Einfluss die Gestaltung der Funktionalitäten in der Testumgebung auf das eingangs skizzierte Spannungsverhältnis aus konstrukt-irrelevanter Varianz und Konstruktunterrepräsentation hat.

# 4 Logdatenanalyse in komplexen Assessments im Spannungsfeld zwischen Benutzerfreundlichkeit und Mehrdeutigkeit

## 4.1 Herausforderungen bei der Logdatenanalyse in offenen Itemformaten

Die Interpretation von Logdaten in komplexen Assessments mit großen Handlungsräumen ist jenseits der mit dem Datenumfang assoziierten Herausforderungen (Schrader/Lawless 2007) in vielerlei Hinsicht anspruchsvoll. Erste notwendige Schritte der Datenaufbereitung und -ag-

gregation sind etwa stark davon abhängig, welche Struktur und welches Format für die Datenspeicherung verwendet wurden (Schmitz/Yanenko 2019). Die Schwierigkeit bei der Interpretation von Logdaten resultiert weiterhin aus der Tatsache, dass sich erfolgsrelevante Interaktionen der Teilnehmenden mit der Testumgebung auf deren Häufigkeit, Dauer, Zeitpunkt oder Reihenfolge beziehen können. So ist etwa die Interpretation der Nutzung von für die Problemlösung irrelevanten Informationen in der Testumgebung durch die Teilnehmenden davon abhängig, wie häufig, wann oder wie lange diese erfolgt. Wenn zu Beginn der Testzeit etwa besonders häufig und lange auf irrelevante Dokumente zurückgegriffen wird, ist dies vermutlich ein Indiz für Fehlvorstellungen oder fehlendes Wissen der Teilnehmenden bezüglich der geeigneten Vorgehensweise bei der Aufgabenlösung. Erfolgt die Beschäftigung mit irrelevanten Informationen oder Dokumenten jedoch gegen Ende des Bearbeitungszeitraums dient sie womöglich nur der Absicherung der bereits erarbeiteten (korrekten) Lösung oder gar der Überbrückung verbleibender Testzeit. Die Bedeutung der Handlungsmuster ist daher je nachdem, wann sie erfolgt mit Blick auf die Kompetenz des Teilnehmenden unterschiedlich zu werten.

Darüber hinaus sind ideale Pfade im Sinne effizienter Bearbeitungsprozesse auf Ebene der Logfiles kaum eindeutig definierbar, da der Detaillierungsgrad der gesammelten Daten höher ist als die dem Item zugrundeliegende domänenspezifische Modellierung sinnvoller Lösungswege. Zur Identifikation erfolgreicher Vorgehensweisen lassen sich im Rahmen explorativer Vorarbeiten grundsätzlich verschiedene Kriterien heranziehen: (a) Vergleiche mit Pfaden von Experten_innen (b) Vergleiche mit Pfaden von mehr oder weniger erfolgreichen Probanden_innen (c) Analyse der Pfade anhand des Vergleichs von Übergangsmatrizen (Sequenzen des Aufrufs von Seiten in der Testumgebung).

Eine adäquate Aufgabenanalyse vorausgesetzt, wären zunächst jeweils die für die Korrektheit der Aufgabenbewältigung relevanten erfassbaren Aktivitäten der Lernenden zu bestimmen. In den computersimulierten Arbeitsproben zur Kostenrechnung des Projekts Arbeitsanaloge Lernaufgaben (AALA; Niegemann 1998) waren das u.a. die Nutzung der Relevanz, Informationen (z.B. BAB), die aktiv gewählt werden mussten, die Entscheidung für eine Vorgehensweise der Berechnung, Berechnungen (mittels Taschenrechner), Aufruf von Hilfen, Zeitdauer von Aktivitäten, Interpretation von (Kosten-)Berechnungen. Die Sequenz des Aufrufs der Seiten wurde per Verlaufsanalyse (Übergangsmatrizen) dargestellt. Ziel des Projekts war der Aufbau integrierter Wissensstrukturen durch selbständig zu bearbeitende arbeitsanaloge Lernaufgaben (Hofer/Niegemann 2000, 60ff.). Ein aktuelles Beispiel (VR-basierter Fahrsimulator), bei dem auch die Reihenfolge von Handlungen (Geschwindigkeitswahl, Bremsverhalten Spurhalten und Blickverhalten/Kopfrotation) automatisch geloggt wird, berichtet ferner Malone (2020, 510).

Zudem ist die Definition von Abweichungskorridoren auf Logdaten-Ebene schwierig, denn die notwendige Sensitivität der Analysen bezüglich erfolgsrelevanter Verhaltensweisen ist ebenfalls unbekannt. So hat die einmalige Öffnung irrelevanter Dokumente bei der Informationsbeschaffung im Rahmen eines Problemlöseprozesses in der Regel keinerlei Effekt auf den Erfolg, eine mehrmalige Beschäftigung mit den Dokumenten womöglich schon. Die Identifikation erfolgs(ir)relevanter Aktivitäten bzw. Aktivitätsmuster der Teilnehmenden ist insofern ein wichtiger Schritt, der im Rahmen vorgeschalteter datengetriebener Explorationsschritte in

den Daten erfolgen sollte, um die anschließende modellgetriebene Analyse der idealen Vorgehensweise zu unterstützen. Hier greifen datengetriebene und modellgetriebene Ansätze im besten Falle ineinander – aus der datengetriebenen Analyse resultieren erste Erkenntnisse über die Erfolgsrelevanz der Häufigkeit und des Zeitpunktes verschiedener Aktivitäten oder Aktivitätsketten und diese Erkenntnisse fließen als Vorannahmen in die modellgetriebene (hypothesenprüfende) Analyse erfolgversprechenden Handelns ein.

## 4.2 Konsequenzen der Funktionsgestaltung in der Testumgebung für Logdaten und Nutzererfahrung

In Assessments, die den Anforderungsgehalt kaufmännischer Probleme realitätsnah nachbilden, umfasst eine zielführende Problembearbeitung neben der Beschäftigung mit authentischen Informations- und Arbeitsmaterialien auch die zielführende Nutzung der für die Aufgabenbearbeitung notwendigen Programme, etwa für Tabellenkalkulation, Textverarbeitung oder E-Mail-Kommunikation nebst einigen Dienstprogrammen bzw. kognitiven Werkzeugen wie Taschenrechner und elektronischem Notizblock (Eigenmann et al. 2015). Darüber hinaus besteht in der Realität die Möglichkeit, mehrere Programmfenster nebeneinander zu öffnen, in ihnen zu scrollen und im Laufe des Bearbeitungsprozesses ohne größeren Aufwand zwischen ihnen zu wechseln, etwa um Daten zu übertragen oder Informationen zu recherchieren.

Logdaten, die in einer realen Handlungsumgebung mit den beschriebenen Funktionalitäten etwa im Rahmen von Arbeitsproben gewonnen werden, sind mehrdeutig und bezüglich der metakognitiven Strategien schwerlich zu interpretieren (siehe oben). So lässt sich bei parallel ausgeführten Tätigkeiten über mehrere Fenster hinweg der Bearbeitungsprozess nicht ohne weiteres nachzeichnen. Wenn zum Beispiel in verschiedenen Dokumenten gleichzeitig Informationen recherchiert werden – ein im Arbeitsprozess häufig anzutreffender Vorgang –, ist der Aufmerksamkeitsfokus der Teilnehmenden unklar. Dieser ließe sich indes eindeutig bestimmen, wenn um den Mauszeiger herum vergleichbar einer Lupe nur ein kleiner Ausschnitt sichtbar wäre, dies wäre aber wiederum für die Teilnehmenden sehr mühsam und zudem weit von der realen Arbeitssituation entfernt. Mit Blick auf die Validität der Logdatenanalyse erscheinen Lösungen sinnvoller, die die Bildschirmteilung und gleichzeitige Öffnung und Sichtung von Dokumenten vermeiden.

Ein wichtiger Aspekt der Funktionsgestaltung mit Blick auf die Mehrdeutigkeit der Logdaten besteht ferner darin, dass bei der Verwendung der zur Verfügung stehenden Funktionalitäten die Handlungsintentionen der Teilnehmenden verborgen bleiben, was die inhaltlich sinnvolle Zuordnung der Interaktionen mit der Testumgebung zu Problemlöseschritten bzw. Phasen erschwert. Mit Blick auf die Interpretierbarkeit von Logdaten sind daher bei der Gestaltung der Werkzeuge und Funktionalitäten in computerbasierten Assessments im Vergleich zu realweltlichen Handlungskontexten Abstriche zu machen. Dabei können grundsätzlich drei Wege beschritten werden: (1) Es kann die Funktionsvielfalt reduziert und z.B. auf bestimmte Programme, Tools oder Dokumente verzichtet werden. Diese Entscheidung ist mit Blick auf domänenspezifische Arbeitsanforderungen vorzunehmen, die im Rahmen der Domänenanalyse identifiziert werden sollten. So wäre es etwa im Bereich des Controllings undenkbar, auf ein Tabellenkalkulationsprogramm zu verzichten, da dieses eines der zentralen Arbeitsmittel

darstellt. (2) Darüber hinaus lässt sich der Funktionsumfang beschränken, beispielsweise indem in der Tabellenkalkulation nur die für die Aufgabenlösung unmittelbar benötigten Funktionen zur Verfügung stehen. (3) Schließlich könnten Funktionen zeitlich oder räumlich eingeschränkt werden, um die Interpretierbarkeit von Logdaten zu erhöhen. Denkbar wäre etwa, dass Werkzeuge wie Taschenrechner und Notizblock nur in einem begrenzten Zeitraum zu Beginn der Testzeit zur Verfügung stehen, keine parallele Nutzung mehrerer Funktionen möglich ist oder beispielsweise nur ein bestimmter Tabellenausschnitt eingesehen werden kann. Allerdings kommen einige dieser Möglichkeiten einer (Vor-)Strukturierung des komplexen Problems gleich, die zwar die Logdaten weniger mehrdeutig macht, die externe Validität der Messung aber beeinträchtigt, da die Komplexität des Problems deutlich reduziert wird. Somit ist eine Abwägung zu treffen, die die Authentizität der Testung, die Interpretierbarkeit der Logdaten und die Benutzerfreundlichkeit für die Teilnehmenden balanciert.

Denn die beschriebenen Möglichkeiten führen nicht nur zu einer gewissen Reduzierung der Authentizität der Testumgebung und sind ob einer etwaigen Beeinträchtigung der externen Validität der Messung auf ein Minimum zu begrenzen, sie gehen zudem auch mit einer Reduzierung der Benutzerfreundlichkeit in der Testumgebung einher. Ausgehend von dem Ansatz des User-Centered Design (UCD; Carroll 1995, 1997) sollen die Funktionalitäten in Testumgebungen möglichst niedrigschwellig, intuitiv und ohne größere Einlernphase zu bedienen sein, nicht von den Testinhalten ablenken und möglichst viele unterschiedliche Handlungsstrategien unterstützen (Harms/Adams 2008). Die grundlegende Zielsetzung des User-Centered Design besteht in der Identifikation und Evaluation von Gestaltungsprinzipien für computerbasierte Anwendungen bzw. Umgebungen, die den jeweiligen Adressaten eine möglichst zielführende und störungsfreie Nutzererfahrung ermöglichen (Garett 2011). Harms/Adams (2008) weisen in diesem Zusammenhang zudem auf den Umstand hin, dass Testteilnehmende sich hinsichtlich ihrer Eingangsvoraussetzungen unterscheiden und in Abhängigkeit dessen die Nutzerfreundlichkeit möglicherweise unterschiedlich wahrgenommen wird. Auch bei der Gestaltung der Funktionalitäten in der Testumgebung ist somit das eingangs angesprochene Spannungsfeld aus konstrukt-irrelevanter Varianz und Konstruktunterrepräsentation zu berücksichtigen: Zu geringe Funktionsumfänge gehen an der Anforderungsrealität vorbei und führen bei der Messung metakognitiver Strategien aufgrund des zu geringen Authentizitätsgrades zu Konstruktunterrepräsentation, ebenso wie vollkommen authentische Funktionsumfänge ihrerseits wiederum aufgrund der Mehrdeutigkeit der resultierenden Logdaten diese Problematik hervorrufen. Eine möglichst weitgehende Reduzierung der Funktionsumfänge führt wiederum zu einer starken Einschränkung der Benutzerfreundlichkeit und erzeugt zudem konstrukt-irrelevante Varianz. Wie sich die entsprechenden Abwägungen bei der Gestaltung der Funktionalitäten darstellen, wird nachfolgend für die metakognitive Kompetenzfacette der Handlungskontrolle exemplarisch illustriert.

### 4.3 Beispiel: Funktionsgestaltung zur Analyse der Handlungskontrolle von Testteilnehmenden mittels Logdaten

Erfolgreiche Problemlöser kontrollieren im Handlungsprozess und nach Abschluss der Problembearbeitung regelmäßig die erzielten Fortschritte und plausibilisieren das Arbeitsergebnis.

Eine wichtige metakognitive Facette von Problemlösekompetenz betrifft insofern die begleitende oder retrospektive Handlungskontrolle (z.B. Betsch/Funke/Plessner 2011; Lester 1994; Rausch/Wuttke 2016). In Assessments müssen die Testleistungen zumeist unter zeitlichen Beschränkungen erbracht werden. Insofern ist es für die Identifikation der begleitenden Handlungskontrolle von Interesse, das Zeitmanagement der Teilnehmenden über die Logdaten zu beobachten. An realen Computerarbeitsplätzen gehört der regelmäßige Blick auf die Uhr zur Alltagsroutine und ist insofern erleichtert, weil diese dauerhaft eingeblendet ist (je nach Betriebssystem am oberen oder unteren Bildschirmrand). Die Frage, ob und wie regelmäßig Teilnehmende ein aktives Zeitmanagement betreiben, ist in den Logdaten aber nur dann beantwortbar, wenn die Testumgebung eine aktive Handlung einfordert. Ist die Uhr dauerhaft sichtbar, kann über die Logdaten nicht beobachtet werden, wie das Zeitmanagement ausfällt. Es bietet sich daher an, ein Uhrsymbol gut sichtbar in der Testumgebung zu platzieren, die verbleibende Testzeit oder aktuelle Uhrzeit aber nur nach Anklicken für einen kurzen Zeitraum preiszugeben. Die sich dadurch ergebende Beeinträchtigung der Benutzerfreundlichkeit ist vergleichsweise gering.

Auch die rückschauende Reflektion der erarbeiteten Lösung in der Testumgebung ist in den Logdaten nicht ohne weiteres zu beobachten, wenn nicht von außen ein entsprechender Impuls erfolgt, der die Teilnehmenden explizit dazu auffordert. Im realen Arbeitsprozess werden Arbeitsergebnisse in der Regel an Kolleg_innen oder Vorgesetzte weitergegeben. Eine entsprechende Plausibilisierung der erarbeiteten Lösung sollte insofern bestenfalls zuvor erfolgen. In einer Testumgebung kann diese in einer erneuten Durchsicht der Informations- und Arbeitsdokumente und gegebenenfalls auch der irrelevanten Informationen bestehen. Der Prozess der Bearbeitung und der nachgängigen Kontrolle gehen insofern von außen betrachtet fließend ineinander über. Für eine bessere Interpretierbarkeit wäre zu überlegen, ob nach Beendigung des letzten für die Problembearbeitung relevanten Arbeitsschrittes die Aufforderung an die Teilnehmenden ergehen sollte, vor der Abgabe der Lösung noch einmal plausibilisierende Schritte zu unternehmen. Hier würde der Problemlöseprozess in gewissem Maße strukturiert, anders ließe sich aber eine die Frage der Handlungskontrolle zumindest in den Logdaten nur schwerlich beantworten. Andere technische Lösungen, die ein Weiterarbeiten ohne entsprechende Plausibilisierungsschritte unmöglich machen, wären mit Blick auf die Benutzerfreundlichkeit problematisch und würden unter Umständen auch zu konstrukt-irrelevanter Varianz führen.

## 5 Ausblick

Dieser Beitrag hat sich mit der Frage beschäftigt, welche Herausforderungen die modellgetriebene Analyse von Logdaten in komplexen Kompetenztests mit offenen Handlungsräumen mit Blick auf Validitätsfragen mit sich bringt und welche Implikationen sich daraus für das Design von Kompetenztests ableiten lassen. Ausgehend von dem Ziel einer prozessnahen Erfassung der Handlungsstrategien von Testteilnehmenden im Rahmen computerbasierter Assessments wurde das Spannungsverhältnis zwischen der Authentizität der Testumgebung, ihrer Benutzerfreundlichkeit und der Interpretierbarkeit der Logdaten skizziert.

Es zeigte sich, dass in offenen und komplexen Itemformaten, in denen die Testteilnehmenden vielfältige Lösungswege beschreiten können, insbesondere die Identifikation idealer Lösungspfade und Abweichungstoleranzen in den Logdaten herausfordernd ist. Hier offenbart sich zum einen der Mehrwert datengetriebener Zugänge, die auf explorativem Wege erfolgversprechende Handlungsmuster aufdecken und eine nachgängige Überprüfung im Rahmen modellgetriebener hypothesenprüfender Ansätze ermöglichen. Zum anderen bleibt mit Blick auf das Ziel einer möglichst intern validen Kompetenzfeststellung zu konstatieren, dass andere methodische Zugänge zur Analyse von Handlungsprozessen, etwa Think-Aloud- oder Eye-Tracking-Verfahren vielversprechende und wichtige ergänzende Validierungsmöglichkeiten bieten.

Für simulierte Handlungsumgebungen haben wir ferner verdeutlicht, inwiefern unterschiedliche Gestaltungsoptionen in der Testumgebung einen Einfluss auf die valide Interpretation der Handlungsmuster der Teilnehmenden nehmen können. Dabei wurde die übergeordnete Bedeutung eines in den Logdaten klar zuordenbaren Aufmerksamkeitsfokus der Testteilnehmenden offenbar – eine Anforderung, die sich in realweltlichen Arbeits- bzw. Testumgebungen ohne starke Eingrenzung der zu erhebenden Logdaten nur schwerlich realisieren lässt. In Testumgebungen, die die Komplexität beruflicher Anforderungssituationen näherungsweise simulieren, sind daher für eine klare Zurechenbarkeit der Aktivitätsmuster zu Kompetenzfacetten Abstriche in der Benutzerfreundlichkeit der Testumgebung notwendig, die allerdings zur Vermeidung konstrukt-irrelevanter Varianz auf das Nötigste zu begrenzen sind. Am Beispiel der begleitenden und retrospektiven Handlungskontrolle als metakognitiver Kompetenzfacette wurde verdeutlicht, dass erfolgsrelevante Handlungen in den Logfiles nur dann sichtbar werden, wenn sie – wie etwa der Blick auf die Uhr im Rahmen des Zeitmanagements per Mausklick – durch eine entsprechende Gestaltung des Funktionalitäten in der Testumgebung auch notwendigerweise von den Testteilnehmenden eingefordert werden.

In der Gesamtschau bleibt die modellgetriebene Logdatenanalyse im Rahmen der Kompetenzmessung ein anspruchsvolles Unterfangen, das aber große Potenziale zur Ergänzung der ansonsten dominierenden produktorientierten Perspektive auf das Kompetenzkonstrukt aufweist. Für ein Gelingen dieses Unterfangens sind externe Validität im Sinne einer größtmöglichen Authentizität des Assessments und interne Validität im Sinne einer möglichst geringen Mehrdeutigkeit der Logdaten unter Berücksichtigung der Nutzerfahrung in der Testumgebung bestmöglich zu balancieren.

Doch auch jenseits der Kompetenzmessung lässt sich der detailreiche Blick auf Handlungsmuster in Gestalt von Logfiles nutzbar machen: Bei der Analyse und individualisierten Förderung von Lehr- und Lern-Prozessen offenbaren sich vielfältige Einsatzfelder. Denn die valide Erfassung und Interpretation von Logdaten ist auch eine Kernaufgabe von Learning Analytics (Ifenthaler/Drachsler 2020). Dort geht es im Wesentlichen um die Verwendung nutzergenerierter Daten für die Optimierung von Lehr-Lernprozessen. “Learning Analytics verwenden statisch und dynamisch generierte Daten von Lernenden und Lernumgebungen, um diese in Echtzeit zu analysieren und zu visualisieren, mit dem Ziel der Modellierung und Optimierung von Lehr-Lernprozessen und Lernumgebungen” (Ifenthaler/Widanapathirana 2014). Die Beschränkung auf “Echtzeit” in dieser Definition grenzt das Gebiet nur vermeintlich von der nachgängigen diagnostischen Zielsetzung der Logdatenanalyse für die Kompetenzbestimmung

ab. Denn im Falle von Learning Analytics besteht eines der Hauptziele in der Identifikation schwächerer Lernender und der Prognose von erfolgversprechenden Handlungsmustern, eine zweifellos diagnostische Aufgabe, die der Kompetenzmessung ebenso zugrunde liegt. Offensichtlich gibt es bei vielen Fragen Überschneidungen, so dass Kompetenzdiagnostik und Learning Analytics hier voneinander profitieren können. Schließlich und endlich dürfte der Stellenwert logdatenbasierter Zugänge zu Lehr-, Lern- wie auch Prüfungsprozessen nicht zuletzt mit Blick auf den erstarkenden Diskurs über die Bedeutung künstlicher Intelligenz in der beruflichen Bildung weiter anwachsen.

## Literatur


Abele, S. (2018): Diagnostic Problem-Solving Process in Professional Contexts: Theory and Empirical Investigation in the Context of Car Mechatronics Using Computer-Generated Log-Files. In: Vocations and Learning, 11(1), 133159. https://doi.org/10.1007/s12186-017-9183-x.

Abele, S./Ostertag, R./Peissner, M./Schuler, A. (2017): Eine Eye-Tracking-Studie zum diagnostischen Problemlöseprozess: Bedeutung der Informationsrepräsentation für das Lösen diagnostischer Probleme. In: Zeitschrift für Berufs- und Wirtschaftspädagogik, 113, 86-109.

Arieli-Attali, M./Ward, S./Thomas, J./Deonovic, B./Davier, A. (2019): The Expanded Evidence-Centered Design (e-ECD) for Learning and Assessment Systems: A Framework for Incorporating Learning Goals and Processes Within Assessment Design. In: Frontiers in Psychology, 1-17. https://doi.org/10.3389/fpsyg.2019.00853.

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

Behrens, J. T./DiCerbo, K. E./Ferrara, S. (2012): Intended and unintended deceptions in the use of simulations. Invitational Research Symposium on Technology Enhanced Assessments, Educational Testing Service, Washington.

Behrens, J. T./DiCerbo, K. E./Foltz, P. W. (2019): Assessment of Complex Performances in Digital Environments. In: The ANNALS of the American Academy of Political and Social Science, 683(1), 217-232. https://doi.org/10.1177/0002716219846850.

Betsch, T./Funke, J./Plessner, H. (2011): Denken - Urteilen, Entscheiden, Problemlösen. Berlin.

Bransford, J./ Stein, B. S. (1993): The ideal problem solver: A guide for improving thinking, learning, and creativity (2nd ed). New York.

Campbell, D. T. (1957): Factors relevant to the validity of experiments in social settings. In: Psychological Bulletin, 54(4), 297-312. https://doi.org/10.1037/h0040950.

Carroll, J. M. (1995): Human-computer interaction: psychology as a science of design. In: Annual Review of Psychology, 48, 61-83.

Carroll, J. M. (1997): Scenario-based design. In: Helander, M./Landauer, T. K. (Eds.): Handbook of Human-Computer Interaction. Second Edition. Amsterdam, 383-406.

Cronbach, L. J. (1971): Test validation. In: Thorndike, R. L. (Ed.): Educational measurement, 2nd ed. Washington, DC, 443-507.

Deutscher, V./Winther, E. (2018): A Conceptual Framework for Authentic Competence Assessment in VET: A Logic Design Model. In: McGrath, S./Mulder, M./Papier, J./Suart, R. (Hrsg.): Handbook of Vocational Education and Training. Basel, 1-14. https://doi.org/10.1007/978-3-319-49789-1_80-1.

DiCerbo, K. E./Behrens, J. T. (2012). From technology-enhanced assessments to assessment-enhanced technology. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.

Dörner, D. (2000): Logik des Misslingens – Strategisches Denken in komplexen Situationen. 13. Aufl. Reinbek bei Hamburg.

Eigenmann, R./Siegfried, C./Kögler, K./Egloffstein, M. (2015): Aufgaben angehender Industriekaufleute im Controlling: Ansätze zur Modellierung des Gegenstandsbereichs. In: Zeitschrift für Berufs- und Wirtschaftspädagogik, 111(3), 417-436.

Flavell, J. H. (1979): Metacognition and cognitive monitoring: A new area of cognitive developmental inquiry. In: American Psychologist, 34, 906-911.

Frey, B. B./Schmitt, V. L./Allen, J. P. (2012): Defining authentic classroom assessment. In: Practical Assessment, Research & Evaluation, 17(2), 1-18.

Garrett, J. J. (2011): The Elements of User Experience. User-Centered Design for the Web and Beyond. Berkeley.

Glaser, R. (1994): Expertise. In: Eysenck, M. W. et al. (Eds.): The Blackwell Dictionary of Cognitive Psychology, Cambridge: Blackwell., 139-142.

Greiff, S./Funke, J. (2009): Measuring complex problem solving – The MicroDYN approach. In: Scheuermann, F./Björnsson, J. (Eds.): The transition to computer-based assessment - Lessons learned from large-scale surveys and implications for testing. Luxembourg: Office for Official Publications of the European Communities,157-166.

Greiff, S./Wüstenberg, S./Avvisati, F. (2015): Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. In: Computers & Education, 91, 1-14.
http://dx.doi.org/10.1016/j.compedu.2015.10.01.

Gschwendtner, T./Abele, S./Nickolaus, R. (2009): Computersimulierte Arbeitsproben: Eine Validierungsstudie am Beispiel der Fehlerdiagnoseleistungen von Kfz-Mechatronikern. In: Zeitschrift für Berufs- und Wirtschaftspädagogik, 4, 557-57.

Gulikers, J. T. M./Bastiaens, T. J./Kirschner, P. A. (2004): A Five-Dimensional Framework for Authentic Assessment. In: Educational Technology Research and Development, 52(3), 67-86.

Harms, M./Adams, J. (2008): Usability and Design Considerations for Computer-Based Learning and Assessment.

He Q./Von Davier, M. (2016): Analyzing Process Data from Problem-Solving Items with N-Grams: Insights from a Computer-Based Large-Scale Assessment. In: Rosen, Y./Ferrara, S./ Mosharraf, M. (Eds.): Handbook of Research on Technology Tools for Real-World Skill Development, 749-776, doi: 10.4018/978-1-4666-9441-5.ch029.

Hofer, M./Niegemann, H. M. (2000): Förderung des Aufbaus integrierter Wissensstrukturen durch selbständig zu bearbeitende arbeitsanaloge Lernaufgaben zur Kostenrechnung in einer computerbasierten komplexen Lernumgebung (Kurzbericht). In: Beck, K. (Ed.): Lehr-Lern-Prozesse in der kaufmännischen Erstausbildung. Ein Schwerpunktprogramm der Deutschen Forschungsgemeinschaft. Kurzberichte und Bibliographie. Landau: Verlag Empirische Pädagogik, 60-67.

Ifenthaler, D./Drachsler, H. (2020): Learning Analystics. In: Niegemann, H./Weinberger, A. (Hrsg.): Lernen mit Bildungstechnologien, Heidelberg, 515-534.

Ifenthaler, D./Widanapathirana, C. (2014): Development and validation of a learning analytics framework: Two case studies using support vector machines. In: Technology, Knowledge and Learning, 19(1–2), 221–240. doi:10.1007/s10758-014-9226-4.

Jonassen, D. H. (2000): Toward a Design Theory of Problem Solving. In: Educational Technology Research and Development, 48(4), 63-85, doi: 10.1007/BF02300500.

Kane, M. T. (2013): Validating the Interpretations and Uses of Test Scores. In: Journal of Educational Measurement, 50(1), 1-73.

Kane, M./Crooks, T./Cohen, A. (1999): Validating Measures of Performance. In: Educational Measurement: Issues and Practice, 18(2), 5-17. https://doi.org/10.1111/j.1745-3992.1999.tb00010.x.

Kinnebrew, J./Mack, D./Biswas, G./Chang, C. (2014): A Differential Approach for Identifying Important Student Learning Behavior Patterns with Evolving Usage over Time. In: Peng, W. C. et al. (Eds.): Trends and Applications in Knowledge Discovery and Data Mining. PAKDD 2014. Lecture Notes in Computer Science, vol 8643. Cham. https://doi.org/10.1007/978-3-319-13186-3_27.

Kinnebrew, J./Segedy, J. R./Biswas, G. (2018): Integrating model-driven and data-driven techniques for analyzing learning behaviors in open-ended learning environments. In: IEEE Transactions on Learning Technologies, 10(2), 140-153.

Lane, S./Stone, C. A. (2006): Performance assessments. In: Brennan, B. (Ed.), Educational Measurement. Westport, CT: American Council on Education & Praeger, 387-431.

Lajoie S. P./Poitras E. G./Doleck T./Jarrell A. (2015): Modeling Metacognitive Activities in Medical Problem-Solving with BioWorld. In: Peña-Ayala, A. (Ed.) Metacognition: Fundaments, Applications, and Trends. Intelligent Systems Reference Library, vol 76. Cham. https://doi.org/10.1007/978-3-319-11062-2_13

Lester, F. K. (1994): Musings about mathematical problem-solving research: 1970–1994. In: Journal for Research in Mathematis Education, 25, 660-675.

Malone, S. (2020): Technologiegestütztes Assessment, Online Assessment. In: Niegemann, H./Weinberger, A. (Hrsg.): Handbuch Bildungstechnologie. Berlin, 493-514.

Messick, S. (1989): Validity. In: Linn, R. L. (Ed.): Educational Measurement. New York, 13-103.

Messick, S. (1994): The Interplay of Evidence and Consequences in the Validation of Performance Assessments. In: Educational Researcher, 23(2), 13-23. https://doi.org/10.3102/0013189X023002013.

Mislevy, R. J./Behrens, J. T./Bennett, R. E./Demark, S. F./Frezzo, D. C./Levy, R./Robinson, D. H./Rutstein, D. W./Shute, V. J./Stanley, K./Winters, F. I. (2010): On the Roles of External Knowledge Representations in Assessment Design. In: Journal of Technology, Learning, and Assessment, 8(2).

Mislevy, R. J./Haertel, G./Riconscente, M./ Rutstein, D. W./Ziker, C. (2017): Evidence-Centered Assessment Design. In: Mislevy, R. J./ Haertel, G./ Riconscente, M./ Wise, D./Rutstein/ Ziker, C. (Eds.): Assessing Model-Based Reasoning using Evidence- Centered Design. Springer International Publishing, 19-24. https://doi.org/10.1007/978-3-319-52246-3_3.

Mislevy, R. J./Steinberg, L. S./Almond, R. G. (2003): On the Structure of Educational Assessments. In: Measurement: Interdisciplinary Research & Perspective, 1(1), 3-62. https://doi.org/10.1207/S15366359MEA0101_02.

Niegemann, H. M. (1998): A case-based learning environment for problem-oriented learning in business education: Design principles and empirical results. In: Merrienboer, J. (Ed.): Proceedings of the Third Workshop of the EARLI SIG Instructional Design, June 26-27, University of Maastricht, The Netherlands, 183-194.

Niegemann, H./Heidig, S. (2020): Interaktivität und Adaptivität in multimedialen Lernumgebungen. In Niegemann, H. M. & Weinberger, A. (Hrsg.): Handbuch Bildungstechnologie. Berlin, 343-368.

OECD (Hrsg.): (2001). Knowledge and skills for life: First results from PISA 2000. Paris.

Pellegrino, J./Chudowsky, N./Glaser, R. (2001): Knowing What Students Know: The Science and Design of Educational Assessment. In: The National Academies Press. https://doi.org/10.17226/10019.

Pellegrino, J./ DiBello, L./Goldman, S. (2016): A Framework for Conceptualizing and Evaluating the Validity of Instructionally Relevant Assessments. In: Educational Psychologist. http://dx.doi.org/10.1080/00461520.2016.1145550.

Rausch, A./Kögler, K. (2016). Authenticity and efficiency in assessing domain-specific problemsolving competence: Conflicting goals in large-scale-assessments? In Mulder, M./Wesselink,R./Biemans, H./Lans, T. (Eds.): International Conference on Competence Theory, Research and Practice in Wageningen, Wageningen, 339-345.

Rausch, A./Wuttke E. (2016): Development of a multi-faceted model of domain-specific problem-solving competence and its acceptance by different stakeholders in the business domain. In: Unterrichtswissenschaft, 44 (2), 164-189.

Rausch, A./Kögler, K./Frötschl, C./Bergrab, M./Brandt, S. (2017): Problemlöseprozesse sichtbar machen: Analyse von Logdaten aus einer computerbasierten Bürosimulation. In: Zeitschrift für Berufs- und Wirtschaftspädagogik, 113(4), 569-594.

Sabourin, J./Rowe, J./Mott, B./Lester, J. (2012): Exploring Inquiry-based Problem-Solving Strategies in Game-based Learning Environments. Berlin.

Schaper, N. (2017): Arbeitsproben und situative Fragen zur Messung arbeitsplatzbezogener Kompetenzen. In: Erpenbeck, J./Rosenstiel, L.v./Grote, S./Sauter, W. (Hrsg.): Handbuch Kompetenzmessung. Erkennen, verstehen und bewerten von Kompetenzen in der betrieblichen, pädagogischen und psychologischen Praxis. Stuttgart, 523-537. https://doi.org/10.34156/9783791035123-523.

Schmitz, A./Yanenko, O. (2019): Web Server Logs und Logfiles. In: Baur, N./ Blasius, J. (Hrsg.): Handbuch Methoden der empirischen Sozialforschung. Wiesbaden, 991-999. https://doi.org/10.1007/978-3-658-21308-4_70.

Schraagen, J./Chipman, S./Shalin, V. (2000): Cognitive Task Analysis. Mahwah, Nj.

Schrader, P. G./Lawless, K. (2007): Dribble files: Methodologies to evaluate learning and performance in complex environments. In: International Society for Performance Improvement 46 (1), 40-48.

Schwantzer, S./Jannaber, S. (2018): Modellierung technischer Serviceprozesse zur Digitalisierung der Aus- und Weiterbildung. In: Thomas, O./Metzger, D./Niegemann, H. (Hrsg.): Digitalisierung in der Aus- und Weiterbildung. Virtual und Augmented Reality für Industrie 4.0. Berlin, 64-74.

Seifried, J./Brandt, S./Kögler, K./Rausch, A. (2020): The Computer-Based Assessment of Domain-Specific Problem-Solving Competence – A three-step scoring procedure. In: Cogent Education, 7(1), 1-20. https://doi.org/10.1080/2331186X.2020.1719571.

Shadish, W. R./Cook, T. D./Campbell, D. T. (2002): Experimental and quasi-experimental designs for generalized causal inference. Boston.

Shute, V. J./Leighton, J. P./ Jang, E. E./ Chu, M.-W. (2016): Advances in the Science of Assessment. In: Educational Assessment, 21(1), 34-59. http://dx.doi.org/10.1080/10627197.2015.1127752.

Swanson, D. B./Norman, G. R./Linn, R. L. (1995): Performance-Based Assessment: Lessons From the Health Professions. In: Educational Researcher, 24(5), 5-11. https://doi.org/10.3102/0013189X024005005.

Sweller, J./Ayres, P./Kalyuga, S. (2011): Cognitive Load Theory. New York.

Veenman, M. V. J./Bavelaar, L./De Wolf, L./Van Haaren, M. G. P. (2014): The on-line assessment of metacognitive skills in a computerized learning environment. In: Learning and Individual Differences, 29, 123-130. https://doi.org/10.1016/j.lindif.2013.01.003.

Weinert, F. E. (2001): Vergleichende Leistungsmessung in Schulen - eine umstrittene Selbstverständlichkeit. In: Weinert, F. E. (Hrsg.): Leistungsmessungen in Schulen. Weinheim, 17-32.

Wuttke, E./Seifried, J./Brandt, S./Rausch, A./Sembill, D./Martens, T./Wolf, K. (2015): Modellierung und Messung domänenspezifischer Problemlösekompetenz bei angehenden Industriekaufleuten: Entwicklung eines Testinstruments und erste Befunde zu kognitiven Kompetenzfacetten. In: Zeitschrift für Berufs- und Wirtschaftspädagogik, 2, 189-207.

## Zitieren dieses Beitrags

Kögler, K./Rausch, A./Niegemann, H. (2020): Interpretierbarkeit von Logdaten in computerbasierten Kompetenztests mit großen Handlungsräumen. In: *bwp@* Profil 6: Berufliches Lehren und Lernen: Grundlagen, Schwerpunkte und Impulse wirtschaftspädagogischer Forschung. Digitale Festschrift für Eveline Wuttke zum 60. Geburtstag, hrsg. v. Heinrichs, K./ Kögler, K./Siegfried, C., 1-21. Online: https://www.bwpat.de/profil6_wuttke/koegler_etal_profil6.pdf (08.09.2020).

## Die Autor*innen

**Prof. Dr. KRISTINA KÖGLER**

Universität Stuttgart, Abteilung Berufs-, Wirtschafts- und Technikpädagogik

Geschw.-Scholl-Str. 24D, 70174 Stuttgart

koegler@bwt.uni-stuttgart.de

https://www.ife.uni-stuttgart.de/bwt/

**Prof. Dr. ANDREAS RAUSCH**

Universität Mannheim, Lehrstuhl für Wirtschaftspädagogik – Lernen im Arbeitsprozess

L 4,1, 68161 Mannheim

rausch@uni-mannheim.de

https://www.bwl.uni-mannheim.de/rausch/

**Prof. Dr. HELMUT NIEGEMANN**

Goethe-Universität Frankfurt, Fachbereich Wirtschaftswissenschaften, Seniorprofessor Wirtschaftspädagogik

Theodor-W.-Adorno-Platz 4, 60629 Frankfurt am Main

niegemann@econ.uni-frankfurt.de

https://www.profniegemann.de/