Validierungsstrategien von Tests zur Erfassung studentischer Kompetenzen

Dissertation

zur Erlangung des Doktorgrades

der Naturwissenschaften

vorgelegt beim Fachbereich 05

der Johann Wolfgang Goethe-Universität

in Frankfurt am Main

von

Christine Aichele

aus Göppingen

Frankfurt 2021

(D30)

vom Fachbereich 05 der

Johann Wolfgang Goethe-Universität als Dissertation angenommen.

Dekanin: Prof'in. Dr. Sonja Rohrmann

Gutachter:

Prof. Dr. Johannes Hartig

Prof. Dr. Andreas Frey

Datum der Disputation: 08. September 2021

Inhalt

Abbildungsverzeichnis

Tabellenverzeichnis

Zusammenfassung

Diese Dissertation befasst sich mit Validierungsstrategien von Tests zur Erfassung studentischer Kompetenzen. Kompetenzen von Studierenden werden zu verschiedenen Zwecken erhoben. Dies beginnt beim Eintritt in das Studium durch Zulassungstests und wird im Studium fortgesetzt z.B. durch Tests zur Zertifizierung von Kompetenz (Benotung von Leistung) oder zur Zuteilung auf bestimmte Kurse (Einteilung in Sprachniveaus). Neben diesen internen Tests zur Erfassung studentischer Kompetenzen werden auch externe Tests genutzt um etwa die Lehre zu verbessern (Evaluation von Veranstaltungen). Die mit dem Einsatz von Tests verbundenen Konsequenzen können sowohl für Studierende als auch Lehrpersonen und Entscheidungsträger*innen schwerwiegend sein. Daher sollten Tests wissenschaftlichen Gütekriterien genügen.

Das wichtigste Kriterium für die Beurteilung von wissenschaftlichen Tests ist *Validität*. In dieser Dissertation wird ein argumentationsbasiertes Validierungsansatz verfolgt. In diesem wird nicht die Validität eines Tests untersucht, sondern die Plausibilität der Interpretation beurteilt, die mit den Testwerten verbunden ist. Bislang fehlt jedoch für viele der wissenschaftlichen Tests für den deutschen Hochschulbereich ein auf die Testwertinterpretation abgestimmtes Validitätskonzept.

In dieser Arbeit wird ein Validierungsschema vorgestellt, in das übliche Testnutzen der Erfassung studentischer Kompetenzen an deutschen Hochschulen eingeordnet werden können. Die Einordnung von Testnutzen in das Schema erlaubt die Ableitung von passenden Validitätsevidenzen. Im Fokus stehen das Verhältnis von Test zu 1) Konstrukt, 2) Lehre und 3) beruflichen Anforderungen.

Das Validierungsschema wird angewandt, um Testwertinterpretationen eines empirischen Forschungsprojektes zur Erfassung von Kompetenz in Nachhaltigkeitsmanagement bei Studierenden zu validieren. Der Schwerpunkt dieser Arbeit liegt auf der Validierung der Interpretation, dass die Testwerte von drei nachhaltigkeitsbezogenen Tests Indikatoren für hochschulisch vermittelte Kompetenz in Nachhaltigkeitsmanagement sind. Die Analysen zur Gewinnung von Validitätsevidenzen konzentrieren sich auf die Grundannahme, dass Lernfortschritte in den nachhaltigkeitsbezogenen Tests vorwiegend hochschulisch vermittelt werden. Dafür wurde ein Messwiederholungsdesign mit zwei Gruppen von Studierenden realisiert. Studierende in der Schwerpunktgruppe besuchten ein Semester lang eine reguläre

Lehrveranstaltungen mit Bezug zu Nachhaltigkeitsthemen und Nachhaltigkeitsmanagement, Studierende der Kontrollgruppe besuchten keine solchen Lehrveranstaltung. Die Einteilung in Schwerpunkgruppe und Kontrollgruppe erfolgte über Analyse von Modulhandbüchern und verwendeten Lehrmaterialien. Die Ergebnisse zeigen, dass Studierende aus der Schwerpunktgruppe in zwei der drei Tests höhere Lernfortschritte zeigen als Studierende der Kontrollgruppe. Selbstberichte der Studierenden zu hochschulischen und außerhochschulischen Lerngelegenheiten lassen darauf schließen, dass Studierende der Schwerpunkgruppe auch außerhochschulisch ein höheres Interesse an Nachhaltigkeitsthemen zeigen, dies schlägt sich jedoch nicht in höherem Vorwissen in den verwendeten Tests nieder. Insgesamt wird daher für die zwei Tests mit höheren Lernfortschritten in der Schwerpunktgruppe die Interpretation als plausibel bewertet, dass die Testwerte hochschulisch vermittelte Kompetenz in Nachhaltigkeitsmanagement abbilden.

Einleitung

1 Zielsetzung der Arbeit

Im deutschen Hochschulsystem werden regelmäßig wissenschaftlich überprüfte Tests bei Studierenden oder Studienplatzinteressierten eingesetzt. Schon 1986 wurde etwa der „Medizinertest“ als Instrument zur Studienplatzselektion eingesetzt. Studienorientierungstests werden angeboten, so dass Studieninteressierte die eigenen Fähigkeiten und Neigungen mit Studienanforderungen abgleichen können (beispielhaft sei hier der Studien-Interessentest (SIT) genannt, herausgegeben von der Hochschulrektorenkonferenz & ZEIT ONLINE). Ziel dieser Tests ist die Senkung von Studienabbruchquoten. Ebenfalls üblich sind Sprachtests, die Personen ein bestimmtes Sprachniveau attestieren und die Studierende häufig für Studienplatzbewerbungen benötigen, insbesondere wenn Muttersprache und Lehrsprache voneinander abweichen.

Neben der Fokussierung von Bewerber*innenauswahl für Studienplätze und der Zertifizierung von Studienleistungen, verschiebt sich in den letzten Jahrzenten der Fokus auf die erworbenen Kompetenzen von Studierenden. Im schulischen Bereich ist die Kompetenzerfassung bei Schüler*innen üblich, um Aussagen über das Bildungssystem zu ermöglichen (z.B. VERA zur Überprüfung der Erreichung von Bildungsstandards). Dies ist im hochschulischen Bereich nicht gängig und, nach Erkenntnissen der AHELO-Machbarkeitsstudie der OECD zur internationalen Vergleichbarkeit studentischer Kompetenzen, nicht einfach umzusetzen (OECD, 2013)[1]. Die Erfassung studentischer Kompetenzen wird aber genutzt, um Studiengänge zu verbessern oder die Effektivität von Lerngelegenheiten zu evaluieren. In der von Bund und Ländern geförderten Qualitätsoffensive Lehrerbildung[2] wurden unter anderem Tests genutzt, um Kompetenzverläufe bei Studierenden zu modellieren und daraus Rückschlüsse auf die Effektivität von Lerngelegenheiten zu ziehen. Beispielhafte Forschungsprogramme, in denen weitere Tests zur Erfassung studentischer Kompetenzen entwickelt wurden sind das DFG-Schwerpunktprogramm

[1] So gelang zum Beispiel schon in der Konzeptionierung der Testinhalte kein Konsens für generische Kompetenzen zwischen den internationalen Experten. Für fachspezifische Kompetenzen in Ingenieurswissenschaften und Wirtschaftswissenschaften war dies zwar möglich. Die Ergebnisse zeigen aber in allen Tests für viele Items deutliche Unterschiede in den Itemschwierigkeiten, was die Vergleichbarkeit der Testwerte einschränkt.

[2] Die Qualitätsoffensive Lehrerbildung ist ein seit 2015 (bis 2023) von Bund und Ländern gefördertes Programm zur inhaltlichen und strukturellen Verbesserung der Lehrer*innenbildung.

Kompetenzmodelle[3]und die KoKo-HS Initiative[4]. Inzwischen liegen in Deutschland viele Tests zur Erfassung von hochschulisch vermittelten fachspezifischen und generischen Kompetenzen bei Studierenden vor (Zlatkin-Troitschanskaia, Pant, Kuhn, Toepper & Lautenbach, 2016).

Ein Testwert kann für Studierende bzw. Studienbewerber*innen schwerwiegende Konsequenzen haben (z.B. bei der Studienplatzvergabe). Die Testwerte können aber auch über Weiterführung oder Beendigung einer Lerngelegenheit entscheiden, wenn der Test von Lehrkräften und Entscheidungsträger*innen zu Evaluationszwecken eingesetzt wird. Daher ist vor Einsatz eines Tests stets dessen Güte zu prüfen. Zu den wichtigsten Gütekriterien zählt Validität. 2014 veröffentlichten American Educational Research Association (AERA), American Psychological Association (APA) und National Council on Measurement in Education (NCME) in den *Standards for Educational and Psychological Testing* (*Standards*) ein gegenüber früheren Versionen der Standards neues Konzept von Validität und ein verändertes Validierungsvorgehen. Um Kompetenzen im Hochschulbereich valide zu erfassen, sind Argumente für die jeweilige Testnutzung und intendierte Testwertinterpretation notwendig (AERA, APA & NCME, 2014; Kane, 2013). Bei jeder Testentwicklung ist zunächst zu definieren, in welchen spezifischen Situationen der Test eingesetzt werden soll und welche Schlussfolgerungen man anhand des Testwertes von Teilnehmer*innen ziehen möchte. Nach dem vorgeschlagenen argumentationsbasierten Validierungsansatz sind die einer Testwertinterpretation zugrundeliegenden Annahmen zu identifizieren. Diese Annahmen müssen, wie Hypothesen, so formuliert sein, dass sie überprüfbar und falsifizierbar sind. Alle verfügbaren Evidenzen hinsichtlich der Grundannahmen werden abschließend zu einem Fazit zur Plausibilität der Testwertinterpretation zusammengefasst.

Inhaltliche Überlegungen müssen dem Validierungsprozess vorangehen, um plausible Argumente für die jeweils vorliegende Testwertinterpretation aufzustellen und dazu passende empirische Evidenzen zu liefern. Soll ein Testwert etwa als Indikator für ein theoretisches

---

[3] Das Schwerpunktprogramm wurde von 2007 bis 2016 gefördert. Zentrale Forschungsergebnisse aus vielen Projekten werden hier vorgestellt: Leutner, Fleischer, Grünkorn und Klieme (2017). *Competence assessment in education. Research, models and instruments*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-50030-0

[4] „Kompetenzmodelle und Instrumente der Kompetenzerfassung im Hochschulsektor – Validierungen und methodische Innovationen“, Laufzeit von 2012 bis 2019. Viele Projekte stellen zentrale Forschungsergebnisse hier vor: Zlatkin-Troitschanskaia, O., Pant, H. A., & Greiff, S. (2019). Assessing Academic Competencies in Higher Education. [Special Issue]. *Zeitschrift für Pädagogische Psychologie, 33*(2).

Konstrukt interpretiert werden, könnte eine prüfbare Grundannahme lauten, dass die Testinhalte von Expert*innen als relevant für das interessierende Konstrukt eingestuft werden. Lautet die Testwertinterpretation jedoch, dass Testwerte auf zukünftiges Handeln schließen lassen, etwa Studienerfolg oder der mehr oder weniger erfolgreicher Umgang mit beruflichen Anforderungen, könnte eine Grundannahme lauten, dass die Testinhalte für das spezifische zukünftige Handeln von Bedeutung sind. Auf diese Weise könnten Testwerte interpretiert werden, auf deren Grundlage Studienplätze vergeben werden. Bislang bleiben jedoch oftmals sowohl Argumente für die Testnutzung als auch ein auf die Testwertinterpretation(en) abgestimmtes Validierungskonzept offen (Kuhn, Zlatkin-Troitschanskaia, Pant & Hannover, 2016). Ein Grund mag darin liegen, dass in der bisherigen Praxis der Testentwicklung die Validität eines Tests festgestellt wird, etwa durch die Angabe von Korrelationskoeffizienten zu anderen Testwerten oder externen Kriterien, um daraus die prognostische, konkurrente oder diskriminierende Validität zu schlussfolgern. Für den „validen Test“ werden dann mögliche Anwendungszwecke angegeben. Dies widerspricht explizit dem in dieser Arbeit verwendeten Validierungskonzept, nach dem ein Test nicht per se valide, sondern nur für einen oder mehrere Einsatzzwecke mehr oder weniger geeignet ist (AERA, APA & NCME, 2014; aber auch schon bei Messick, 1989).

Ein Ziel dieser Arbeit ist daher, ein Schema zur Abstimmung von Validierungskonzepten auf übliche Testnutzen im deutschen Hochschulsystem zu entwickeln. Das Schema soll für zukünftige Testentwicklungen als Handreichung dienen, Validitätsevidenzen besser auf den vorgesehenen Einsatzzweck eines Tests abzustimmen. Als Einstieg werden in den folgenden Kapiteln die verwendeten theoretischen Konzepte definiert. In Kapitel 2 wird der Kompetenzbegriff definiert und auf Besonderheiten bei der Erfassung studentischer Kompetenzen eingegangen. In Kapitel 3 wird das in dieser Arbeit verwendete Validitätsverständnis und die Bedeutung für Testentwicklung und Testnutzung beschrieben. Die Entwicklung des Validierungsschemas für standardisierte Tests zur Erfassung studentischer Kompetenzen wird in Kapitel 4 beschrieben.

Ein weiteres Ziel dieser Arbeit ist die exemplarische Nutzung des entwickelten Validierungsschemas. Dazu wird in Kapitel 5 das empirische Forschungsprojekt und die verwendeten Testinstrumente vorgestellt, dessen Testwertinterpretationen validiert werden sollen. Um das Schema möglichst vollständig anzuwenden, wird in den Kapiteln 6 bis 8 für alle im Forschungsprojekt verfolgten Testwertinterpretationen das Vorgehen der Validierung

beschrieben. Diese umfassen auch Forschungsfragen des Projekts, die von Projektkolleg*innen verantwortet wurden und werden. Diese Fremdleistungen werden explizit gekennzeichnet und es wird auf bereits veröffentlichte Schriften verwiesen, die unter Mitarbeit der Autorin im Projekt entstanden. Schwerpunkt dieser Arbeit bildet die Validierung der Testwertinterpretation in Kapitel 8.

Abschließend wird in Kapitel 9 die Anwendbarkeit des Schemas sowie die Nutzung des argumentationsbasierten Validierungsansatzes diskutiert.

2 Studentische Kompetenzen und deren Erfassung

In den folgenden Kapiteln geht es um studentische Kompetenzen. Dazu wird zunächst der Kompetenzbegriff definiert. Anschließend werden Ziele erläutert, die mit der Erfassung studentischer Kompetenzen verbunden sind. Abschließend wird kurz auf unterschiedliche Arten von Tests eingegangen, die zur Erfassung studentischer Kompetenzen genutzt werden.

2.1 Definition von Kompetenz

In der vorliegenden Arbeit wird Kompetenz auf die weitverbreitete Definition von Weinert (2001) zurückgegriffen:

> „Dabei versteht man unter Kompetenzen die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können … “ (S.27f).

Zum einen geht es um kognitive Aspekte, die zur Lösung eines Problems befähigen. Der *Problembezug* zeigt, dass Kompetenz in Relation zu einer Anforderungssituation steht. Eine „kompetente“ Person zeigt sich darin, dass sie die Anforderungssituation erfolgreich bewältigt. Um dieses kompetente Verhalten zeigen zu können, sind laut Weinert auch *motivationale, volitionale und soziale Aspekte* notwendig. Diese fasst er als Handlungskompetenzen zusammen, da sie beeinflussen, ob und wie vorhandenes Wissen und Fertigkeiten in einer Anforderungssituation zur Problembewältigung eingesetzt werden (Weinert, 2001).

Die Fähigkeiten und Fertigkeiten sind gemäß der Definition *verfügbar oder erlernbar*, das heißt, sie können einer Person ohne eigenes Zutun zur Verfügung stehen aber auch durch Training erworben sein. Dieses Kriterium ist insbesondere für die Verwendung des Kompetenzbegriffs im Bildungsbereich relevant. Hier steht die Vermittlung und Weiterentwicklung von Kompetenzen im Vordergrund. Sind Kompetenzen veränderlich und trainierbar, lassen sie sich gezielt durch Lerngelegenheiten in Bildungsinstitutionen entwickeln (Hartig & Klieme, 2006). Daher ist im Bildungsbereich die Identifikation von Kompetenzstrukturen wichtig. Sind die Einzelkomponenten, besser noch die Wirkweise dieser auf erfolgreiches Handeln in einer

Anforderungssituation bekannt, lassen sich Lerngelegenheiten zur Förderung einer Kompetenz entwickeln.

Nach Blömeke, Gustafsson und Shavelson (2015) wird in diesem Verständnis Kompetenz als nicht direkt beobachtbare kognitive, motivationale und affektive Personencharakteristik verstanden, welche Handeln in bestimmten Situationen vorhersagen. Das Handeln in einer Anforderungssituation stellt hier einen Indikator für Kompetenz dar. Blömeke et al. (2015) sehen dies als Endpunkt eines Kontinuums von zwei Perspektiven auf den Kompetenzbegriff. Der zweite Endpunkt des Kontinuums wird als Verständnis beschrieben, in dem Kompetenz als Handlung in einer relevanten Anforderungssituation selbst verstanden wird. Dieses Kompetenzverständnis mag häufiger in Auswahlentscheidungen oder Zertifizierung von Kompetenz vorherrschen. Hier zählt vorrangig, ob eine Person eine Anforderungssituation erfolgreich bewältigt oder nicht. Eher untergeordnet ist die Frage, wie die Kompetenz erworben wurde und welche Einzelkomponenten zum erfolgreichen Handeln beitragen.

2.2 Erfassung studentischer Kompetenzen

2.2.1 Rechtlicher Rahmen

Im *Hochschulrahmengesetz (HRG),* zuletzt geändert durch Art. 1 G v. 2019 I 1622 §7, sind die Ziele eines Studiums in Deutschland grundsätzlich geregelt:

> „Lehre und Studium sollen den Studenten auf ein berufliches Tätigkeitsfeld vorbereiten und ihm die dafür erforderlichen fachlichen Kenntnisse, Fähigkeiten und Methoden dem jeweiligen Studiengang entsprechend so vermitteln, daß er zu wissenschaftlicher oder künstlerischer Arbeit und zu verantwortlichem Handeln in einem freiheitlichen, demokratischen und sozialen Rechtsstaat befähigt wird.“

Die im HRG notwendigerweise breit gefassten Ziele eines Studiums lassen sich mit dem Verständnis von Kompetenz, wie in dieser Arbeit definiert, vereinbaren. Auch hier wird über die reine Vermittlung von Wissen und kognitiven Fertigkeiten hinaus formuliert. Die fachspezifischen Kenntnisse, Fähigkeiten und Methoden sollen den Studierenden nicht als Selbstzweck vermittelt werden, sondern um sie zu erfolgreichen Handeln im Beruf und als Individuum in einer demokratischen Gesellschaft zu befähigen.

2.2.2 Ziele von Kompetenzerfassung

Pellegrino, Chudowsky und Glaser (2001) unterscheiden drei grobe Ziele von (schulischer) Kompetenzerfassung:

1) Individuelle Leistung erfassen

Im regulären Studienbetrieb steht dieses Ziel von Kompetenzerfassung sicher im Vordergrund. Schon beim Eintritt in ein Studium wird die Erfassung individueller Leistung als Kriterium für die *Studienplatzvergabe* genutzt. Grundsätzlich gilt in Deutschland freie Berufswahl und eine freie Wahl von Ausbildungsstätten (Artikel 12 im Grundgesetz der Bundesrepublik Deutschland). Gibt es mehr Studieninteressierte als Plätze, müssen die Ressourcen möglichst fair verteilt werden. Zum Wintersemester 2020/21 waren ein Drittel aller grundständigen Studiengänge in Deutschland örtlich zulassungsbeschränkt mit einer Studienplatzvergabe mit Numerus Clausus (Hochschulrektorenkonferenz, 2020). Die dafür verwendete Note der Hochschulzugangsberechtigung spiegelt bisherige schulische Leistungen. Strittig ist jedoch, ob dies als vorherrschendes Kriterium für potenziellen Studienerfolg dienen kann. Zumindest für Studienplätze in medizinischen Studiengängen wurde die vorwiegende Nutzung der Hochschulzugangsberechtigung als Auswahlkriterium gerichtlich eingeschränkt (siehe Diskussion zur „Abiturbestenquote“ in Kapitel 3.2.1).

Als zentraler Nutzen von Tests im Hochschulbetrieb gilt die *Zertifizierung von Kompetenz* durch Prüfungen (Tremp & Eugster, 2006). In der Regel werden Prüfungen für Module oder einzelne Lehrveranstaltungen von den Personen erstellt, welche die entsprechende Lehrveranstaltung durchführen oder dafür verantwortlich sind.[5] Daraus ergibt sich auch, dass das Abschneiden in diesen Prüfungen wenig über Hochschulen oder auch nur über unterschiedliche Messzeitpunkte hinweg vergleichbar sind (vgl. Frey, Spoden, Fink & Born, 2020). Diesen Schluss legen auch Untersuchungen nahe, welche die Notengebung in einzelnen Fächern über Hochschulstandorte und die Veränderung von Notengebung im Längsschnitt vergleichen (Tsarouha, 2019). Eine Ausnahme bildet hier das Projekt „Kat-HS“, in dem ein Test entwickelt wurde, der als Klausur für

[5] An der Johann-Wolfgang Goethe Universität Frankfurt am Main z.B. geregelt in der Rahmenordnung für gestufte und modularisierte Studiengänge (UniReport Satzungen und Ordnungen, vom 22. Dezember 2020). Im Fachbereich 05 der Johann-Wolfgang Goethe Universität Frankfurt am Main ist dies in §21 (4) in der aktuell gültigen Prüfungsordnung vom 14. Juli 2020 für den Bachelorstudiengang Psychologie mit dem Abschluss „Bachelor of Science (B.Sc.)“ umgesetzt (UniReport Satzungen und Ordnungen, vom 08. September 2020).

eine Lehrveranstaltung zur Einführung in Forschungsmethoden verwendet wird (Frey, Spoden & Born, 2020). Der Test erfasst studentische Leistung kriterienorientiert, das bedeutet, dass Noten nicht im Verhältnis zur Leistung anderer Studierender vergeben werden (normorientiert), sondern den Umfang bewertet, in dem die für die Lehrveranstaltung definierten Lernzielen erreicht wurden. Die Testwerte werden mit IRT-Modellen geschätzt und über Semester vergleichbar gemacht. Eine hohe Vergleichbarkeit studentischer Leistungen wird auch in Studiengängen angestrebt, die auf durch den Staat reglementierte Berufsfelder hinführen und daher staatlichen Qualitätsstandards genügen müssen. Dazu zählen Studiengänge in Jura, Medizin und Pharmazie und Lehramt. Gemeinsam ist diesen Studiengängen, dass es eine Abschlussprüfung nach der universitären Ausbildung (umgangssprachlich „1. Staatsexamen“) und eine weitere nach einer praktischen Ausbildung (umgangssprachlich „2. Staatsexamen“) gibt. Den höchsten Standardisierungsgrad von Abschlussprüfungen lässt sich im deutschen Hochschulbereich beim Medizinstudium ausmachen. In der Approbationsordnung für Ärzte (Abk. ÄApprO; Bundesministerium für Gesundheit) wird geregelt, welche Inhalte Studierende während ihres Studiums behandelt haben müssen, um zur Prüfung zugelassen zu werden. Weiter wird beschrieben, welche Inhalte in welchen Prüfungen behandelt werden müssen. Zudem sind die Aufgaben in den Abschlussprüfungen für alle Teilnehmenden gleich (§14 ÄApprO 2002). Daher müssen die schriftlichen Tests an bundesweit einheitlichen Terminen stattfinden. Dementsprechend werden die Prüfungsaufgaben nicht von einzelnen Lehrpersonen sondern von einer staatlich beauftragten Institution entwickelt (seit 1972 vom „Institut für medizinische und pharmazeutische Prüfungsfragen“). Zumindest für die schriftlichen Tests[6] lässt sich eine hohe Vergleichbarkeit studentischer Leistungen über Hochschulstandorte hinweg feststellen.

Ein weiterer Zweck von Tests zur Erfassung individueller Leistung ist die *Zuteilung von Studierenden auf Lerngelegenheiten passend zu ihrem Kompetenzstand.* Dies geschieht etwa regelmäßig bei sprachlichen Einstufungstests, deren Ergebnisse genutzt werden um Studierende in Sprachkurse mit unterschiedlichem Niveau einzuteilen.

Darüber hinaus werden Maße individueller Leistung zu *Selektionszwecken* genutzt. Dies geschieht etwa bei der Vergabe von Studienplätzen eines weiterführenden Studiengangs nach

[6] Bei zwei von insgesamt drei Abschlussprüfungen wird auch mündlich-praktisch geprüft. Die Prüfungskommission für die mündlichen Prüfungen setzt sich hauptsächlich aus der Hochschule angehörenden Lehrpersonen zusammen, für die dritte Prüfung können auch außerhochschulische Ärzte der Prüfungskommission angehören.

Abschlussnoten des grundständigen Studiengangs. In diesem Fall werden die im Laufe eines Studiums angesammelten (gewichteten) *zertifizierten Kompetenzen* als Kriterium verwendet.

2) Lernen erleichtern

Spätestens seit dem Umsetzen der Bologna-Reform und der damit einhergehenden Modularisierung von Studiengängen sollen Tests im Hochschulbetrieb nicht nur der regulären Leistungsüberprüfung dienen. Die Studierenden sollen auch Rückmeldung über ihren aktuellen Lernstand und gegebenenfalls Informationen zu Entwicklungsfeldern erhalten (Schaper & Hilkenmeier, 2013). Da es zur Prüfungspraxis an deutschen Hochschulen allgemein wenig Forschung gibt (Schindler, 2015), ist auch wenig darüber bekannt, in wie fern die Prüfungspraxis in Hochschulen diesen Zweck erfüllen. Für externe Tests lässt sich festhalten, dass im Rahmen des DFG Schwerpunktprogramms *Kompetenzmodelle* Instrumente entwickelt und geprüft wurden mit dem Ziel, Individuen Rückmeldung über ihren Lernstand zu geben und die zur Förderung individueller Lernergebnisse genutzt werden können (Kunina-Habenicht, Wilhelm, Matthes & Rupp, 2010).

3) Evaluation von Programmen

Verbesserung von Lehre durch Evaluation. Prinzipiell sind Evaluationen von Lehrveranstaltungen über die Erfassung studentischer Kompetenzen möglich, häufig erfolgt dies jedoch über Einschätzungen zur Zufriedenheit mit der Lehrveranstaltung (Braun, Gusy, Leidner & Hannover, 2008). Braun et al. (2008) entwickelten ein Instrument zur Lehrevaluation, welches auf Kompetenzzuwächsen der Studierenden beruht. Die Kompetenzzuwächse der Studierenden werden jedoch über Selbsteinschätzungen erfasst und nicht über Veränderungsmessung der in einem Test gezeigten Kompetenz. In der Qualitätsoffensive Lehrerbildung ist ein Themenbereich „Forschung in der Lehrkräftebildung“. Als zentrale Fragestellung wird darin die Wirksamkeit von Lehrkräftebildung untersucht, um wissenschaftlich gesicherte Befunde zu erfolgreichem Handeln im Unterricht zu generieren.[7]

Aussagen über Bildungssysteme ermöglichen. Während im Schulbereich standardisierte Tests als Instrument zur Qualitätsentwicklung gängig sind (z.B. VERA zur Überprüfung der Erreichung

[7] Beispielsweise wurde untersucht, welche Lernmethoden Lehramtsstudierenden dabei helfen, zukünftig Schüler*innen auf Anforderungen eines Lebens in Smart Cities vorzubereiten (Dorsch, 2018) oder wie die Verzahnung von Theorie- und Praxisphasen für angehende Englisch-Lehrkräfte besser gelingen kann, um deren diagnostische Kompetenz zu stärken (Kemmerer, 2019).

von Lernzielen), ist dies im Hochschulbereich bislang nicht üblich. Eine Machbarkeitsstudie zur Erfassung studentischer Kompetenzen im internationalen Vergleich wurde von der OECD durchgeführt (OECD, 2013). In dieser Studie wurden Tests zu generischen, ingenieurswissenschaftlichen und wirtschaftlichen Kompetenzen entwickelt und geprüft, die Leistungen von Studierenden aus unterschiedlichen Staaten vergleichbar machen sollen. Untersuchungen zur Messinvarianz zeigen jedoch, dass sich die Itemschwierigkeiten nicht nur zwischen Staaten, sondern auch innerhalb eines Staates zwischen Institutionen häufig unterscheiden. In allen Tests traten Items auf, deren Schwierigkeit in Abhängigkeit der Größe, des höchsten verliehenen Abschlusses oder der Forschungsorientierung einer Institution variierten. Dies schränkt das eigentliche Ziel der Studie ein, studentische Leistungen vergleichbar zu machen um darüber Rückschlüsse auf Effektivität von Hochschulinstitutionen ziehen zu können. Darüber hinaus nahm an der Machbarkeitsstudie keine Hochschulinstitution aus Deutschland teil.

2.3 Unterscheidung von Tests zur Erfassung studentischer Kompetenzen

Die im vorherigen Kapitel aufgeführten Testbeispiele für die einzelnen Zwecke studentischer Kompetenzerfassung lassen sich nach verschiedenen Kriterien unterscheiden.

Zunächst finden einige Tests systemimmanent statt, sind also Teil des regulären Studienbetriebs. Dazu zählen etwa mündliche Abfragen, Klausuren oder Abschlussprüfungen, die als *interne* Tests bezeichnet werden. Tests, die nicht Teil des regulären Studienbetriebs sind werden *externe* Tests genannt. Beispiele aus dem deutschen Schulbereich sind PISA, VERA oder der IQB-Ländervergleich. Interne und externe Tests lassen sich auf einem Kontinuum der „Nähe zum Unterricht“ anordnen, wobei interne Tests typischerweise näher am Unterricht konzipiert sind als externe Tests (Ruiz-Primo, Shavelson, Hamilton & Klein, 2002).

Ruiz-Primo et al. (2002) argumentieren, dass unterschiedliche Formen von Tests notwendig sind, um verschiedene Ziele von Kompetenzerfassung zu bedienen. So scheinen Tests, die *nahe* bis *proximal* am Unterricht orientiert sind, effektiver über Lernfortschritte von Studierenden zu informieren, als *distale* und *ferne* Tests[8]. Dies wird auf die bessere Abstimmung – *alignment* – von Assessment und Unterrichtsgeschehen bei nahen Assessments zurückgeführt. Den Aspekt

[8] *Anmerkung*. Eigene Übersetzung. Im Original: *close*, *proximal*, *distal*, *remote* (Ruiz-Primo et al., 2002).

der Abstimmung von Test auf Unterrichtsgeschehen findet sich auch bei Pellegrino et al. (2001). Diese fordern neben der Abstimmung auf Unterrichtsgeschehen die Berücksichtigung curricularer Vorgaben: „[E]ducational assessment does not exist in isolation, but must be aligned with curriculum and instruction if it is to support learning“ (Pellegrino et al., 2001, S. 3). [9]

Gleichzeitig wird aus der Klassifikation unmittelbar ersichtlich, dass sich Tests, die mit hoher Ähnlichkeit zu einer bestimmten Lehreinheit erstellt wurden, nicht eignen, um Kompetenzen von Studierenden auf nationaler oder internationaler Ebene vergleichen lassen. Aussagen über Bildungssysteme lassen sich mit Ergebnissen aus diesen Tests und daraus generierten Kennzahlen (z.B. Absolventenquoten, durchschnittliche Abschlussnoten, durchschnittliche Studienzeit, etc.) daher auch nicht ableiten. Um Aussagen über Bildungssysteme treffen zu können, bedarf es demnach Tests, die studentische Kompetenzen über Hochschulstandorte hinweg vergleichbar machen.

Unabhängig davon, ob ein interner oder externer Test verwendet wird: Viele davon werden zu Zwecken verwendet, die für Studierende, Lehrpersonen oder Entscheidungsträger*innen mit Konsequenzen verbunden sind. Für Studierende beeinflusst die Notenvergabe die Erfolgsaussichten bei Bewerbungen für Arbeitsstellen oder weiterführenden Studiengängen. Lehrpersonen entscheiden sich aufgrund von Evaluationen für oder gegen eine bestimmte Form von Lehrveranstaltung und Entscheidungsträger*innen aus der Bildungsverwaltung oder Politik können Testergebnisse als Grundlage für Entscheidungen über Entwicklungen im Hochschulsystem nutzen. Daher sollten die verwendeten Tests bestimmten Gütekriterien genügen.

Wissenschaftliche Tests sind solche, die bestimmte Kriterien bei der Testentwicklung und der Datenerhebung erfüllen und für die Testeigenschaften untersucht und festgehalten werden (Moosbrugger & Kelava, 2020). Für interne Tests lässt sich vermuten, dass viele davon nicht wissenschaftlich überprüft wurden. Zumindest gibt es kaum Erkenntnisse darüber, welche

[9] Die Abstimmung des Tests auf Curriculum und Lehre gilt nicht nur für die Inhalte. Auch das Itemformat kann mehr oder weniger gut geeignet zur Erfassung einer bestimmten Kompetenz sein. So fordern etwa Shavelson, Ruiz-Primo und Wiley (2005) ein *Performance Assessment* statt der Abfrage von Wissensinhalten, wenn strategisches Wissen erfasst werden soll.

Kriterien Hochschullehrende und –prüfende bei der Prüfungserstellung und der Leistungsbewertung anlegen (Schindler, 2015).[10]

Zu den zentralen Gütekriterien von Tests zählen Objektivität, Reliabilität und Validität (Moosbrugger & Kelava, 2020).[11] Während die Objektivität hauptsächlich durch Planung und entsprechende Vorbereitung erreicht werden kann, stellen Reliabilität und Validität die Hauptmerkmale wissenschaftlicher Untersuchungen dar.

Objektivität bedeutet, dass Testwerte unabhängig von der Person sind, die den Test durchführt bzw. auswertet. Sie soll meist durch die Standardisierung von Prozessen erreicht werden. Dazu gehören Abläufe der Testung an sich (z.B. einheitliche Anweisungen, Bearbeitungszeiten und räumliche Gegebenheiten) aber auch Regeln zur standardisierten Kodierung und zum Scoring von Antworten. Das Ziel dieser Standardisierung ist, allen Testpersonen die gleiche Chance zu geben, ihre Fähigkeiten im Test zu demonstrieren (AERA et al., 2014).

Reliabilität beschreibt die Messgenauigkeit eines Instruments, genauer den Anteil der wahren Varianz von Testwerten (bzw. der Anteil der wahren Varianz in den Personenparametern in der Item-Response-Theorie; Rose, 2020) im Verhältnis zur gesamten Varianz (Gäde, Schermelleh-Engel & Werner, 2020). Je reliabler ein Instrument, desto höher ist der Anteil der wahren Varianz und umso geringer ist der Anteil der Varianz der Messfehler. Die Reliabilität wird im Rahmen der klassischen Testtheorie häufig über die Stabilität von Testwerten bestimmt, also wie zuverlässig die Testwerte über verschiedene Messungen oder unterschiedliche Testformen sind (vgl. AERA et al., 2014). Bekannte Maße sind etwa die interne Konsistenz und die Retest-Reliabilität. Die interne Konsistenz ist das Ausmaß der Interkorrelation der Items einer Skala und wird interpretiert als Maß dafür, ob Items einer Skala das gleiche messen. Die Retest-Reliabilität ist die Korrelation zwischen Testwerten, die mit dem gleichen Test zu unterschiedlichen Zeitpunkten erfasst wurden, und wird als Stabilität der Testwerte bei wiederholter Messung

10 Schindler (2015) verweist auf zwei Studien aus dem deutschen Hochschulbereich. Eine qualitative Arbeit untersucht die Abstimmung von Prüfungsanforderungen und Lehrzielen aus Modulbeschreibungen für Module zu höherer bzw. Angewandter Mathematik I / II aus Maschinenbaustudiengängen von drei Hochschulstandorten (Ștefănică, 2013). Eine weitere Publikation identifiziert aus Studierendenbefragungen Cluster von Prüfungsanforderungen an der Technischen Universität München, die zwischen den Polen „wiedergabeorientiert“ und „anforderungsorientiert“ liegen (Schulz, Zehner, Schindler & Prenzel, 2014). Beide Arbeiten untersuchen nur wenige Prüfungspraxen ausgewählter Veranstaltungen in einzelnen Hochschulen.

11 Verschiedene Organisationen veröffentlichen weitere Gütekriterien von Tests und Standards zur Testentwicklung und –anwendung, z.B. für Tests zur berufsbezogenen Eignungsbeurteilung (DIN, 2016) oder die *Standards for Educational and Psychological Testing (AERA, APA & NCME, 2014).*

interpretiert. Da die Messgenauigkeit in der Item-Response-Theorie (IRT) von der Testinformationsfunktion abhängt, werden hier durchschnittliche Reliabilitätsmaße über die Verteilung der geschätzten Personenparameter angegeben (vgl. Rose, 2020).

Objektivität und Reliabilität bilden eine notwendige Voraussetzung für Validität (Moosbrugger & Kelava, 2020). Nur wenn Testwerte objektiv und reliabel erfasst wurden, können diese zwischen unterschiedlichen Personen und unterschiedlichen Messungen verglichen werden. Diese Vergleichbarkeit ist jedoch irrelevant, wenn sich herausstellt, dass ein Test nicht *valide* ist, also nicht das misst, was er messen soll. Daher wird Validität als das wichtigste Gütekriterium eines Tests betrachtet. Aufgrund der besonderen Bedeutung der Validität für diese Dissertation werden im folgenden Kapitel unterschiedliche Perspektiven auf Validität und das in dieser Arbeit verwendete Validitätskonzept vorgestellt.

3 Validität

Im vorherigen Kapitel wurde Validität eingeführt als Frage danach, ob ein Test misst, was er eigentlich messen soll. In der Psychologie und in der Bildungsforschung interessieren meist Merkmale, Kompetenzen oder Einstellungen von Lernenden, die nicht direkt beobachtbar sind, sogenannte latente Konstrukte. Um Rückschlüsse über das Merkmal zu ziehen, beobachtet man stattdessen Indikatoren, von denen man auf das latente Merkmal schließt. Daher ist die Frage, ob ein Test misst, was er messen soll, nicht einfach zu beantworten.

Über das angemessene Vorgehen bei der Validierung eines Tests wird seit Jahrzehnten diskutiert. Der folgende Abschnitt gibt – ohne Anspruch auf Vollständigkeit - einen Überblick über die historische Perspektive auf Validität. Anschließend wird das aktuelle Validitätsverständnis nach AERA et al. (2014) beschrieben, welches dieser Dissertation zugrunde liegt.

3.1 Historische Perspektive

Ab den 1950er Jahren war die Frage dominant, ob der Test das interessierende Konstrukt (z.B. Intelligenz) abbildet. Zur Beantwortung dieser Frage konnten verschiedene Arten von Validität betrachtet werden, die in den *Technical recommendations for psychological tests and diagnostic techniques* (American Psychological Association, 1954) als Inhaltsvalidität, konkurrente Validität, prädiktive Validität und Konstruktvalidität beschrieben sind. *Inhaltsvalidität* bedeutet, dass die Testitems als repräsentative Stichprobe des „Itemuniversums“ gelten. Diese Validitätsart ist für Tester wichtig, die an den Inhalten (oder Verhaltensweisen) interessiert sind, die das Testverhalten auslösen. *Prädiktive* und *konkurrente Validität* werden unter kriterienorientierter Validität subsumiert. Hier spielt die Vorhersage des Kriteriums eine zentrale Rolle, der konkrete Testinhalt ist von nachrangiger Bedeutung. Bei konkurrenter Validität werden Kriterium und Merkmal gleichzeitig erfasst, bei prädiktiver Validität wird das Kriterium vor dem Merkmal erfasst. Beides wird durch einen Korrelationskoeffizienten zwischen Kriterium und Merkmal angegeben.

Dem in den *Technical recommendations* neu eingeführten Begriff *Konstruktvalidität* gehen Cronbach und Meehl (1955) weiter nach. Ihnen zufolge steht bei Konstruktvalidität die Frage im Mittelpunkt, welches Konstrukt Varianz im Testverhalten erklärt. Cronbach und Meehl (1955) zufolge muss Konstruktvalidität untersucht werden, wenn es kein direktes Maß für das

interessierende Konstrukt gibt und fraglich ist, ob ein Test als Maß für das Konstrukt interpretiert werden kann. Dies ist bei den Formen der kriterienorientierten Validität dann der Fall, wenn es kein geeignetes Kriterium gibt, gegen das der Test geprüft werden kann. Bei der Inhaltsvalidität muss Konstruktvalidität untersucht werden, wenn das Itemuniversum als nicht definiert angesehen werden muss. Daher betrachten die Autoren Konstruktvalidität als wichtigstes Validitätskriterium.

Unter Konstrukt verstehen Cronbach und Meehl (1955) ein postuliertes Attribut von Menschen, dass sich im Testverhalten widerspiegeln soll (S. 283). Eine numerische Entsprechung der Konstruktvalidität wäre demnach ein Koeffizient der ausdrückt, welcher Varianzanteil im Testwert durch das Konstrukt hervorgerufen wird (S. 289). Dabei wird angenommen, dass die Person zu jeder Zeit das Attribut / die Struktur besitzt oder nicht besitzt (qualitative Ausprägung) oder ein bestimmtes Ausmaß des Attributs aufweist (quantitative Ausprägung).

Um die Konstruktvalidität eines Tests zu untersuchen, schlagen Cronbach und Meehl (1955) ein nomologisches Netzwerk vor. Aus diesem können Hypothesen bezüglich des interessierenden Konstrukts abgeleitet und überprüft werden. Ein nomologisches Netzwerk ist ein System an Gesetzen, in dem das Konstrukt existiert. Das System von Gesetzen bildet die Theorie um das Konstrukt, das erfasst werden soll. Die Gesetze können beobachtbare Größen untereinander verbinden, theoretische Konstrukte mit beobachtbaren Größen verbinden oder theoretische Größen untereinander verbinden. Um in der Wissenschaft anwendbar zu sein, ist eine notwendige Voraussetzung für ein nomologisches Netzwerk, dass zumindest ein Teil der Gesetze beobachtbare Größen beinhalten („… unless the network makes contact with observations, and exhibits explicit, public steps of inference, construct validation cannot be claimed“ S. 291).

Entsprechend dem Forschungsstand zu einer Theorie kann ein nomologisches Netzwerk unterschiedlich stark definiert sein. Je besser es definiert ist, desto stärker und präziser wird das interessierende Konstrukt definiert. Negative Evidenzen aus der Überprüfung der Gesetzmäßigkeiten („claims“) können bedeuten, dass der Test nicht das gewünschte Konstrukt misst oder das experimentelle Design nicht für die Überprüfung der Hypothesen angemessen war. Vor allem bei schwachen nomologischen Netzwerken sollten negative Evidenzen aber auch als Hinweis auf eine Fehlspezifikation betrachtet werden. Gleichzeitig können neue Beobachtungen dazu dienen, das nomologische Netzwerk zu verfeinern und zu erweitern. Änderungen sind insbesondere dann sinnvoll, wenn dadurch die Anzahl der Gesetzmäßigkeiten reduziert wird, die für Vorhersagen notwendig sind.

Die Herausforderung der Erstellung eines nomologischen Netzwerks besteht darin, empirisch prüfbare Gesetzmäßigkeiten theoretisch fundiert abzuleiten. Vor allem bei schwachen nomologischen Netzwerken stellt sich bei negativen Evidenzen zu den Gesetzmäßigkeiten die Frage, ob der Test oder das nomologische Netzwerk fehlspezifiziert sind.

Ab den 1970er Jahren verschiebt sich der Fokus weg von der Frage, ob ein Test misst, was er messen soll, hinzu „Welche Aussagen erlauben die Testwerte eines Tests?“ (vgl. Cronbach, 1971; Cronbach, 1986). Explizit wird in den Standards (AERA, APA & NCME, 1985) und bei Messick (1989) der Fokus auf den Testzweck und die Interpretation der Testwerte gelegt. Es entwickelt sich ein argumentationsbasierter Validierungsansatz, der dem Forschungsprozess von Hypothesenbildung über die Falsifikation/Verifikation der Hypothesen hin zu einem vorläufigen Erkenntnisschluss entspricht (AERA et al., 2014; Kane, 1992, 2013).

3.2 Aktuelles Validitätskonzept

Ab den 1980er Jahren entwickelt sich ein argumentationsbasierter Validitätsansatz. In diesem wird nicht die Validität eines Tests untersucht, sondern die mit einem Testwert verbundene Interpretation auf ihre Gültigkeit hin überprüft. Nach Kane (1992) lassen sich praktische/ interpretative Schlussfolgerungen nicht endgültig als wahr oder falsch, also valide oder invalide beurteilen. Demnach kann eine Testwertinterpretation, bzw. die Argumentation für diese Interpretation (das sogenannte „interpretative argument“), nur mehr oder weniger plausibel sein. Testwertinterpretationen weisen nach Kane folgende Eigenschaften auf:

1) Testwertinterpretationen sind Artefakte, sie ergeben sich nicht aus einem Test heraus, sondern werden von Testentwickler*innen oder Testnutzer*innen einem Testwert zugewiesen.

2) Sie sind dynamisch. Durch neue Ergebnisse können interpretative Argumente weiter oder enger gefasst werden, anspruchsvollere Methoden können ein besseres Verständnis für Testwerte oder neue Evidenzen liefern. Veränderungen in der Gesellschaft können neue Interpretationen notwendig machen.

3) Testwertinterpretationen müssen möglicherweise für einzelne Gruppen mit spezifischen Besonderheiten oder für besondere Umstände angepasst werden, etwa für Substichproben mit bestimmten körperlichen Einschränkungen (Lesbarkeit) oder für Testumstände (Lärm, Hitze).

4) Testwertinterpretationen werden nach ihrem Ausmaß an Plausibilität beurteilt, nicht in einem einfachen wahr-falsch Urteil.

Die Plausibilität einer Testwertinterpretation wird nach Kane (1992, 2013) anhand der Klarheit der Testwertinterpretation beurteilt. Die Argumentation der Interpretation muss klar benannt sein und die Details der Annahmen sollen expliziert werden. Dieser Schritt ist für Kane der wichtigste, da für „versteckte“ Annahmen von Testwertinterpretationen keine Evidenzen gesucht werden, welche die Annahmen stützen oder dieser widersprechen könnten. Außerdem müssen die Schlussfolgerungen aus den spezifizierten Annahmen nachvollziehbar sein, die Testwertinterpretation soll also kohärent sein. Drittes Kriterium ist schließlich die Plausibilität der Annahmen: Werden die Annahmen durch empirische Evidenz unterstützt oder sind die Annahmen natürlich plausibel?

Die argumentationsbasierte Validierung von Testwertinterpretation ist in den aktuellen *Standards for Educational and Psychological Testing* (AERA et al., 2014) vertreten. Die interpretative Argumentation selbst erfordert keinen spezifischen Aufbau oder bestimmte Methoden. Wichtig ist vielmehr, dass die Validitätsevidenzen auf die jeweilige Testwertinterpretation abgestimmt sind und die im Argument spezifizierten Annahmen empirisch gestützt werden.[12]

Im Vergleich zur historischen Perspektive auf Validität legt das aktuelle Validitätskonzept den Fokus auf die Anwendung von Tests. Das Vorgehen in beiden Konzepten ist ähnlich. Bei Cronbach und Meehl (1955) werden aus den im nomologischen Netzwerk definierten Gesetzmäßigkeiten Hypothesen abgeleitet („claims“). Im argumentationsbasierten Validierungsansatz werden Annahmen identifiziert, welche der gewünschten Testwertinterpretation zugrunde liegen. Für beides werden Evidenzen gesucht, welche die Behauptungen stützen oder widerlegen können. Der argumentationsbasierte Validierungsansatz grenzt das interessierende Merkmal („Konstrukt“ im Sinne von Cronbach & Meehl, 1955) jedoch auf bestimmte Anwendungsfälle ein. Es wird also nicht die Validität eines Tests untersucht, sondern die Interpretation eines Testwerts in vorab definierten Testnutzen.

[12]In den Standards for Educational and Psychological Testing werden Maße zur internen Struktur, Antwortprozesse, Maße zur Angemessenheit des Testinhalts und Zusammenhänge zu anderen Variablen als Evidenzquellen benannt. Schon Cronbach und Meehl (1955) nennen als konkrete Methoden u.a. die Identifikation von Gruppenunterschieden, Korrelationsmatrizen und Faktoranalysen um Konstruktvalidität zu untersuchen; Interkorrelationen von Items und Reliabilität für die Analyse der internen Struktur; oder die Untersuchung von Antwortprozessen der Testpersonen.

3.2.1 Konsequenzen von Testeinsätzen

Das beschriebene Validitätskonzept wurde in den USA entwickelt und bezieht sich explizit auf den Bildungsbereich. Da Tests dort vielfach eingesetzt werden um Ressourcen zu verteilen, hat sich schon früh die Perspektive auf die Konsequenzen eines Testeinsatzes geweitet. Cronbach (1986) beschreibt etwa einen Abschlusstest für High Schools in einem Staat der USA, der für die Bewerbung von Colleges (Universitäten) genutzt wurde, gegen dessen Einsatz eine Absolventin klagte weil sie der Meinung war, dass ihre High School sie nicht ausreichend auf diesen Test vorbereitet hatte. Zu einem Validitätsargument gehört deshalb auch die Untersuchung von nicht intendierten Konsequenzen des Testeinsatzes.

Ein Beispiel aus dem deutschen Hochschulbereich soll das Verdeutlichen. Szenario: Ein Test soll eingesetzt werden, um Studierende in unterschiedliche Englischkurse einzuteilen. Die Kurse werden bei erfolgreichem Bestehen auf bestimmten Levels zertifiziert. Höhere Levels in Bewerbungen werden von ausländischen Hochschulen oder Arbeitgebern mit internationaler Ausrichtung positiv bewertet.

Eine Testwertinterpretation des Tests könnte lauten, dass Studierende mit einem bestimmten Testwert über ausreichend Kompetenz verfügen, um einem Kurs der nächst höheren Kompetenzstufe folgen zu können. Dieser Interpretation würden dann (mindestens) die Annahmen zugrunde liegen, dass 1) bestimmte Kompetenzstufen erreicht werden müssen um einem Kurs folgen zu können, 2) der Testwert kriterienorientiert verankert ist und 3) der Test auf die Lehrinhalte der Sprachkurse abgestimmt ist.

Kognitive Interviews mit Testpersonen könnten Evidenzen zur ersten Annahme aus dem Bereich „Antwortprozesse“ liefern. Wenn Studierende in kognitiven Interviews bei der Aufgabenbearbeitung berichten, dass sie die Aufgaben aufgrund des theoretisch erwarteten Vorwissens bearbeiten, stützt dies Annahme 1. Weitere Evidenz könnten Expertenurteile aus dem Bereich „Inhalt“ liefern. Denkbar sind etwa Einstufungen von Sprachdidaktikern, welche Kompetenzen in welchem Sprachniveau beherrscht werden. Annahme 2 könnte durch die Vorhersage von Aufgabenschwierigkeiten durch bestimmte Itemmerkmale gestützt werden, diese Evidenz wäre dem Bereich „interne Struktur“ zuzuordnen. Annahme 3 könnte wiederum durch Expertenurteile zur Passung von Test und Item gestützt werden.

Die gesammelten Evidenzen würden dann zu einem Gesamturteil über die Plausibilität der Testwertinterpretation zusammengefasst werden. Wird der Schluss gezogen, dass die

Testwertinterpretation (vorläufig!) plausibel ist, könnten die Testwerte dazu genutzt werden, Studierende auf Sprachkurse mit unterschiedlichen Sprachniveaus zuzuweisen.

Zusätzlich sollte untersucht werden, ob der Testeinsatz unerwünschte Konsequenzen haben kann. Denkbar sind z.B. Vorteile für Muttersprachler*innen der Testadministrationssprache oder Übungseffekte der Testformate, die zu besseren Testwerten bei den Studierenden führen, die mehr Vorbereitungszeit für den Test aufwenden. So könnte die Höhe der Testwerte im englischen Sprachverständnis beispielsweise vom Deutsch-Leseverständnis abhängen. Damit würde ein konstruktirrelevantes Merkmal die Testwerte beeinflussen und bestimmte Gruppen systematisch benachteiligen. Die Schwierigkeit bei der Frage nach den Testkonsequenzen stellt die Fülle an potentiellen Merkmalen und Kompetenzen dar, die zu Benachteiligungen führen könnten. Grundsätzlich muss nicht jedes Gruppenmerkmal, welches die Testwerte beeinflusst, eine Benachteiligung darstellen. Wäre im Beispiel das Deutsch-Leseverständnis Teil der Konstruktdefinition (weil der Test z.B. eingesetzt wird bei Übersetzer*innen von Deutsch nach Englisch), sollten die Testwerte im Englisch-Sprachverständnis gerade nicht unabhängig davon sein. Wichtig ist also die Unterscheidung zwischen konstruktrelevanten und –irrelevanten Merkmalen, welche die Testwerte beeinflussen.

Wann genau nicht intendierte Konsequenzen eines Testeinsatzes zu vernachlässigen oder nicht mehr akzeptabel sind, ist sicher vom Einzelfall abhängig und Teil eines gesellschaftlichen Verständigungsprozesses. Im Beispiel bei Cronbach (1986), wurde der Testeinsatz gerichtlich untersagt, bis die Schulen des Bundesstaates nachwiesen, dass alle Schüler*innen ausreichend Lerngelegenheiten erhielten. Im Beispiel mit dem Testeinsatz für die Zuweisung zu Sprachkursen könnte man argumentieren, dass hochschulische Sprachkurse in jedem Semester und kostenlos – oder, im Vergleich zu außerhochschulischen Sprachkursen, relativ günstig – angeboten werden. Selbst bei einer möglichen Benachteiligung der Nicht-Muttersprachler*innen könnten diese während des Studiums weitere Sprachkurse belegen. Die unerwünschten Konsequenzen wären also gering (das nächsthöhere Sprachniveau kann erreicht werden, allerdings müssen Nicht-Muttersprachler*innen dafür möglicherweise mehr Kurse belegen).

Ein anderes Beispiel aus dem deutschen Bildungssystem mit schwerwiegenderen Konsequenzen ist die Zulassungsbeschränkung von Studiengängen. 2017 bemängelte das Bundesverfassungsgericht die teilweise verfassungswidrige Vergabe von Studienplätzen für Humanmedizin (BVerfG, Urteil des Ersten Senats 1 BvL 3/14, Rn. 1-253) die sich auf die vorrangige Berücksichtigung der Abiturnote stütze, welche über die Bundesländer hinweg nicht

vergleichbar seien. Seit 2020 gilt nun ein neues Zulassungsverfahren für das Medizinstudium in Deutschland (vgl. Stiftung für Hochschulzulassung, 2020). Für die Abiturbestenquote, über die 30% der Studienplätze vergeben werden, werden die Studienbewerber*innen anhand ihrer Abiturnote zunächst innerhalb der Bundesländer in eine Rangfolge gebracht. Anschließend werden die Länderreihenfolgen in eine Bundesreihenfolge gebracht, bei der der Anteil des Landes an der Gesamtbewerberzahl sowie der Anteil der an der Gesamtzahl der Achtzehn- bis unter Einundzwanzigjährigen in der Bevölkerung berücksichtigt wird. Dieses Verfahren soll die Chancengerechtigkeit bezüglich der Vergleichbarkeit von Abiturnoten verbessern. Im Sprachgebrauch der Validierung werden nicht intendierte Konsequenzen für Abiturient*innen aus Bundesländern mit höheren schulischen Anforderungen vermieden.

3.2.2 Bedeutung des Validierungskonzeptes für Testentwicklung und Testeinsatz

Ausgehend vom Validitätskonzept der aktuellen Standards ergeben sich für die Testentwicklung von Kompetenztests vor allem zwei Anforderungen. Zunächst müssten Testentwickler*innen zu Beginn der Testentwicklung definieren, wie die Testwerte interpretiert werden sollen und zu welchen Nutzen der Test verwendet werden kann. Ein Test kann auch für mehrere Testnutzen verwendet werden. Je mehr Einsatzzwecke ein und derselbe Test bedienen soll, desto wahrscheinlicher ist jedoch, dass er keinen einzigen Zweck wirklich gut erfüllt (AERA et al., 2014; S. 188).

Zweitens muss das Validierungskonzept auf die definierte Testwertinterpretation abgestimmt sein. Sind für einen Test mehrere Testwertinterpretationen definiert, muss jede davon durch ein eigenes Validitätsargument mit spezifischen Evidenzen gestützt werden. Testanwender*innen wiederum sind dafür verantwortlich, einen Test nur für die Anwendungssituationen einzusetzen, die sich aus den validierten Testwertinterpretationen ergeben (AERA et al., 2014).

Gleichzeitig können neue Forschungsergebnisse oder gesellschaftliche Entwicklungen das Validitätsargument einer Testwertinterpretation schwächen und weitere Evidenzen oder eine Änderung der Testwertinterpretation erforderlich machen.

Modell zur Einordnung bisheriger Testwertinterpretationen bei Testentwicklungen im deutschen Hochschulsystem

4 Modell zur Einordnung bisheriger Testwertinterpretationen bei Testentwicklungen im deutschen Hochschulsystem.

Das in dieser Arbeit verwendete Validitätskonzept fordert für jeden Zweck, für den ein Test eingesetzt werden kann (= *Testnutzen*), eine spezifische Validierung der damit einhergehenden Testwertinterpretation (AERA et al., 2014). Obwohl in Deutschland viele Publikationen zu Tests zur Erfassung studentischer Kompetenzen vorliegen, stellten Kuhn et al. (2016) fest, dass für die meisten dieser Testentwicklungen ein auf den Testnutzen abgestimmtes Validierungskonzept fehlt.

Ein Grund mag darin liegen, dass in der bisherigen Forschungspraxis ein Test validiert wurde, um dann bei zufriedenstellenden Gütekriterien mögliche Anwendungsfälle des Tests aufzuzählen. Dieses Vorgehen widerspricht jedoch dem in dieser Arbeit verwendeten Validitätsansatz grundlegend. Ein weiterer Grund könnte darin liegen, dass in den *Standards* zwar Prinzipien zum Vorgehen bei der Validierung, und Quellen aus denen Validitätsevidenzen stammen können, genannt werden. Im Text selbst ist jedoch nicht beschrieben, wie Testentwickler*innen dabei konkret vorgehen sollen. Weiter wird in den *Standards* explizit darauf hingewiesen, dass die einzelnen Prinzipien nicht als Checkliste verstanden werden sollen, bei deren Einhaltung man automatisch auf die Validität einer Testwertinterpretation schließen könne. Vielmehr komme es auf den Einzelfall der Testentwicklung, gesellschaftlichen Entwicklungen und gesetzlichen Regelungen an. Ein Ziel dieser Arbeit ist daher, übliche Testnutzen und damit verbundene Testwertinterpretationen von Tests zur Erfassung studentischer Kompetenzen im deutschen Hochschulbereich zu identifizieren[13]. Anschließend soll herausgearbeitet werden, wie auf diese Testwertinterpretationen abgestimmte Validitätsevidenzen aussehen können.

Das Schema wird an späterer Stelle in dieser Arbeit auf empirische Daten angewandt (ab Kapitel 6).

13 Die Einschränkung auf den deutschen Hochschulbereich wird vorgenommen, um den Unterschiedlichkeiten verschiedener Hochschulsysteme Rechnung zu tragen. Die Eigenheiten verschiedener Hochschulsysteme können zu unterschiedlichen Testnutzen und damit einhergehenden Testwertinterpretationen führen, denen in anderen Ländern möglicherweise geringere Relevanz beigemessen wird (Zlatkin-Troitschanskaia, Pant, Kuhn, Toepper & Lautenbach, 2016).

4.1 Verhältnis von Test zu Konstrukt, Lehre und beruflichen Anforderungen

Um Testnutzen im deutschen Hochschulbereich klassifizieren zu können, wird auf das Konzept des *Alignments* von Test, Lehre und Curriculum zurückgegriffen (Pellegrino et al., 2001). Darin wird die Wichtigkeit der Abstimmung dieser drei Aspekte aufeinander betont, da Tests andernfalls ihren Zweck verfehlen könnten. Ruiz-Primo et al. (2002) zeigen jedoch, dass die Abstimmung eines Tests auf konkrete Lehrinhalte oder auf ein definiertes Curriculum variieren sollte, in Abhängigkeit des Testnutzens (vgl. Kapitel 2.3). Das Verhältnis von Test zu Konstrukt, Lehre und beruflichen Anforderungen, soll daher als *Validierungsschema* verwendet werden.

Die Idee der Abstimmung von Tests auf Lehre und Curriculum nach Pellegrino et al. (2001) kommt aus der Testentwicklungen für Kompetenzerfassung bei Schüler*innen in den USA. Für den schulischen Bereich existieren dort wie auch in Deutschland übergreifende Curricula.[14] Im Hochschulbereich sind solche übergreifenden Curricula jedoch nicht üblich (vgl. Kapitel 2.2.2). Um diesem Umstand Rechnung zu tragen, wird stattdessen der allgemeinere Begriff *Konstrukt* verwendet. Dieser Begriff soll verdeutlichen, dass kein Konsens über zu vermittelnde Inhalte und zu erreichende Kompetenzen der Studierenden über Hochschulen hinweg bestehen muss. Kernfrage ist also nicht mehr, ob der Test auf ein Curriculum abgestimmt ist, sondern ob der Test geeignet ist, das intendierte Konstrukt zu erfassen.

Das Verhältnis von Test zu *Lehre* wird auch für studentische Kompetenzen übernommen. Dies kann sowohl die tatsächlich implementierte Lehre, als auch die intendierte Lehre, im Sinne von Modulhandbüchern umfassen. Damit fasst die Abstimmung von Test zu Lehre im Validierungsschema die Abstimmung von Test und Lehre und die Abstimmung von Test zu *Curriculum* zusammen. In den Worten von Anderson (2002) beschreibt die Abstimmung zwischen Test und curricularen Vorgaben das, was typischerweise unter *Inhaltsvalidität* eines Tests verstanden wird. Die Abstimmung zwischen einem Test und tatsächlichen Lehraktivitäten wird hingegen durch die Begriffe *Content Coverage* und *Opportunities-to-Learn* beschrieben.

Die größte Änderung im Vergleich zum Konzept des Alignments für den Schulbereich besteht in der Hinzunahme der *beruflichen Anforderungen*. Rein gesetzlich soll ein Studium auch zur Handlungsfähigkeit im Beruf befähigen (HRG §7). Darüber hinaus sind die Curricula einiger Studiengänge explizit auf berufliche Anforderungen ausgerichtet, etwa im Studium der

[14] In Deutschland verabschieden die Länder eigene Kerncurricula für Schulfächer, welche die nationalen Bildungsstandards implementieren sollen.

Humanmedizin. Daher wird zusätzlich die Abstimmung zwischen Test und *beruflichen Anforderungen* berücksichtigt.

Abbildung 1 zeigt das auf die Erfassung studentischer Kompetenzen angepasste Validierungsschema. Die durchgezogenen Pfade in der Abbildung stehen für die Beziehungen der drei Eckpunkte untereinander. Die Definition des zu messenden Konstruktes kann unterschiedlich stark von Lehre und beruflichen Anforderungen geprägt sein.

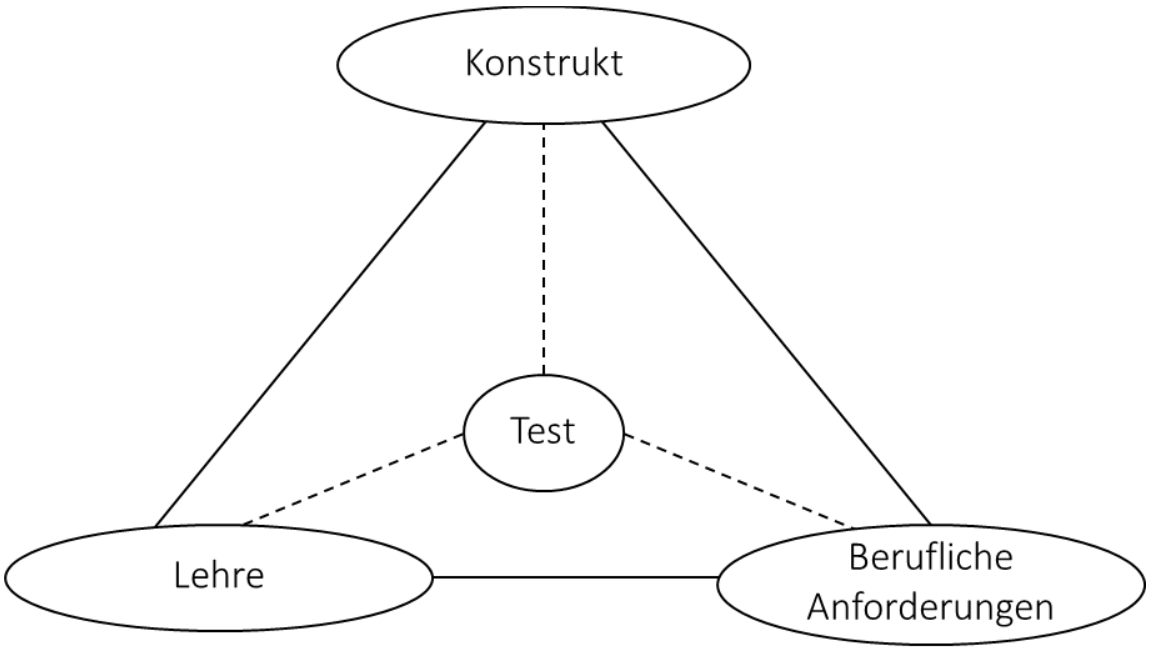


Abbildung 1 Abstimmung von Test auf Konstrukt, Lehre und berufliche Anforderungen

Auch wenn es nur für wenige Studiengänge übergreifende Curricula gibt, besteht vermutlich in einigen Studiengängen ein übergreifendes Verständnis davon, welche Kompetenzen Studierende erreichen sollten. So gelang es etwa in der AHELO-Studie für wirtschaftswissenschaftliche und ingenieurswissenschaftliche Inhalte einen internationalen Konsens über zu vermittelnde Kompetenzen zu definieren (OECD, 2013). Wie eng der Zusammenhang zwischen Konstrukt und Lehre ist, kann jedoch variieren. Bei regulären Modulabschlussprüfungen dürfte der Zusammenhang zwischen Konstrukt und Lehre sehr eng sein (*nahe* Tests im Sinne von Ruiz-Primo et al., 2002). Unter der Annahme, dass in Lehrveranstaltungen unterschiedliche Schwerpunkte gesetzt werden, sollte der Zusammenhang von Konstrukt und Lehre etwas weiter sein, wenn Kompetenzen hochschulübergreifend erfasst

werden (*proximale* oder *ferne* Tests). Ein Konstrukt kann ebenso durch berufliche Anforderungen geprägt sein. Insbesondere bei Tests in Studiengängen, die auf ein konkretes Berufsfeld vorbereiten, wird der Zusammenhang von Konstruktdefinition und zukünftigen beruflichen Anforderungen eng sein. Beispielsweise sollten Tests zur Erfassung professioneller Lehrkompetenz die zu erfassenden Konstrukte insbesondere durch die Bedeutung für zukünftiges Lehrerhandeln definieren. In diesen Studiengängen dürfte auch die Lehre besser auf zukünftige berufliche Anforderungen abgestimmt sein. In der Qualitätsoffensive Lehrerbildung etwa wird explizit die Verbesserung der Lehrkräftebildung gefördert, um Studierende besser auf ihren Berufsalltag vorzubereiten.

Im Zentrum der drei Aspekte *Konstrukt*, *Lehre* und *berufliche Anforderungen* steht der *Test*. Dieser kann zu unterschiedlichen Zwecken entwickelt worden sein und verfolgt dementsprechende Testwertinterpretationen. Diese Testwertinterpretationen sollten durch Evidenzen gestützt werden, die sich auf den Zusammenhang des Tests zu einem der drei Eckpunkte beziehen, repräsentiert durch die gestrichelten Pfade in Abbildung 1.

Verhältnis von Test zu Konstrukt

Tests zur Erfassung individueller Leistung (z.B. zur Notenvergabe bei Hochschulprüfungen) könnten Testwertinterpretationen verfolgen, die den Zusammenhang von Test und Konstrukt betonen. Eine konkrete Testwertinterpretation wäre beispielsweise „Testwerte werden als Indikatoren für die individuelle Ausprägung der Kompetenz interpretiert." Ein entsprechendes Validitätsargument könnte vor allem Evidenzen präsentieren, die typischerweise unter Konstruktvalidität verstanden werden. Die Grundannahmen würden z.B. nach der internen Struktur der Testwerte fragen und als Validitätsevidenz könnte geprüft werden, ob diese den theoretisch erwarteten Strukturen entspricht (*interne Struktur* als Quelle der Validitätsevidenz nach den *Standards*).

Verhältnis von Test zu Lehre

Tests zur Evaluation (z.B. zur Überprüfung der Wirksamkeit unterschiedlicher Lehrformen) würden in Testwertinterpretationen eher den Zusammenhang von Test und Lehre betonen, etwa „Unterschiede in den Testwerten spiegeln Unterschiede in den erlebten Lerngelegenheiten wider." Diese Testwertinterpretation sollte durch Evidenzen gestützt werden, die sich auf den Zusammenhang von Test und Lehre beziehen. Eine überprüfbare Grundannahme könnte sein, dass Studierende mit mehr oder qualitativ höherwertigen Lerngelegenheiten höhere Testwerte aufweisen sollten, als Studierende ohne diese

Lerngelegenheiten (*Zusammenhang zu anderen Variablen* als Quelle der Validitätsevidenz nach den *Standards*).

Verhältnis von Test zu beruflichen Anforderungen

Ein Test zur Erfassung individueller Leistung könnte auch genutzt werden, um Aussagen über zukünftige Leistungen im Berufsleben zu treffen. Eine entsprechende Testwertinterpretation wäre „Testwerte sind Indikatoren für erfolgreiches Handeln im Beruf“. Eine prüfbare Grundannahme dieser Interpretation lautete, ob Personen mit höheren Testwerten beruflich erfolgreicher handeln als Personen mit niedrigen Testwerten (*Zusammenhang zu anderen Variablen* als Quelle der Validitätsevidenz nach den *Standards*).

Abgrenzung der Testwertinterpretationen

In den obigen Beispielen von Testwertinterpretationen wird jeweils nur auf den Zusammenhang von Test zu einem Aspekt des Schemas eingegangen. Dies dürfte in der Praxis allerdings selten vorkommen, da die Punkte des Dreiecks in der Regel nicht getrennt voneinander betrachtet werden können. Wie in der Abbildung dargestellt und weiter oben beschrieben, wird es mehr oder weniger enge Zusammenhänge von Konstrukt, Lehre und beruflichen Anforderungen geben. Diese wirken sich auch auf die Testwertinterpretationen aus.

Ein Beispiel soll dies verdeutlichen. Der oben genannte Test zur Evaluation einer Lehrveranstaltung verfolgt die Interpretation „Unterschiede in den Testwerten spiegeln Unterschiede in den erlebten Lerngelegenheiten wider.“ Die genannte Grundannahme und entsprechende Validitätsevidenzen beziehen sich auf den Zusammenhang von Testwert und Lerngelegenheiten. Jedoch wird die Lehrveranstaltung nur dann als vorteilhaft gelten, wenn sie nicht nur für höhere Testwerte verantwortlich ist, sondern auch belegt wurde, dass die Testwerte das intendierte Konstrukt erfassen. Die Testwertinterpretation beinhaltet also neben dem Zusammenhang von Test zu Lehre auch den Zusammenhang von Test zu Konstrukt.

Ein Extremfall, in dem tatsächlich nur der Zusammenhang von Test zu einem Eckpunkt interessiert, könnte die Prognose von Berufserfolg sein. Wenn die Interpretation verfolgt wird „Testwerte sind Indikatoren für beruflich erfolgreiches Handeln“, und dieser Test z.B. zur Bewerber*innenauswahl eingesetzt wird, könnte irrelevant sein, was genau der Test erfasst.[15]

[15] Vgl. Kapitel 2.1: Diese Perspektive entspricht dem Endpunkt des Kontinuums bei Blömeke, Gustafsson und Shavelson (2015), nach dem Kompetenz als Handeln in einer Anforderungssituation selbst verstanden wird.

Diese Perspektive erlaubt jedoch nicht, zugrundeliegende Prozesse erfolgreichen Handelns zu identifizieren. Das wäre notwendig um zu verstehen, wie diese Prozesse gefördert werden können und ist damit für die Verbesserung von Lehre relevant.[16] Für Tests zur Erfassung studentischer Kompetenzen wird daher erwartet, dass sich Testwertinterpretationen häufig nicht nur einem Zusammenhang im Schema zuordnen lassen.

4.2 Forschungsfragen

Im vorherigen Abschnitt wurde ein Schema vorgestellt, mit dem Testnutzen für die Erfassung studentischer Kompetenzen eingeordnet werden sollen. Nun wird geprüft, ob sich die in bisherigen Testentwicklungen genannten Testnutzen in dieses Schema einordnen lassen. Daher müssen bisherige Testnutzen identifiziert werden und überprüft werden, ob sich diese in das Validierungsschema einordnen lassen.

1. Beziehen sich bisherige Zwecke von Tests für die Erfassung studentischer Kompetenzen auf das Verhältnis von Test zu Konstrukt, Test zu Lehre und Test zu beruflichen Anforderungen und lassen sich damit in das Validierungsschema einordnen?

Das Validierungsschema soll auch zur Ableitung passender Validitätsevidenzen dienen. Validitätsevidenzen beziehen sich jedoch nicht auf einen Testnutzen. Vielmehr sind mit einem Testnutzen bestimmte Testwertinterpretationen verbunden, auf die ein Validitätsargument abgestimmt wird (vgl. Kapitel 3.2). Daher sollen auch die mit den Testnutzen verbundenen Testwertinterpretationen identifiziert werden und überprüft werden, ob sich diese in das Validierungsschema einordnen lassen:

2.1 Welche Testwertinterpretationen werden bisher für Tests zur Erfassung studentischer Kompetenzen in Deutschland genannt?

2.2 Lassen sich die Testwertinterpretationen in das Validierungsschema einordnen?

Anschließend werden zu den jeweiligen Testwertinterpretationen die Validitätsevidenzen identifiziert. Unter der Voraussetzung, dass sich die Testwertinterpretationen in das

[16] Vgl. Kapitel 2.1: Diese Perspektive entspricht dem anderen Ende des Kontinuums nach Blömeke, Gustafsson und Shavelson (2015), in dem Handeln in relevanten Anforderungssituationen als Indikator für die nicht direkt beobachtbare Kompetenz betrachtet wird.

Validierungsschema einordnen lassen, sollten sich Validitätsevidenzen auf die für die Testwertinterpretation identifizierten Zusammenhänge beziehen. Untersucht wird, ob Testwertinterpretationen, die zu unterschiedlichen Kategorien zugeordnet wurden, spezifische Muster von Validitätsevidenzen zeigen.

3.1 Welche Validitätsevidenzen werden für spezifische Testwertinterpretationen geliefert?

3.2 Lassen sich Muster von Validitätsevidenzen für Testwertinterpretationen einzelner Kategorien identifizieren?

4.3 Methode

Um die Forschungsfragen zu beantworten, wurde zunächst eine Literaturrecherche für bisherige Testentwicklungen zur Erfassung studentischer Kompetenzen in Deutschland durchgeführt. Basierend auf den als relevant erachteten Literaturen werden die von Testentwicklern vorgeschlagenen Testnutzen (Forschungsfrage 1), Testwertinterpretationen (Forschungsfrage 2.1) und Validitätsevidenzen (Forschungsfrage 3.1) identifiziert. Die Zuordnung von Testnutzen und Testwertinterpretationen in das Validierungsschema (Forschungsfrage 1 und 2.1) sowie die Passung von Validitätsevidenzen zu Testnutzen (Forschungsfrage 3.2) wird über eine qualitative Inhaltsanalyse (Mayring, 2015) kodiert. Nach Kuckartz (2018) zeichnet sich die qualitative Inhaltsanalyse unter anderem durch das systematische Vorgehen (etwa die Definition von Kategorien und die Erstellung eines Codierleitfaden), die Verwendung von Gütekriterien (etwa die Übereinstimmung von Beurteilern) und die zentrale Bedeutung von Kategorien aus. Dies ermöglicht die Quantifizierung der von Testentwickler*innen vorgeschlagenen Testnutzen, Testwertinterpretationen und Validitätsevidenzen und ermöglicht Aussagen über (relative) Häufigkeiten.

Die Literaturrecherche zur Identifikation relevanter Beiträge und das Vorgehen der qualitativen Inhaltsanalyse werden im Folgenden beschrieben.

4.3.1 Literaturrecherche

Aufgrund der Fragestellung sollen Publikationen analysiert werden, die standardisierte Tests zur Erfassung von Kompetenzen bei Studierenden an deutschen Hochschulen vorstellen.

Die Literaturrecherche wird eingegrenzt auf wissenschaftliche Literaturdatenbanken. Damit wird bewusst auf mögliche Treffer aus Internetsuchmaschinen oder öffentlich zugänglichen

Wissensdatenbanken (z.B. Wikipedia) verzichtet. Diese Entscheidung wurde getroffen, weil davon ausgegangen wird, dass nach wissenschaftlichen Kriterien entwickelte und überprüfte Kompetenztests auch in entsprechenden Organen publiziert werden. Die Dokumentation von neu entwickelten Tests in einschlägigen Zeitschriften, Portalen, etc. soll die Tests der Wissenschaftsgemeinde zur Prüfung und Weiternutzung zugänglich machen. Für die Recherche wurden für die Bildungsforschung wichtige Datenbanken ausgewählt. Dazu zählten ERIC[17], FIS Bildung[18], Web of Science (disziplinenübergreifend) und EBSCOHost, über das auf PsycINFO[19], PsycArticles[20] und ERC[21] zugegriffen wurde.

Die Literaturrecherche wurde als Volltextsuche[22] im April 2017 durchgeführt. Die für die Volltextsuche verwendeten Suchbegriffe wurden in einem mehrschrittigen Verfahren bestimmt. Zunächst wurde ein Wörterpool mit Begriffen generiert, die aus deutsch- und englischsprachigen Publikationen zum Thema bekannt waren. Für diese Begriffe wurden Synonyme gesucht und ebenfalls in die Suchen eingeschlossen. Nachdem die Suchbegriffe feststanden, wurde in den Thesauri der ausgewählten Literaturdatenbanken nach passenden Stichwörtern gesucht, die dann ebenfalls in den Wörterpool für die Volltextsuche eingingen. Die finalen Suchbegriffe für die deutsch- sowie die englischsprachige Suche sind in Tabelle 1 abgebildet.

Um die Treffer der Suchanfragen einzugrenzen, wurden bestimmte Wortkombinationen über sachlogische Operatoren verbunden. Begriffe, die sich auf das Messen von Kompetenzen (z.B. „competence“, „competency“, „assessment“, „Tests“) oder auf Hochschulen (z.B. „universitä*“, „Hochschul*“, bzw. „universit*“, „higher education“) bezogen, wurden so mit Begriffen verknüpft, welche die Einschränkung auf das deutsche Hochschulsystem sicher stellen sollten (z.B. „deutsch“, „german“). In jeder Datenbank wurden alle möglichen Wortkombinationen als

[17] ERIC ist eine US-amerikanische Datenbank für Bildungsforschung und –information, in der über 1000 Zeitschriften sowie Hochschulschriften und Konferenzbeiträge gelistet sind.

18 FIS Bildung ist die Datenbank zum Bildungswesen in Deutschland. Literaturnachweise zum Themengebiet der Hochschulen bilden mit knapp 125.000 Treffern die drittgrößte Themengruppe (Stand: 17.04.2017).

19 PsycINFO ist nach eigenen Angaben die weltweit größte Datenbank für peer-reviewed Texte zu „behavioral science and mental health” (Stand: 17.04.2017)

20 PsycArticles listet APA-Journals.

[21] ERC deckt den Bereich Bildungsforschung und -information ab. Es sind mehr als 2000 Zeitschriften zu allen Bildungsbereichen gelistet.

[22] Die unterschiedlichen Datenbanken verwenden keinen einheitlichen Thesaurus. Eine Literaturrecherche über die einzelnen Thesauri hätte zu unterschiedlichen Suchanfragen in den Datenbanken geführt. Deshalb wurde die Literatursuche als Volltextsuche durchgeführt.

Volltextsuche durchgeführt. Alle durchgeführten Suchanfragen mit der jeweiligen Trefferzahl sind in Anhang A dokumentiert. Eine beispielhafte englische Suchanfrage in der Datenbank EBSCOhost lautete:

> (german* AND "higher education") AND ("test development" OR competenc* N5 measur* OR competenc* N5 assess*)

Die identifizierten Texte wurden nach Ausschluss von Duplikaten nach folgenden Kriterien selektiert:

1) Detaillierte Beschreibung der verwendeten Tests

Als erstes Kriterium wurde die Beschreibung der verwendeten Tests und Angabe von Gütekriterien überprüft.[23] Da die *Standards* auch Nachnutzer*innen von Tests in der Pflicht sehen, die Angemessenheit des Tests für den jeweiligen Einsatz zu prüfen, sollten in Testpublikationen hinreichend Informationen zu Gütekriterien des Tests zu finden sein. Publikationen, die keine Gütekriterien zu den verwendeten Tests berichten, wurden daher nicht berücksichtigt.

2) Stichprobe

Diese Arbeit legt den Fokus auf *studentische Kompetenzen* im *deutschen* Hochschulbereich. Daher werden im Folgenden nur Tests berücksichtigt, welche im Studium erworbene Kompetenzen erfassen. Die Stichprobe sollte daher aus Studierenden bestehen. Tests zur Vergabe von Studienplätzen bleiben damit unberücksichtigt.[24] Außerdem wurden nur Artikel berücksichtigt, in denen Kompetenzen von Studierenden an deutschen Hochschulen erfasst wurden.[25]

---

[23] Hier wurden vermehrt historischen Betrachtungen über Bildungssysteme; Positionspapiere zu Bildungsthemen; Steuern; weitere Buchbesprechungen und Trainingsevaluationen ohne Aussagen zu den verwendeten Outcome-Maßen ausgeschlossen.

[24] Neben dem Medizinertest zur Vergabe von Studienplätzen betrifft das auch aktuelle Entwicklungen von Zulassungstests für Pharmaziestudiengänge („Pharmazeutischer Studierfähigkeitstest“ [PhaST] in Baden-Württemberg und „Hamburger Naturwissenschaftstest“ [HamNat] an der Uni Greifswald). Darüber hinaus sind bei zwei-stufigen Studiengängen auch Tests beim Übergang zum höherwertigen Studium denkbar. In der Praxis werden hier jedoch meist im Grundstudium erworbene Credit Points in definierten Themengebieten gefordert.

[25] In der Datenbank PsycArticles wurden etwa einige Artikel zur Didaktik des Deutschstudiums in den USA gefunden.

3) Kompetenztest mit Leistungsitems

Es wurde nur Literatur berücksichtigt, in der Kompetenz von Studierenden durch Leistungstestitems erfasst wird, die Antworten von Studierenden also (abgestuft) beurteilt werden als „richtig“ oder „falsch“.

4) Veröffentlichung ab 2010 in englischer oder deutscher Sprache

Die Literatur sollte Testentwicklungen und Testeinsätze der letzten Jahre umfassen. Dies war zum einen der Aktualität von Testanwendungen geschuldet, zum anderen wurde die Wahrscheinlichkeit von auf Testnutzen abgestimmte Validitätskonzepte bei neueren Publikationen als höher eingeschätzt. Daher wurde das Selektionskriterium pragmatisch (angelehnt an das Review von Zlatkin-Troitschanskaia et al., 2016) auf 2010 festgelegt. Da der Fokus auf dem deutschen Hochschulsystem liegt, wurde die Veröffentlichungssprache auf Deutsch und Englisch, als dominante Wissenschaftssprache, eingeschränkt. Es ist zu beachten, dass hierdurch möglicherweise relevante Literatur in anderen Veröffentlichungssprachen verlorengeht.

Die so identifizierten Treffer wurden erweitert um Literaturen aus dem Review von Zlatkin-Troitschanskaia et al. (2016), das eine Übersicht über internationale und nationale Tests zur Erfassung studentischer Kompetenzen bietet, sowie der Anwendung des Schneeballsystems. Der Ablauf der Identifikation relevanter Texte ist in Abbildung 2 dargestellt.

Tabelle 1 Suchbegriffe für die Literaturrecherche

Kriterium	Deutsche Suchbegriffe	Englische Suchbegriffe
Einschränkung auf deutsche Hochschulen:	Deutsch	German
	Hochschule	higher education
	Universität	university
Fokus Testentwicklung	Testentwicklung	test development
	Testkonstruktion	
Kompetenzerfassung	Kompetenzmessung	competence
	Kompetenzerfassung	competency
	Wissen	knowledge
	Kompetenz	skill
	Fähigkeit	measure
	messen	assess
	erfassen	

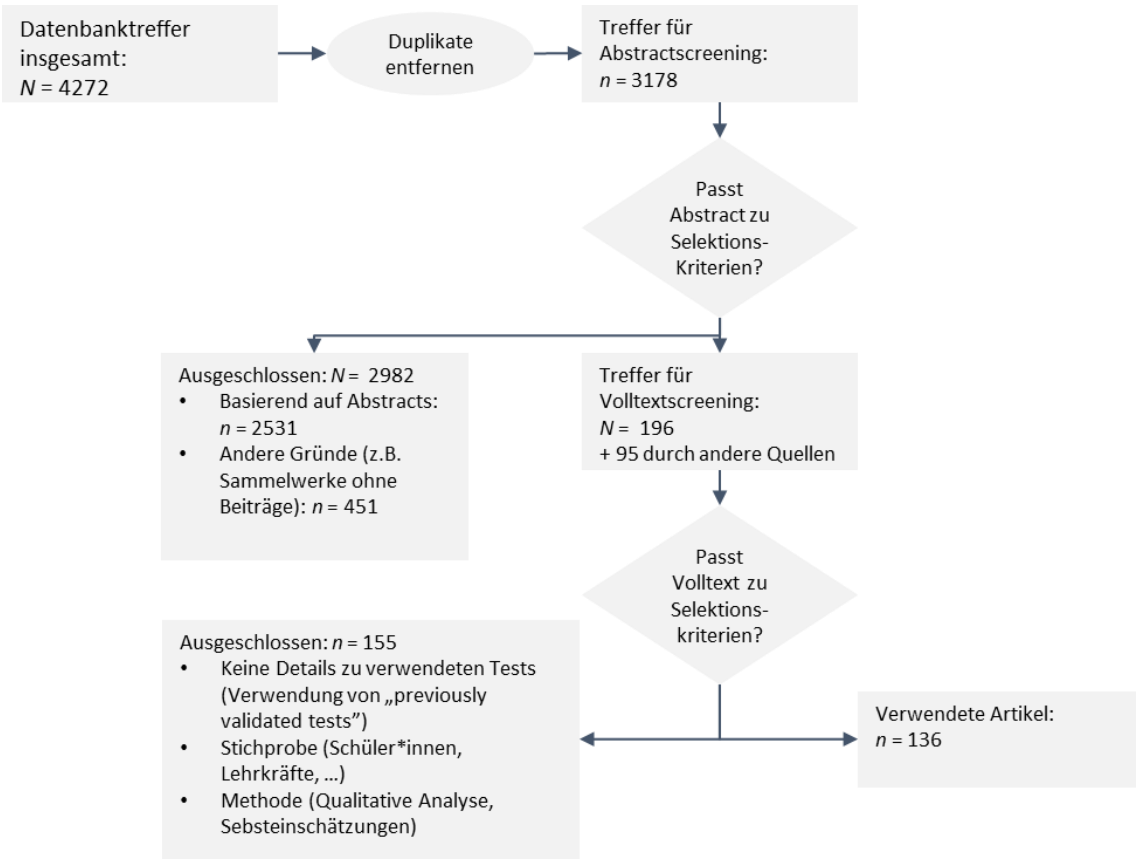


Abbildung 2 Flussschema zur Selektion der Volltexte aus den Treffern der Literaturrecherche

Anmerkungen. Zusätzliche Treffer (*n* = 95), für welche die Volltexte recherchiert wurden, stammen aus Verweisen von vorherigen Treffern (Schneeballsystem) und aus einem Abgleich mit der Literatur des Reviews von Zlatkin-Troitschanskaia et al. (2016).

4.3.2 Kodierung von Textstellen

Wenn in einer Literatur der Testnutzen oder die Testwertinterpretation nicht explizit genannt waren, sollten diese mittels qualitativer Inhaltsanalyse identifiziert werden. Dazu ist es notwendig, zunächst das Ausgangsmaterial zu bestimmen. Anschließend wird die Kodierung des Materials beschrieben.

4.3.2.1 Ausgangsmaterial

Aufgrund der Menge der Treffer aus der Literaturrecherche, wurde eine Volltextanalyse als zu aufwendig eingeschätzt. Deshalb wurden aus jedem Text einzelne Abschnitte ausgewählt. Diese wurden anschließend auf Testnutzen, Testwertinterpretation und Validitätsevidenzen hin analysiert und kodiert.

In der *Einleitung* eines Textes sollten die Relevanz und das Ziel des Vorhabens deutlich werden, wodurch man auf den Testnutzen oder die Testwertinterpretation eines Testes schließen kann. Die entsprechenden Textstellen sollten Formulierungen enthalten wie „Dieser Test soll ... erfassen"; „Der Test zur Erfassung von …"; „Ziel dieser Arbeit ist die Entwicklung eines Tests zur…". Aus den *Hypothesen* lassen sich wie auch aus den berichteten *Ergebnissen* die für die Validierung genutzten Quellen von Evidenzen ableiten („Wie hängen die Testwerte mit … zusammen?"; „Lassen sich die Testwerte durch … vorhersagen?"). In der Diskussion eines Textes sollten Textabschnitt zu finden sein, welche auf die mit Testwerten verbundene Interpretationen schließen lassen. Entsprechende Formulierungen könnten lauten: „Deshalb werden die Ergebnisse als Indikatoren für … interpretiert", „Daraus schließen wir, dass der Test … erfasst.". Der *Ausblick* eines Textes, häufig auch in den Diskussionsteil integriert, wurde auf relevante Abschnitte hin untersucht, da die Autorin dieser Dissertation feststellte, dass dort mögliche Testzwecke des entwickelten Instruments genannt werden und daraus auf den Testnutzen geschlossen werden kann („Als Einsatz ist … denkbar.").

4.3.2.2 Kodierung

Nach Auswahl der zu kodierenden Textstellen erfolgten zunächst eine Sichtung des Materials und der Aufbau eines Pools an typischen Begriffen und Formulierungen, die einer bestimmten Kategorie zugeordnet wurden (Definition von Ankerbeispielen). Anschließend kodierten drei geschulte Kodierer*innen getrennt voneinander eine unterschiedliche Anzahl an Textstellen. Die Kodierung erfolgte im Zeitraum von Oktober 2017 bis August 2018. Das Kategorienschema und die Kodieranweisungen sind in Tabelle 2 abgebildet.

Tabelle 2 Kodierschema für die Identifikation von Testwertinterpretationen

Kategorie (Code)	Ankerbeispiele	Kodierregeln
Konstrukt (1)	"Prior to the construction of an adequate test instrument the substance of PPK must be specified. What exactly is PPK as we understand it today?" (Hohenstein, 2015; S. 107 "Findet sich die theoretisch angenommene eindimensionale Struktur der professionellen Wahrnehmung inhaltlicher Strukturierung in den Daten wieder?" (Meschede, 2013; S. 91)	Die Textstelle verbindet Testwerte mit dem Konstrukt, welches der Test erfassen soll. Beispielsweise könnte das Ziel einer Studie lauten, ein Konstrukt besser zu verstehen. Es werden Hypothesen über interne Struktur des Konstrukts und über Zusammenhänge von Testwerten zu Kriterien aus einer Theorie abgeleitet.
Lehre (2)	"Sind Annahmen über das Fachdidaktische Wissen im Sinne des hier verwendeten Konstrukts als ein Wissen, das primär an der Universität erwerbbar ist, möglich?“ Gramzow, 2015; S. 156 "Bei Lehramtsstudierenden ist ferner von Interesse, ob diese Kompetenzen durch fachbezogene und fachdidaktische Lehrveranstaltungen im Verlauf des Studiums beeinflusst werden." Kotzebue & Nerdel (2012), S. 189 "Die Ergebnisse der Studie können dahingehend interpretiert werden, dass es derzeit den Universitäten in NRW nur unzureichend gelingt, bildungswissenschaftliches Wissen systematisch und kumulativ aufzubauen." (Kunina-Habenicht et al., 2013; S. 18)	Die Textstelle verbindet Testwerte mit hochschulischen Lerngelegenheiten. Die Testleistung wird auf hochschulische Lerngelegenheiten zurückgeführt, z.B. wird angenommen, dass Studierende mit mehr Lerngelegenheiten (in höheren Semestern, in bestimmten Studiengängen) höhere Testwerte erzielen als andere Studierende. In diesem Fall werden quantitative Unterschiede in Lerngelegenheiten angenommen. Bei Evaluationen von Lehrformen werden qualitative Unterschiede in Lerngelegenheiten angenommen. Die qualitativ hochwertigere Lehrform wird mit höheren Testwerten in Verbindung gebracht.
Berufliche Anforderungen (3)	"Eine besondere Herausforderung hingegen besteht darin, Wissen zu erfassen, das besonders eng mit der tatsächlichen Bewältigung spezifischer beruflicher Anforderungen verknüpft ist …" (König & Lebens, 2012; S. 4) "Ist das Instrument sensitiv für Unterschiede in der professionellen Wahrnehmung zwischen Lehramtsstudierenden des Sachunterrichts zu Beginn sowie am Ende ihres Studiums (Novizen) und naturwissenschaftsdidaktisch erfahrenen Lehrpersonen (Experten)?" Meschede, 2013; S. 93)	Die Textstelle bringt das Konstrukt mit beruflichen Anforderungen in Verbindung. Dazu zählen die Verbindung zu beruflichen Anforderungen (Test spiegelt realitätsnahe Anforderungen) und die Prognose von erfolgreichem Handeln in Abhängigkeit der Testwerte. (höhere Testwerte → beruflich erfolgreicher) Hinweis zum zweiten Beispiel: Die Kontrastierung von Studierenden, unabhängig vom Studienfortschritt, und erfahrenen Lehrpersonen spricht für berufliche Anforderungen. Würde der Gruppenvergleich auch zwischen Studienanfängern und fortgeschrittenen Studierenden gezogen, läge eine Zuordnung zur Kategorie 2 nahe. In dem Fall sollte „Nicht eindeutig zuzuordnen“ mit der Ergänzung „(2, 3)“ kodiert werden.
Keine der Kategorien (0)		Die Textstelle liefert keine Hinweise auf eine der Kategorien.
Nicht eindeutig zuzuordnen (A, B)	"Das in der Lehrerausbildung angeeignete Wissen schlägt sich bei angehenden Lehrkräften in der classroom expertise management (CME) nieder. Beide Maße sind mit der Qualität des gehaltenen Unterrichts korreliert." Bsp. Zuordnung: (2,3)	Die Textstelle ist nicht eindeutig einer Kategorie zuzuordnen. In Klammern bitte die Kategorien angeben, zwischen denen das Rating schwankt.

4.3.3 Gütekriterien der qualitativen Inhaltsanalyse

Ein Gütemerkmal der qualitativen Inhaltsanalyse ist, dass die Arbeit durch fachlichen Austausch begleitet wird (*Forschungswerkstatt;* Mayring, 2015). Die Verfasserin dieser Arbeit nahm an der DIPF-internen Forschungswerkstatt zu qualitativen Analysen von Februar 2018 bis November 2018 teil. Dort stellte sie das geplante Kategorienschema vor und stellte die Anwendbarkeit des Schemas zur Identifikation ausgewählter Textstellen zur Diskussion.

Als wichtiges Gütekriterium wird die Reliabilität von Zuweisungen des Materials zu einzelnen Kategorien angegeben. Dies kann über die Übereinstimmung verschiedener Rater geschehen (Interrater-Reliabilität), was als Maß für die Reproduzierbarkeit der Ergebnisse interpretiert wird. Die Interrater-Reliabilität der Zuordnung von Textstellen in Kategorien wurde in R (R Core Team, 2020) mit Hilfe des Psych Pakets (Revelle, 2020) berechnet.

Für die Berechnung von Cohen's Kappa wurden nur Zuordnungen verwendet, die eindeutig waren. Die durchschnittliche Interrater-Reliabilität liegt bei $\kappa = .78$ für 121 Textstellen, die von allen drei Ratern zugeordnet wurden. Tabelle 3 gibt die Beurteilerübereinstimmung zwischen einzelnen Ratern an. Diese liegen alle mindestens in einem Bereich, der als substanzielle Übereinstimmung zwischen Ratern interpretiert wird (Cohen, 1960).

Tabelle 3 Beurteilerübereinstimmung der Zuordnung von Textstellen

	Rater 1	**Rater 2**	**Rater 3**
Rater 1	n = 732	$n_{1,2}$ = 138	$n_{1,3}$ = 140
Rater 2	.79 [.69 - .89]	n = 168	$n_{2,3}$ = 129
Rater 3	.80 [.71 - .90]	.75 [.65 - .86]	n = 181

Anmerkungen. Werte unterhalb der Diagonalen geben Cohen's Kappa als Maß für die Beurteilerübereinstimmung und das 95% Konfidenzintervall an. In der Diagonalen ist die Anzahl kodierter Textstellen eines Raters angegeben. Werte oberhalb der Diagonalen enthalten die Anzahl gemeinsam kodierter Textstellen der zwei Rater.

4.4 Ergebnisse

Im Folgenden werden die Ergebnisse nach Kodierung des Raters 1 beschrieben, da dieser als einziger alle Textstellen zuordnete.

Entsprechend der Forschungsfragen sollten zunächst die Testnutzen und dann die Testwertinterpretationen identifiziert und dem Validierungsschema zugeordnet werden.

Anschließend sollten die Validitätsevidenzen für einzelne Testwertinterpretationen überprüft werden. Es stellte sich jedoch heraus, dass eine eindeutige Trennung von Testnutzen und Testwertinterpretation kaum möglich war. Dies lag vor allem daran, dass in den seltensten Fällen explizit Testnutzen oder Testwertinterpretationen genannt wurden. Unter den identifizierten Literaturen waren 71 Publikationen, die 2014 oder später veröffentlicht wurden. Von diesen bezogen sich jedoch nur 12 bei der Validierung auf Kane (2013) oder auf die aktuellen *Standards*. Auch in diesen wurde jedoch nicht in allen Fällen eine explizite Testwertinterpretation formuliert. Diese musste dann aus ausgewählten Textstellen „rekonstruiert" werden (siehe Tabelle 4).

Für Forschungsfrage 1 werden daher Textabschnitte alle Treffer der Literaturrecherche berücksichtigt. Für die Forschungsfragen 2 und 3 werden jedoch nur diejenigen Literaturen betrachtet, die sich explizit auf einen argumentationsbasierten Validierungsansatz beziehen.

Lassen sich bisherige Testnutzen in das Validierungsschema einordnen? (Forschungsfrage 1)

Die Treffer der Literaturrecherche umfassten vorwiegend Zeitschriftenartikel und Monografien (Dissertationen). Die Anzahl der ausgewählten Textstellen pro Text sollte mit der Länge des Textes zusammenhängen. Während in Monografien mehr Platz für Ausführungen ist, müssen Zeitschriftenartikel bestimmte Umfänge einhalten. Um Verzerrungen der Ergebnisse aufgrund des Textformates vorzubeugen, wurde ein Testnutzen pro Literatur nur einmal gezählt, auch wenn dieser für mehrere Textstellen kodiert wurde.

Für die 136 Literaturen wurden so 278 Testnutzen zugeordnet. Je zu einem Drittel der Fälle wurden das Verhältnis von Test zu Konstrukt und das Verhältnis von Test zu Lehre kodiert. In nur etwa 14% der Fälle wurde das Verhältnis von Test zu beruflichen Anforderungen kodiert. In ähnlichem Umfang wurden Textstellen als nicht eindeutig zuzuordnen eingeschätzt (16%). Eine Übersicht über diese Textstellen unter Angabe der Kategorien, zwischen denen die Zuordnung liegt, findet sich in Anhang C. Für acht Textstellen wurde eine Zuordnung zu keiner der Kategorien als passend eingeschätzt. Für zwei weitere wurde eine Zuordnung zu Lehre vorgenommen bei gleichzeitiger Kodierung von Kategorie 0. Diese Textstellen sind in Anhang D aufgeführt.

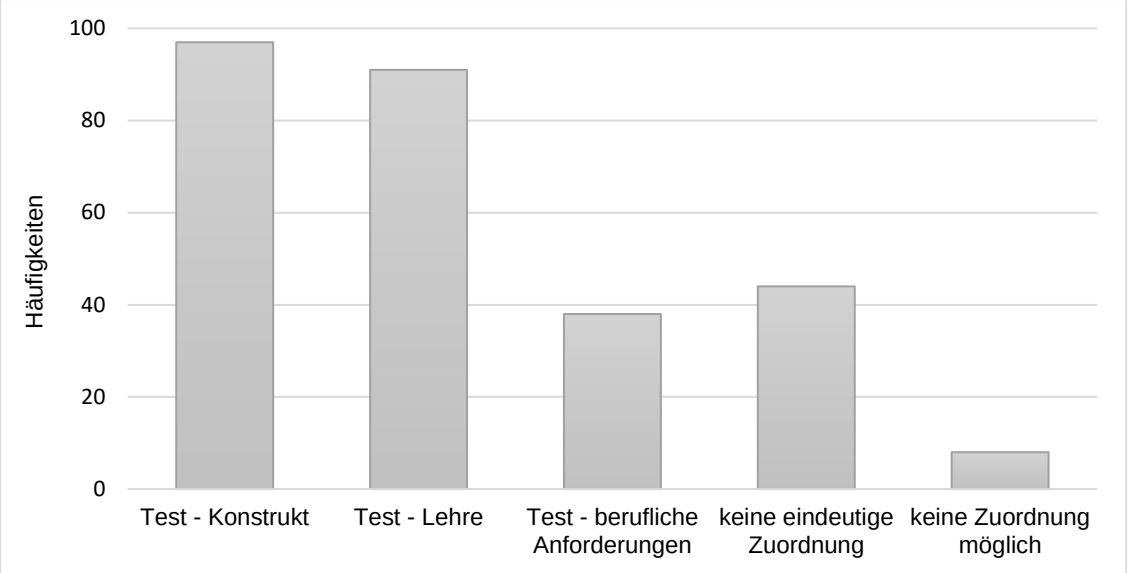


Abbildung 3 Zuordnung von Testnutzen in das Validierungsschema

Welche Testwertinterpretationen werden in Tests zur Erfassung studentischer Kompetenzen genannt und lassen sich diese in das Validierungsschema einordnen? (Forschungsfrage 2)

Die 12 Literaturen, die sich auf ein argumentationsbasiertes Validierungsverständnis beziehen, erfassen fast alle Kompetenzen aus dem Bereich der (Vorschul-) Lehrkräftebildung oder wenden die Tests auch bei Studierenden aus Lehramtsstudiengängen an (z.B. die Erfassung wirtschaftswissenschaftlichen Wissens bei Kuhn et al., 2014). Ausnahme ist eine Literatur, die einen Test zur Erfassung der fachübergreifenden Kompetenz *Wissenschaftliches Denken* vorstellt (Stiller et al., 2016).

Tabelle 4 zeigt die in diesen Texten identifizierten Testwertinterpretationen und deren Kodierung durch Rater 1. Auch hier fällt auf, dass nicht in allen Texten eine explizite Testwertinterpretation genannt wird (vgl. Tabelle 4: Brückner et al., 2015). Diese Texte haben keine Validierung einer Testwertinterpretation zum Ziel. Vielmehr werden die Tests verwendet, um Zusammenhänge von Testwerten zu anderen Variablen zu untersuchen (z.B. bei Blömeke, Jenßen, Grassmann, Dunekacke & Wedekind, 2017).

Die Testwertinterpretationen konnten alle in das Validierungsschema eingeordnet werden. Auch hier liegt der Schwerpunkt auf dem Zusammenhang von Test zu Konstrukt (in 8 Fällen). Es wird jedoch im Vergleich zu allen Textstellen häufiger der Zusammenhang von Test zu beruflichen Anforderungen betont (in 4 von 12 Fällen).

Tabelle 4 Textstellen die als Testwertinterpretation identifiziert wurden und deren Einordnung in das Validierungsschema

Quelle	Textstelle	Kodierung	Anmerkungen
Blömeke et al. (2017), S. 340	„To the best of our knowledge, there is no conceptual framework that specifically describes the structure of preschool teachers' knowledge. To avoid a purely operational definition, we therefore applied basic educational-psychological dimensions of primary teachers' knowledge to preschool teachers but operationalized these on the basis of research on 3 to 6-year-old children's development and learning."	Konstrukt	Der Text versteht sich nicht als Validierung sondern verwendet einen zuvor entwickelten Test um den Zusammenhang von Fachwissen und hochschulischen Lerngelegenheiten zu untersuchen.
Brückner et al. (2015), S. 442	„Before testing the hypotheses empirically, we ensured that we had a valid crossnational measure of economic knowledge."	Konstrukt	Text versteht sich nicht als Validierung; untersucht Einfluss von schulischem Vorwissen auf Testwerte internationale Messinvarianz als Aspekt von Konstrukt. Test wird bei Studienanfängern eingesetzt, für diese ist er relative schwer (S. 444)
Brückner und Pellegrino (2016), S. 294	„Students' response processes are viewed as an especially relevant source of evidence in that the items call upon appropriate mental operations associated with processing the intended forms of knowledge and skill—that is, the nature of a student's thinking and reasoning that lead to correct or incorrect item responses can be attributed to construct-relevant processes."	Konstrukt	Text stellt Methode zur Modellierung von Antwortprozessen dar, Anwendungsbeispiel ist Analyse von mentalen Bearbeitungsprozessen bei Test aus WiWiKom (dt. Adaption des TUCE Tests)
Gramzow (2015), S. 156 S. 156	„Angelehnt an diese Unterscheidung soll das hier entwickelte Testinstrument primär eine Testwertinterpretation, die auf die Erklärung von Leistungen abzielt, legitimieren. Für die Erklärung der Leistung im Test sollen Annahmen über das theoretische Konstrukt des (an der Universität erwerbbaren) Fachdidaktischen Wissens und seiner (möglichen) Facetten herangezogen werden können." „Sekundär soll eine Testwertinterpretation der Bewertung von Leistungen möglich gemacht werden. Die Leistungen einer Person sollen mit den Leistungen einer anderen Person verglichen und dadurch bewertet werden können (vgl. Hartig et al., 2012)."	Konstrukt, Lehre Konstrukt	Mit den Testwerten sind keine Konsequenzen verbunden: "Im Fokus dieser Arbeit steht die intendierte Testwertinterpretation der Erklärung von Leistungen und im zweiten Schritt der Bewertung von Leistungen, Entscheidungen auf Grundlage der Testergebnisse sind nicht angestrebt." S. 177
Hammer et al. (2015), S. 39	„In der vorliegenden Validierungsstudie, in der die Annahme validiert werden soll, dass die Ergebnisse des DaZ-Tests als Indikator für ein entsprechend kompetentes Handeln zukünftiger Fachlehrkräfte gelten können … "	Berufliche Anforderungen	In der Diskussion wird Ausblick auf möglichen Testeinsatz gegeben: "… Dafür müssten gezielt Fortbildungsveranstaltungen entwickelt werden. Der in dieser Arbeit vorgestellte DaZ Kompetenztest könnte dabei als Instrument für die Evaluation eingesetzt werden." S. 51

Fortsetzung auf der nächsten Seite

Quelle	Textstelle	Kodierung	Anmerkungen
Jahn (2014), S. 49	„Die in das Instrument integrierten Videoclips müssen als authentische Unterrichtssequenzen wahrgenommen werden und repräsentativ für die lernwirksamen Unterrichtskomponenten Zielorientierung, Lernbegleitung und Lernatmosphäre sein, vor deren Hintergrund professionelle Unterrichtswahrnehmung erfasst wird."	Berufliche Anforderungen	
Kuhn (2014), S. 7	„Die Bemühungen zur Validierung richten sich im Rahmen dieser Arbeit insbesondere darauf zu untersuchen, inwieweit die Testergebnisse bzw. das zugrundeliegende Antwortverhalten durch das theoretisch definierte Konstrukt erklärt werden können (i. S. einer theoriebasierten Testwertinterpretation, Hartig et al., 2007, S. I 60f)."	Konstrukt	Fokus Forschung, keine Verwendung der Testwerte vorgesehen: "Der zuletzt genannte Aspekt wird im Folgenden als vernachlässigbar erachtet, da mit dieser Studie derzeit rein grundlagenorientierte Zwecke verfolgt werden, die keine direkte Verwertung des Tests bzw. der Testergebnisse (bspw. zur Feststellung der Eignung für den Lehrerberuf) vorsehen." S.116
S. 114f	„Eine theoretische und empirische Fundierung entlang der ersten vier Aspekte trägt zu einer umfassenden Einschätzung bei, inwieweit vom Testscore bzw. von den Itemantworten auf das theoretisch definierte Konstrukt „wirtschaftsdidaktisches Wissen" geschlossen werden kann."	Konstrukt	
Kuhn et al. (2014), S. 161	„Die Ergebnisse aus dem ILLEV-Projekt schließen forschungsmethodisch und inhaltlich an den vorliegenden Stand zur Lehrerforschung in den allgemein bildendenden Domänen an. Sie liefern erste empirisch fundierte Hinweise zur spezifischen Struktur der professionellen Kompetenz von (angehenden) Lehrkräften im kaufmännisch-verwaltenden Bereich."	Konstrukt	
Riese et al. (2015), S. 68	„Dabei sollen die entwickelten Testinstrumente primär eine Testwertinterpretation legitimieren, die auf die Erklärung von Leistungen abzielt (angelehnt an die Unterscheidung von Hartig, Frey & Jude, 2012)."	Lehre	Test soll Wissen erfassen, das durch Hochschulstudium erlernt wird, daher Kodierung als Lehre.
Stiller et al. (2016), S. 8	„We conclude that the results support the validity of our interpretation of the test scores as measures of scientific reasoning competencies."	Konstrukt	
Vogelsang (2014), S. 494	„Sollte diese Hypothese als nicht zutreffend abgelehnt werden, wird dies dahingehend interpretiert, dass das Paderborner Instrument in seiner Gesamtheit nicht der Forderung nach Handlungsvalidität genügt. Demnach würde es keine ‚direkte' Handlungsressource für das Unterrichten, also keine Kompetenz im eigentlichen Sinne erfassen."	Konstrukt, berufliche Anforderungen	
Wiesbeck (2015), S. 84	„If the pre-service teachers perceive the conversations as authentic, they probably behave as if in a real situation. A high perceived authenticity would speak for the validity of the results and their prognostic value regarding real parent-teacher conversations."	Berufliche Anforderungen	Handeln erfasst bei Lehramtsstudierenden in simulierten Gesprächen mit Schauspieler*innen

Welche Validitätsevidenzen werden geliefert und lassen sich Muster für Testwertinterpretationen einzelner Kategorien erkennen? (Forschungsfrage 3)

Für die Literaturen, die sich explizit auf einen argumentationsbasierten Validierungsansatz beziehen, wurden die Validitätsevidenzen geprüft. Die in den Artikeln präsentierten Validitätsevidenzen sind in Anhang E aufgeführt und sortiert nach der Quelle der Validitätsevidenz entsprechend der *Standards*.

Auffallend ist, dass alle der Texte, unabhängig von der zugeordneten Kategorie für die Testwertinterpretation, Angaben zur internen Struktur der Testwerte machen oder auf Vorarbeiten verweisen, in denen diese untersucht wurden (Brückner et al., 2015; Hammer et al., 2015). Als Evidenzen werden häufig die Passung von Testwerten zu IRT-Modellen und konfirmatorischen Faktorenanalysen präsentiert sowie Modellvergleiche angestellt. Weitere Methoden sind die Untersuchung von Messinvarianz (Blömeke et al., 2017) und die Vorhersage von Itemschwierigkeiten (Stiller et al., 2016).

Evidenzen zu internalen Antwortprozessen werden nur in vier Texten präsentiert. Die entsprechenden Testwertinterpretationen bezogen sich auf alle Kategorien des Validierungsschemas. Bei Brückner und Pellegrino (2016) und Kuhn (2014) wurden Think-Aloud Studien genutzt, um konstruktrelevante von –irrelevanten Bearbeitungsprozessen zu trennen. Diese Validitätsevidenzen passen zu einer Testwertinterpretation, die auf den Zusammenhang *Test-Konstrukt* abzielt. Bei Gramzow (2015) wird die Methode des lauten Denkens eingesetzt, um die Wissensquellen zu identifizieren, auf Grundlage derer die Items bearbeitet werden. Überprüft wird, ob Studierende die Items mithilfe universitär erworbenen Wissens bearbeiten. Diese Validitätsevidenz passt zu einer Testwertinterpretation, welche auf den Zusammenhang *Test-Lehre* abzielt. Und schließlich nutzte Wiesbeck (2015) kognitive Interviews, um die Authentizität der Items einzuschätzen, also ob diese beruflich relevanten Anforderungssituationen nahekommen. Diese Validitätsevidenz passt zu einer Testwertinterpretation, welche den Zusammenhang *Test-berufliche Anforderungen* fokussiert.

Auch die Evidenzquellen zu Testinhalten und zum Zusammenhang zu anderen Variablen werden für Testwertinterpretationen aus allen Kategorien im Validierungsschema verwendet. Insgesamt lässt sich daher kein Muster für spezifische Quellen von Validitätsevidenzen je nach Testwertinterpretation erkennen.

Erkennbar ist lediglich ein Zusammenhang zwischen der Testlänge und der Anzahl an Quellen für Validitätsevidenzen. Bei den sieben Zeitschriftenbeiträgen berichteten sechs nur ein bis zwei Quellen von Validitätsevidenzen. Nur im Artikel von Blömeke et al. (2017) werden mehr Quellen berichtet, dafür wird aber auf unterschiedliche Vorarbeiten verwiesen. In den fünf Monografien wurden hingegen mindestens drei Quellen, in vier Fällen sogar vier Validitätsevidenzen berichtet.

4.5 Diskussion zur Klassifikation von Testwertinterpretationen

Einordnung von Testnutzen in das Validierungsschema

Der Großteil der in den Textstellen identifizierten Testnutzen bzw. Interpretationen konnte in das Validierungsschema eingeordnet werden. Am häufigsten wurden der Zusammenhang von Test zu Konstrukt und der Zusammenhang von Test zu Lehre zugeordnet. Wenn alle Textstellen betrachtet werden, scheinen Testnutzen und Testwertinterpretationen, die sich auf berufliche Anforderungen beziehen, bislang eine untergeordnete Rolle zu spielen. Bei den Tests, die sich auf einen argumentationsbasierten Validierungsansatz beziehen, wurde in vier Fällen (also zu einem Drittel) die Kategorie *berufliche Anforderungen* kodiert. Alle dieser Tests erfassen Kompetenzen von Lehrkräften. Wie in Kapitel 4.1 diskutiert, beziehen sich diese Tests damit auf Studiengänge, in denen ein stärkerer Bezug zu konkreten Berufsfeldern besteht.

In einigen Textstellen fällt der Aspekt der internationalen Vergleichbarkeit von Kompetenzen auf (z.B. bei Brückner et al., 2015; Buchholtz, 2014). Dieser wurde nicht explizit im Validierungsschema berücksichtigt. Andererseits kann die Invarianz des Konstrukts über Studiengänge, Hochschulen, oder Nationen als Aspekt der Konstruktdefinition verstanden werden. Damit kann dieser Testnutzen eingeordnet werden auf den Zusammenhang von Test zu Konstrukt.

Abgrenzung von Testwertinterpretationen

Wie erwartet, konnten nicht alle Textabschnitte eindeutig zugeordnet werden. Bei den nicht eindeutig zuordenbaren Textstellen aus allen Texten, werden am häufigsten die Kategorien *Konstrukt* (1) und *Lehre* (2) gemeinsam kodiert (35 Textstellen). Seltener tritt eine gleichzeitige Kategorisierung von *Konstrukt* (1) und *berufliche Anforderungen* (3; 6 Textstellen) bzw. von

Lehre (2) und *berufliche Anforderungen* (3; 4 Textstellen auf). In drei Fällen wurde die Textstelle sogar allen drei Kategorien zugeordnet (1,2,3).

Bei Riese et al. (2015) wird etwa die gleichzeitig Kodierung der Kategorien 1 und 2 deutlich: „*Die Ziele des vorgestellten Projekts liegen in der Modellierung und Messung domänenspezifischer und generischer Kompetenzen, die Lehramtsstudierende der Physik im Hochschulstudium erwerben sollen.*" (Abstract). Die zu erfassenden Kompetenzen sind hier zwar nicht näher spezifiziert („…domänenspezifischer und generischer Kompetenzen…"), weisen aber darauf hin, dass die Tests ein bestimmtes Konstrukt erfassen sollen (Zusammenhang Test-Konstrukt). Des Weiteren sollen diese Kompetenzen hochschulisch erwerbbar sein, was den Zusammenhang von Test zu Lehre betont.

Ein Beispiel für die gleichzeitige Kategorisierung von 1 und 3 findet sich bei Vogelsang (2014): „*Bilden die Kennwerte des Paderborner Instruments zur Erfassung professioneller Handlungskompetenz (angehender) Physiklehrkräfte Prädiktoren für die Performanz von (angehenden) Physiklehrkräften im Physikunterricht?*" (S. 280). Zum einen wird die zu erfassende Kompetenz explizit genannt („professionelle Handlungskompetenz angehender Physiklehrkräfte"), zum anderen wird der Zusammenhang von Testwert zu beruflichem Handeln als Validitätsevidenz für die zu erfassende Kompetenz interpretiert. Die Kompetenzdefinition scheint in diesem Fall maßgeblich durch berufliche Anforderungen geprägt zu sein. Der Nähe von Konstrukt und beruflichen Anforderungen führt in diesem Fall zu einer gleichzeitigen Kodierung der zwei Kategorien. Wie in Kapitel 4.1 ausgeführt, wird dies vor allem bei Tests zur Kompetenzerfassung in Studiengängen mit starkem Berufsbezug erwartet, wie z.B. im Lehramtsstudium. Tatsächlich finden sich die gleichzeitige Kodierung von Kategorie 1 und 3 auch in weiteren Tests zur Kompetenzerfassung (angehender) Lehrkräfte (z.B. bei Kirschner, 2013, S. 31; Linninger et al., 2015, S. 80f; Lohse-Bossenz, Kunina-Habenicht, Dicke, Leutner & Kunter, 2015, S. 39f).

Bei Lauterbach (2015) wird eine Textstelle gleichzeitig mit den Kategorien 2 und 3 kodiert: „*Das als Online-Befragung durchgeführte Expertenrating diente der Prüfung der Aufgaben hinsichtlich ihrer Eignung, … die Relevanz entsprechender Kompetenzen für Studium und Beruf zu erfassen*" (S.7). Die Einschätzung der Testinhalte liefert Evidenzen zum Zusammenhang von Test und Lehre sowie zum Zusammenhang von Test und beruflichen Anforderungen.

Ein Beispiel für die Kodierung aller Kategorien findet sich bei Linninger et al. (2015): „*Thus, EK is supposed to provide the theoretical subject-unspecific foundation for teachers' professional*

behavior both inside and outside the classroom and should therefore be fostered by academic teacher education“ (S. 73). Hier wird zunächst das Konstrukt „EK“ benannt (*Konstrukt*), welches professionelles Handeln von Lehrkräften beeinflussen soll (*berufliche Anforderungen*). Daraus wird geschlussfolgert, dass das definierte Konstrukt in hochschulischer Bildung gefördert werden sollte (*Lehre*).

Wie die Beispiele zeigen, ist die Zuordnung von mehreren Kategorien keine Frage der Länge eines Textabschnittes. Das Beispiel für die Zuordnung zu allen drei Kategorien umfasst etwa nur einen Satz. Die Mehrfachkodierungen verdeutlichen die Beziehungen der Außenseiten des Validierungsschemas. Die Aspekte *Konstrukt*, *Lehre* und *berufliche Anforderungen* können nicht unabhängig voneinander betrachtet werden. Damit wird die Relevanz des Validierungsschemas zur Erfassung studentischer Kompetenzen, in dem eben diese drei Aspekte in Bezug zueinander gesetzt werden, als gegeben betrachtet. Gleichzeitig erschwert dies die Ableitung von Validierungsstrategien für einzelne Testwertinterpretationen.

Validierungsstrategien

Muster für spezifische Testwertinterpretationen. Auf Basis der identifizierten Testwertinterpretationen lassen sich keine Muster von Quellen von Validitätsevidenzen auf spezifische Testwertinterpretationen erkennen. Die Einordnung von Testwertinterpretationen in das Validierungsschema können aber genutzt werden, um zu erkennen, für welche Zusammenhänge Evidenzen im Validitätsargument benötigt werden. Eine beispielhafte Anwendung findet sich in den Kapiteln 6 bis 8.

Testnutzen Forschung. Sowohl für alle Textstellen als auch für die Subgruppe, die sich auf einen argumentationsbasierten Validierungsansatz beziehen, steht der Zusammenhang von Test zu Konstrukt im Vordergrund. Validitätsevidenzen die sich auf diesen Zusammenhang konzentrieren werden typischerweise als *Konstruktvalidität* bezeichnet.

Wenn Evidenzen zum Zusammenhang von Test zu Konstrukt als gesichert gelten, werden mögliche Testnutzen des „validen“ Tests vorgeschlagen, z.B. bei Hammer et al. (2015), die sich an anderer Stelle explizit auf den argumentationsbasierten Validierungsansatz nach Kane (2013) beziehen:

> „Inhaltlich wirft dies die Forschungs- und Entwicklungsfrage auf, inwieweit und welche Lernarrangements hier für die fachdidaktische

> Lehramtsausbildung an Universitäten konzipiert werden könnten. Dafür müssten gezielt Fortbildungsveranstaltungen entwickelt werden. Der in dieser Arbeit vorgestellte DaZ Kompetenztest könnte dabei als Instrument für die Evaluation eingesetzt werden." (S. 51)

Zu betonen ist auch, dass in einigen Texten gar nicht der Anspruch verfolgt wurde eine Testwertinterpretation zu validieren. Vielmehr sollten mithilfe „valider" Tests die eigentlichen Forschungsfragen adressiert werden. Kirschner (2013) formulieren etwa:

> „Nachdem gezeigt wird, dass die Tests reliabel und valide sind, werden die fachspezifischen Bereiche des Professionswissens näher beleuchtet. Es wird untersucht, welche demographischen Variablen mit CK und PCK von Physiklehrkräften zusammenhängen. Des Weiteren wird analysiert, wie sich CK und PCK und ihr Zusammenhang zwischen Lehramtsstudierenden, Lehrkräften im Vorbereitungsdienst und Lehrkräften unterscheiden." (S. 38)

In diesen Texten steht der Forschungsaspekt im Vordergrund, der Test ist z.B. ein Mittel um ein besseres theoretisches Verständnis zu entwickeln. Der eigentliche Testnutzen müsste hier also „Forschungsinteresse" lauten. Mit dem *Test Use* in den *Standards* ist dahingegen ein Testeinsatz gemeint, bei dem anhand der erfassten Testwerte Entscheidungen getroffen werden. Diese Entscheidungen sind mit Konsequenzen verbunden, weshalb der Untersuchung nicht erwünschter Konsequenzen von Testeinsätzen ein eigenes Kapitel in den *Standards* gewidmet ist. Diese Form von Testnutzen ist in den untersuchten Textstellen häufig nicht gegeben.[26] Deutlich wird dies bei Kuhn (2014): „Der zuletzt genannte Aspekt wird im Folgenden als vernachlässigbar erachtet, da mit dieser Studie derzeit rein grundlagenorientierte Zwecke verfolgt werden, die keine direkte Verwertung des Tests bzw. der Testergebnisse (bspw. zur Feststellung der Eignung für den Lehrerberuf) vorsehen." (S.116)

[26] Ausnahmen bilden Tests zur Evaluation von Lehrveranstaltung oder Trainingsmaßnahmen, z.B. bei Wiesbeck (2015). Im Text selbst wird allerdings nur klar, dass der Test als geeignet für Trainingsevaluationen beurteilt wird. Ob die Testwerte tatsächlich genutzt wurden, um über die Fortführung des entwickelten Trainings zu entscheiden, bleibt offen.

Das obige Zitat aus Kirschner (2013) bietet noch einen weiteren Diskussionspunkt. In der Arbeit werden die „reliablen und validen Test“ verwendet, um deren Zusammenhänge in Stichproben mit unterschiedlichem Studienfortschritt bzw. unterschiedlicher Berufserfahrung zu untersuchen. In anderen Studien der Lehrerbildung werden die Veränderungen interner Strukturen in Abhängigkeit des Studienfortschritts dahingegen als Evidenz für Konstruktvalidität präsentiert (z.B. für die Unterrichtsplanung bei [angehenden] Physiklehrkräften, Stender et al., 2014). Hier stellt sich die Frage, wo die Validierung einer Testwertinterpretation endet und wo Forschung im Sinne des Erkenntnisgewinns beginnt. Die Perspektiven lassen sich jedoch beide vereinen. Die Plausibilität einer Testwertinterpretation wird nicht endgültig, sondern nur bis zum Vorliegen neuer Erkenntnisse bewertet. Auch ein nomologisches Netzwerk unterliegt Veränderungen. Neue Forschungsergebnisse können Anpassungen und Präzisierungen ermöglichen oder einzelne Teile des Netzwerks in Frage stellen (Cronbach & Meehl, 1955). Bei einer Testwertinterpretation können auch gesellschaftliche Entwicklungen eine Neubewertung der Plausibilität erforderlich machen (AERA, APA & NCME, 2014). Folgt man dieser Argumentation, ist die Validierung einer Testwertinterpretation mit einem Forschungsprozess gleichzusetzen.

Damit bleibt jedoch die Frage, wie mit widersprüchlichen Ergebnissen im Validierungsprozess umzugehen ist. Sprechen widersprüchliche Validitätsevidenzen gegen die Plausibilität einer Testwertinterpretation oder gegen die den Interpretationen zugrundeliegenden theoretischen Annahmen? Cronbach & Meehl (1955) empfehlen, dies von der Stärke der bisherigen Evidenzen für eine Theorie bzw. ein nomologisches Netzwerk abhängig zu machen. Je mehr stützende Evidenzen vorliegen, desto glaubwürdiger ist die durch das Netzwerk abgebildete Theorie. Umgekehrt sollten gerade bei noch schwachen Netzwerken die zugrundeliegenden Theorien überprüft werden, wenn widersprüchliche Evidenzen präsentiert werden.

Einschränkungen

In den folgenden Abschnitten wird auf inhaltliche und methodische Einschränkungen der Analysen aus Kapitel 4 eingegangen.

Anpassung des Validierungsschemas. Für insgesamt 10 Textstellen wurde (teilweise) keine Zuordnung in das Validierungsschema kodiert. Sechs der Textstellen fragten nach Einflüssen von individuellen Unterschieden der Studierenden auf Ausprägung der Kompetenz

oder Entwicklungsverläufe der Kompetenz. Diese umfassen Faktoren, die bereits vor Beginn des Studiums ansetzen (Bsp.: „In welchem Umfang beispielsweise unterscheiden sich angehende Lehrkräfte bereits am Anfang ihres Studiums in bereichsspezifischen Lernvoraussetzungen? Haben Unterschiede in solchen Lernvoraussetzungen differenzielle Entwicklungen in der Lehrerausbildung zur Folge?“, König & Herzmann, 2011; S. 187), als auch studienbegleitende Faktoren. Beispielsweise betrachten Wolter und Schiener (2014) die Frage, „ob Jobben neben dem Studium den Studienerfolg – mutmaßlich aus Zeitgründen – beeinträchtigt" (S.67). Aus der Perspektive eines Angebots-Nutzungs-Modells (z.B. Seidel, 2014) zur Erklärung von Leistungen von Schüler*innen (bzw. Studierenden), betrachten die oben angesprochenen Literaturen mit dem Zusammenhang von Testwerten zu individuellen Faktoren die Seite der „Nutzung“. Das Validierungsschema umfasst jedoch durch die Kategorie *Lehre* ausschließlich die Angebotsseite von Unterricht (bzw. Lehre). Die individuellen Voraussetzungen der Studierenden und die Frage, ob und in welchem Umfang angebotene Lerngelegenheiten genutzt werden, werden nicht im Validierungsschema abgebildet.

Angesichts der wenigen Textstellen, die nicht in das Validierungsschema eingeordnet werden können, scheint diese Einschränkung jedoch

Datenlage. Die Datenlage für die Untersuchung der Forschungsfragen weist einige Informationslücken auf. Die Literaturrecherche ergab einige Treffer zur Evaluation von Trainings. Wenn diese das Instrument zur Erfassung des Lernerfolgs nicht beschrieben und keine Angaben zu Gütekriterien machten, wurden die Texte nicht in der weiteren Analyse berücksichtigt. Der mögliche Testnutzen „Evaluation von Lehrveranstaltungen“ taucht daher bislang im Schema nicht auf, da in den Publikationen zur Evaluation von Trainingsmaßnahmen nicht die Entwicklung oder Validierung der verwendeten Lernerfolgsmaße beschrieben wurden. Zum anderen kann auch in den eingeschlossenen Texten selbst ein Informationsdefizit bestehen, da den Autor*innen weniger wichtig erscheinende Einzelheiten möglicherweise gekürzt wurden, um Zeichenbeschränkungen der Publikationsorgane einzuhalten. Allerdings kann bei Veröffentlichungen davon ausgegangen werden, dass Leser*innen ausreichend Informationen erhalten, um die Plausibilität einer Testwertinterpretation einschätzen zu können. Zum anderen werden Tests auch zur Nachnutzung veröffentlicht. Einer der *Standards* fordert Testentwickler*innen dazu auf, Informationen zum intendierten Einsatzzweck, den dafür aufgestellten Testwertinterpretationen und den dazugehörigen Validitätsevidenzen bereitzustellen. So können Nachnutzer*innen prüfen, ob der Test für den gewünschten

Einsatzzweck geeignet ist. Um die Nachnutzung zu ermöglichen, sollten also in der Publikation eines neu entwickelten Tests immer entsprechende Angaben bereitgestellt werden.

Auffallend ist, dass selbst bei den Artikeln, die sich explizit auf Standards oder Kane (2013) beziehen, nicht immer eine klare Formulierung der Testwertinterpretationen oder Grundannahmen zu finden ist. Für Kane selbst stellt die Identifikation der Grundannahmen den wichtigsten Schritt im Validierungsprozess dar, da für „versteckte" Grundannahmen auch keine Evidenzen gesucht werden. Dies stellte die größte Herausforderung in der Analyse der Daten und gleichzeitig die wichtigste Einschränkung dar. Mögliche Testwertinterpretationen mussten aus den vorhandenen Informationen für einen Test abgeleitet werden, um diese einordnen zu können. Auch wenn sich aus dem Testnutzen, dem Studiendesign und z.B. der gewählten Stichprobe plausible Testwertinterpretationen rekonstruieren ließen, beruhen diese doch auf einer Interpretation eines Raters. Welche Testwertinterpretationen die Verfasser*innen der Texte verfolgten, ist nicht eindeutig. Aufgrund des Umfangs wurde keine Volltextanalyse durchgeführt sondern lediglich einzelne Textabschnitte analysiert. Daher könnten Testwertinterpretationen bei der Selektion relevanter Textabschnitte übersehen worden sein.

Zudem basieren die Ergebnisse auf Literaturen bis Sommer 2017. Ob und wenn ja wie aktuellere Testentwicklungen den argumentationsbasierten Validierungsansatz umsetzen, kann in dieser Arbeit nicht beurteilt werden.

Beurteilung der Qualität von Validitätsevidenzen. Eine weitere Einschränkung bezüglich des methodischen Vorgehens ergibt sich aus der fehlenden Analyse der Qualität der bereitgestellten Validitätsevidenzen. Die Bandbreite der untersuchten Tests umfasst mehrere Fachrichtungen der Geistes-, Sozial- und Naturwissenschaftler und deren Fachdidaktiken. Die theoretischen Grundlagen der spezifischen Kompetenzen in den Fachgebieten sind für die Autorin dieser Dissertation meist unbekannt. So kann etwa nicht beurteilt werden, ob es angemessen ist, für pädagogisches Fachwissen in Physik ein psychometrisch besser passendes zwei-dimensionales Modell zugunsten eines theoriekonformen drei-dimensionalen Modells zu verwerfen. In dieser Arbeit wurde daher nur betrachtet, ob die im Text beschriebenen Validitätsargumente zu den von den Autor*innen beschriebenen (wenn auch nicht explizit genannten) Testwertinterpretationen passen. Die Beurteilung der Angemessenheit der Validitätsevidenzen und der Qualität des Validitätsarguments kann nicht in allen Fällen geleistet werden.

Anwendungsbeispiel: Validierung der Testwertinterpretationen im Ko-NaMa-Projekt

5 Das Ko-NaMa-Projekt

Das vorgestellte Modell zur Einordnung von Testwertinterpretationen wird für das Validierungskonzept des Projekts „Ko-NaMa - Simulationsbasierte Messung und Validierung eines Kompetenzmodells für das Nachhaltigkeitsmanagement“ (Ko-NaMa) angewandt. Das Projekt wurde vom Bundesministerium für Bildung und Forschung gefördert (Förderkennzeichen 01PK15010) und im Rahmen der Initiative „Kompetenzmodellierung im Hochschulsektor“ (KoKoHS) durchgeführt. Ko-NaMa war ein Verbundprojekt zwischen der Georg-August-Universität Göttingen (GAU) und dem DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation (DIPF). Das Projektziel war ein Kompetenzmodell für Nachhaltigkeitsmanagement zu validieren und Erklärungsfaktoren für die Kompetenzausprägung und Kompetenzentwicklung bei Studierenden sowie im Vergleich von Hochschulinstitutionen zu bestimmen (Seeber, Hartig, Dierkes & Schumann, 2016). Die Projektpartner an der GAU waren schwerpunktmäßig für die Instrumentenentwicklung und die technische Umsetzung der Testinstrumente verantwortlich. Die Arbeitsschwerpunkte des DIPF lagen auf der psychometrischen Modellierung und den zur Beantwortung der Projektfragestellungen notwendigen Analysen.

Dieses Kapitel stellt zentrale Arbeiten im Projekt vor, welche die Grundlage für die Validierung von Testwertinterpretationen bilden. Daher wird zunächst Kompetenz in Nachhaltigkeitsmanagement definiert und das Kompetenzmodell beschrieben. Dann wird das im Projekt verfolgte Validitätskonzept einschließlich der Testwertinterpretationen dargestellt. Abschließend werden die entwickelten Instrumente beschrieben. Diese bilden die Basis für Validierung der im Projekt verfolgten Testwertinterpretationen, welche in den Kapiteln 6 bis 8 besprochen wird. Anschließend werden empirische Evidenzen für die im Projekt verfolgten Testwertinterpretationen vorgestellt. Dabei wird nach dem in der Einleitung vorgestellten argumentationsbasierten Validitätsansatz vorgegangen. Die Evidenzen je Testwertinterpretation wurden aus dem ab Kapitel 4 in dieser Arbeit entwickelten Schema abgeleitet.

5.1 Kompetenz in Nachhaltigkeitsmanagement

Im Ko-NaMa Projekt wird Kompetenz in Nachhaltigkeitsmanagement definiert als die komplexe Fähigkeit, die zum Teil widerstreitenden Ziele von sozialen, ökologischen und ökonomischen Dimensionen wirtschaftlichen Handelns zu berücksichtigen (Seeber et al., 2019). Im klassischen

Triple-Bottom-Line Ansatz werden die drei Dimensionen Ökonomie, Ökologie und Soziales gleichbedeutend gewichtet (Elkington, 1998). Im ökonomischen Triple-Bottom-Line Ansatz wird die ökonomische Dimension im wirtschaftlichen Handeln stärker gewichtet (Lehmann, 2014), da ein gewinnbringendes Unternehmen die Voraussetzung für die Berücksichtigung von Nachhaltigkeitsaspekten darstellt. Im Ko-NaMa-Projekt wird die Annahme zugrunde gelegt, dass die Gewichtung der drei Dimensionen nicht festgelegt ist, sondern je nach Entscheidungssituation zwischen den Gewichtungen im klassischen und ökonomischen Ansatz wechselt.

In einem vorherigen Projekt konnte gezeigt werden, dass primär kognitive Aspekte Handlungsintentionen zu Nachhaltigkeitsmanagement beeinflussen (Seeber, Fischer, Michaelis & Müller, 2014; Michaelis, 2017). Die kognitiven Aspekte von Kompetenz in Nachhaltigkeitsmanagement im Ko-NaMa-Projekt setzen sich zusammen aus deklarativem Wissen über 1) Betriebswirtschaftslehre, 2) Nachhaltigkeit aus gesamtgesellschaftlicher Sicht und 3) Nachhaltigkeitsmanagement, sowie dem 4) strategischen und Entscheidungswissen über Nachhaltigkeitsmanagement (vgl. Seeber, Hartig, Dierkes & Schumann, 2016). Die ersten drei Aspekte unterscheiden sich nach ihren Inhalten. Der Unterschied zwischen dem dritten und vierten Aspekt liegt in der qualitativen Form der Wissensrepräsentation, die auf der Klassifikation von Shavelson, Ruiz-Primo und Wiley (2005) basieren. Die Autoren klassifizieren Wissen in vier Bereiche: Deklaratives Wissen, Prozedurales Wissen, Schematisches Wissen und Strategisches Wissen. Diese bilden eine geordnete Rangreihe mit deklarativem Wissen als Kategorie, welche die niedrigste Kategorie von Wissen repräsentiert, die Studierende erlernen können. Deklaratives Wissen wird als domänenspezifisches Wissen über Fakten, Definitionen und Beschreibungen verstanden. Als höchste Fähigkeitsstufe wird strategisches Wissen verstanden, welches konzeptualisiert ist als das Wissen darüber, wann, wo und wie Strategien und domänenspezifische Heuristiken angewandt werden.

Im Ko-NaMa-Projekt werden die deklarativen Wissensdimensionen als Bedingungsfaktoren für das strategische und Entscheidungswissen über Nachhaltigkeitsmanagement angesehen. Grundlegendes deklaratives Wissen über Betriebswirtschaftslehre muss vorhanden sein, um ein Unternehmen profitabel zu führen. Erst dann können auch soziale und ökologische Aspekte wirtschaftlichen Handelns berücksichtigt werden. Dafür müssen grundlegende soziale und ökologische Wirkzusammenhänge im globalen Wirtschaftssystem (deklaratives Wissen über Nachhaltigkeit aus gesamtgesellschaftlicher Perspektive) und Methoden zum Einbezug sozialer und ökologischer Aspekte in die unternehmerische Entscheidungsfindung (deklaratives Wissen

zu Nachhaltigkeitsmanagement) bekannt sein. Sind diese drei deklarativen Wissensdimensionen vorhanden, können unternehmerische Entscheidungen unter Berücksichtigung der ökonomischen, ökologischen und sozialen Dimension getroffen werden (strategisches und Entscheidungswissen zu Nachhaltigkeitsmanagement).[27] Die Rationale dahinter lautet, dass Nachhaltigkeitsmanagement nur bei einer auf Dauer erfolgreichen wirtschaftlichen Unternehmung gelingt (Seeber et al., 2019).

5.2 Validierungskonzept im Ko-NaMa-Projekt

Im Ko-NaMa-Projekt wurden Testinstrumente zur Erfassung der Kompetenz Nachhaltigkeitsmanagement bei wirtschaftswissenschaftlichen Studierenden entwickelt. Als Testnutzen ist das Abbilden von Lernerfolgen bei Studierenden vorgesehen. Es wurden drei Testwertinterpretationen verfolgt:

1) Unterschiede in den Testwerten spiegeln interindividuelle Unterschiede in der Ausprägung der Nachhaltigkeitsmanagementkompetenz wider.

2) Testwertinterpretation: Die Testwerte sind Indikatoren für beruflich relevante Kompetenzen im Bereich Nachhaltigkeitsmanagement.

3) Testwertinterpretation: Die Testwerte sind Indikatoren für hochschulisch vermittelte Kompetenzen.

Die Testwertinterpretationen lassen sich in das ab Kapitel 4 entwickelte Schema einordnen (siehe Abbildung 4). Testwertinterpretation 1 bezieht sich ausschließlich auf den Zusammenhang von Test zu Konstrukt. Die anderen beiden Testwertinterpretationen lassen sich nicht eindeutig einem Zusammenhang zuordnen. Beide beinhalten neben ihrem Schwerpunkt auf dem Zusammenhang von Test zu beruflichen Anforderungen (Testwertinterpretation 2) bzw. zu Lehre (Testwertinterpretation 3) auch den Zusammenhang von Test zu Konstrukt.

[27] In dieser Arbeit stehen die kognitiven Aspekte der Kompetenz Nachhaltigkeitsmanagement im Vordergrund. Im Kompetenzmodell werden auch motivational-affektive Aspekte berücksichtigt. Bisherige Forschungsergebnisse zeigen, dass die 1) *Einstellung zu Nachhaltigkeit* einen Einfluss auf die Intention hat, ökologische und soziale Aspekte in der Unternehmenspraxis zu berücksichtigen (Seeber et al., 2014; Michaelis, 2017). Im Kompetenzmodell werden ebenfalls 2) das *Interesse an nachhaltigkeitsrelevanten Themen*, 3) *epistemologische Überzeugungen zu Nachhaltigkeit*, 4) die *wahrgenommene Verhaltenskontrolle im nachhaltigen Handeln*, 5) die *Motivation nachhaltig zu Handeln* und 6) *allgemeine Selbstwirksamkeit und Selbstregulation* als mögliche Einflussfaktoren auf Kompetenz in Nachhaltigkeitsmanagement konzipiert.

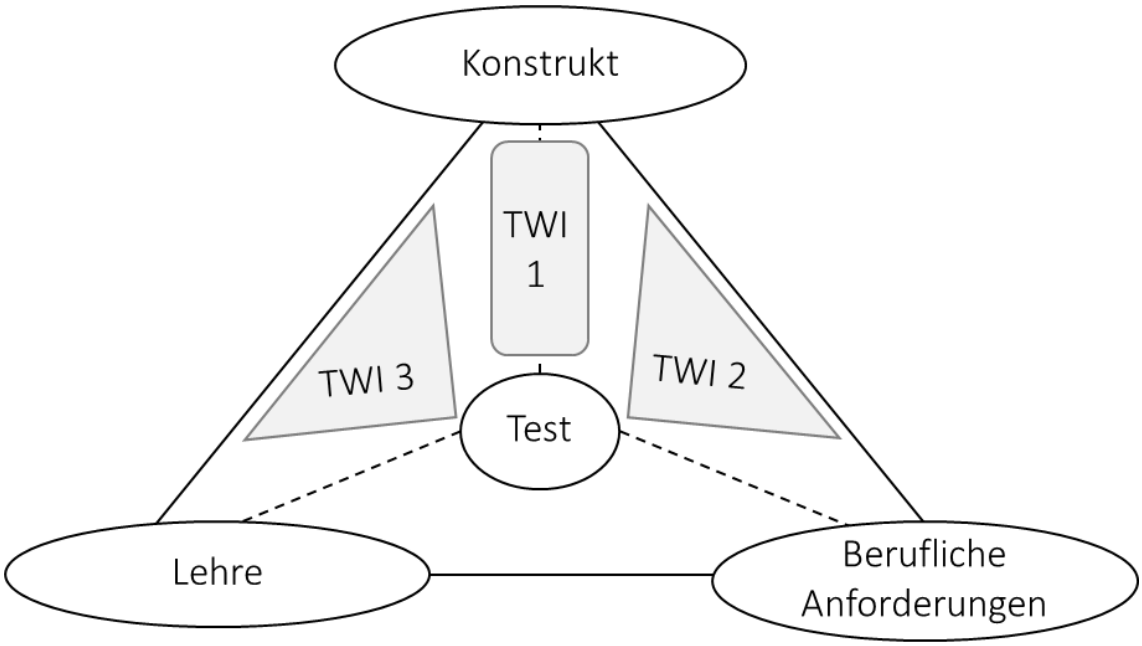


Abbildung 4 Einordnung der Testwertinterpretationen des Ko-NaMa-Projektes in das Validierungsschema

Anmerkungen. *TWI* = Testwertinterpretation.

Zur Validierung der Testwertinterpretationen wurde ein argumentationsbasierter Validierungsansatz gewählt (AERA et al., 2014). Dieser gliedert sich in vier Schritte (Hartig, Frey & Jude, 2020):

1) **Formulierung der intendierten Testwertinterpretationen.** Zunächst formulieren die Testentwickler*innen, welche Interpretationen sie mit den Testwerten verknüpfen und welche Schlüsse aus den Testwerten gezogen werden können.

2) **Formulierung der Grundannahmen**. Anschließend werden die Annahmen expliziert, auf denen eine Testwertinterpretation basiert. Die Grundannahmen müssen so formuliert sein, dass sie empirisch prüfbar sind.

3) **Evidenzen für Grundannahmen prüfen**. Für jede der Grundannahmen werden unterstützende oder widerlegende Evidenzen untersucht. Kane (2013) sieht hierin den wichtigsten Schritt der Validierung von Testwertinterpretationen, da für nicht explizierte Grundannahmen keine Evidenzen gesucht werden.

4) **Zusammenfassung der Plausibilität der Testwertinterpretation**. Basierend auf den Evidenzen zu den Grundannahmen wird ein Fazit bezüglich der Plausibilität der Testwertinterpretation gezogen. Dieses Fazit ist als vorläufig zu betrachten und gilt bis zum Vorliegen weiterer Evidenzen oder gesellschaftlicher Entwicklungen, die eine Veränderung der Testwertinterpretation notwendig machen können.

In den Kapiteln 6 bis 8 wird das Validierungsvorgehen für je eine der Testwertinterpretation beschrieben. Der Schwerpunkt dieser Arbeit liegt auf der Validierung von Testwertinterpretation 3, die in Kapitel 8 erfolgt. Evidenzen für Testwertinterpretation 3 kommen aus Analysen, die Lernfortschritte in den Leistungstests auf unterschiedliche Lerngelegenheiten zurückführen. Daher wird im folgenden Unterkapitel die Operationalisierung von Lerngelegenheiten vorgestellt. Anschließend werden die Testinstrumente, in denen Lernfortschritte erklärt werden sollen, beschrieben.

5.3 Operationalisierung von Lerngelegenheiten

In Testwertinterpretation 3 sollen Testwerte als Indikatoren hochschulisch vermittelter Kompetenzen validiert werden. Grundlegende Annahme des Validitätsarguments ist, dass Lernfortschritte vorwiegend auf hochschulische Lerngelegenheiten zurückzuführen sind. Außerhochschulische Lerngelegenheiten sollten keinen Erklärungswert für Lernfortschritte haben.

5.3.1 Hochschulische Lerngelegenheiten

Die Operationalisierung hochschulischer Lerngelegenheiten erfolgte auf zwei Arten.

Klassifikation auf Kursebene

Zum einen beurteilte das Ko-NaMa Projektteam die Lehrveranstaltungen danach, ob Lerngelegenheiten zu Nachhaltigkeitsthemen angeboten wurden. Dazu wurden geplante Lehrmaterialien analysiert und Rücksprache mit der verantwortlichen Lehrperson gehalten. Lehrveranstaltungen mit solchen Lerngelegenheiten wurden als *Schwerpunktgruppe* klassifiziert, Lehrveranstaltungen ohne Lerngelegenheiten zu Nachhaltigkeitsthemen wurden als *Kontrollgruppe* klassifiziert.

Diese Klassifikation erfolgte im Zuge der Stichprobenrekrutierung. Sie basiert auf dem geplanten Curriculum einer Lehrveranstaltung, was vom implementierten Curriculum abweichen kann

(Naumann, Musow, Aichele, Hochweber & Hartig, 2019). Außerdem gibt diese Einteilung keine Auskunft darüber, ob Studierende in anderen Lehrveranstaltungen bereits relevante Lerngelegenheiten erfuhren. Deshalb wurden auch die Studierenden um ihre Einschätzung gebeten.

Selbstberichte der Studierenden

Die Fragebogenskalen zu hochschulischen Lerngelegenheiten wurden im Ko-NaMa-Projekt entwickelt (Seeber et al., 2020). Die Verfasserin der Arbeit unterstützte die Entwicklung durch die Bereitstellung von Gütekriterien zu Items und Skalen und faktoriellen Strukturprüfungen nach Pilotierung und Haupterhebung.

Die Studierenden wurden gebeten, ihre bisherigen hochschulischen Lerngelegenheiten zu unterschiedlichen Nachhaltigkeitsthemen auf einer vierstufigen Likert-Skala von 1 (*Trifft gar nicht zu*) bis 4 (*Trifft voll zu*) einzuschätzen. Die Skalen lauten „Normative Leitideen von Nachhaltigkeit“ (*NoNa*), „allgemeiner Gesellschaftlicher Nachhaltigkeitsdiskurs“ (*GeNa*) und „betriebliches Nachhaltigkeitsmanagement“ (*BetNa*). Die Inhalte wurden dabei so gewählt, dass sie den Inhalten der Leistungstests zuzuordnen sind. Während die Skalen NoNa und GeNa Inhalte aus dem Test zur gesamtgesellschaftlichen Perspektive zu Nachhaltigkeit thematisieren, werden in der Skala BetNa Inhalte abgefragt, die in den Tests zu Nachhaltigkeitsmanagement zu finden sind. Die Itemtexte und deskriptive Statistiken auf Itemebene sind in Tabelle 5 abgebildet. In Tabelle 7 sind die Skalenmittelwerte getrennt nach Schwerpunkt- und Kontrollgruppe pro Messzeitpunkt abgebildet.

Die Skalen zu selbstberichteten hochschulischen Lerngelegenheiten wurden auf ihre Plausibilität hin geprüft. Die Antworten der Studierenden aus der Schwerpunkt- und Kontrollgruppe sollten die Informationen aus der Einteilung von Schwerpunkt- und Kontrollgruppe replizieren. Die Werte in den Skalen werden als valide Indikatoren hochschulischer Lerngelegenheiten betrachtet wenn

1) Studierende aus der Schwerpunktgruppe zum zweiten Messzeitpunkt mehr hochschulische Lerngelegenheiten berichten als Studierende der Kontrollgruppe, $t(140.82) = 9.80$, $p < .001$, 95% Konfidenzintervall für die wahre Mittelwertdifferenz [0.738 – 1.110].

2) Studierende aus der Schwerpunktgruppe zum zweiten Messzeitpunkt signifikant mehr bisherige hochschulische Lerngelegenheiten berichten als zum ersten Messzeitpunkt, $t(38) = 9.33$, $p < .001$, 95% Konfidenzintervall für die wahre Mittelwertdifferenz [0.84 – 1.31].

3) Studierende der Kontrollgruppe zum zweiten Messzeitpunkt nicht mehr hochschulische Lerngelegenheiten berichten als zum ersten Messzeitpunkt, $t(71) = 0.82$, $p = .42$, 95% Konfidenzintervall für die wahre Mittelwertdifferenz [-0.06 – 0.15].

Alle drei Ergebnisse sprechen für die Plausibilität der Interpretation, dass die Fragebogenwerte der Skalen zu hochschulischen Lerngelegenheiten als Indikatoren für hochschulische Lerngelegenheiten zu Nachhaltigkeitsthemen interpretiert werden können.

5.3.2 Außerhochschulische Lerngelegenheiten

Die Fragebogenskalen basieren auf Vorarbeiten von Michaelis (2017) und wurden im Ko-NaMa-Projekt weiterentwickelt. Die Verfasserin der Arbeit unterstützte die Entwicklung durch die Bereitstellung von Gütekriterien zu Items und Skalen und faktoriellen Strukturprüfungen nach Pilotierung und Haupterhebung.

Die außerhochschulischen Lerngelegenheiten wurden durch Selbstberichte der Studierenden erfasst. Die Studierenden schätzten ihre bisherigen Lerngelegenheiten auf einer vierstufigen Likert-Skala von 1 (*Trifft gar nicht zu*) bis 4 (*Trifft voll zu*) ein. Die Items beziehen sich auf das Informationsverhalten der Studierenden in Bezug auf den Themenbereich „Natur und Umwelt“ (*NaUm*), zu „gesellschaftlichen Themenbereichen wie Problemen und Arbeitsbedingungen in Entwicklungsländern“ (*InGe*) und zur „Informationshäufigkeit zu nachhaltigkeitsrelevanten Themen“ (*NaIn*). Die Itemtexte sowie deskriptive Statistiken sind in Tabelle 6 abgebildet. In Tabelle 7 sind die Skalenmittelwerte getrennt nach Schwerpunkt- und Kontrollgruppe pro Messzeitpunkt abgebildet.

Tabelle 5 Skalen und Items der selbsteingeschätzten hochschulischen Lerngelegenheiten sowie deskriptive Statistiken auf Itemebene je Messzeitpunkt

Bitte bewerten Sie, inwieweit die angegebenen Themen in Ihrem Studium bisher thematisiert wurden	MZP 1					MZP 2				
	N	*M* (*SD*)	*Min*	*Max*	fehlende Werte in %	*N*	*M* (*SD*)	*Min*	*Max*	fehlende Werte in %
Normative Leitideen von Nachhaltigkeit (NoNa)		1.84 (1.38)					2.23 (1.16)			
Dimensionen der Nachhaltigkeit (Ökonomie, Ökologie & Soziales)	300	2.01 (0.96)	1	4	40%	155	2.52 (1.1)	1	4	69%
Intergenerationelle Gerechtigkeit (für zukünftige Generationen)	296	1.8 (0.88)	1	4	41%	155	2.16 (1.01)	1	4	69%
Intragenerationelle Gerechtigkeit (zwischen derzeit lebenden Generationen)	294	1.78 (0.89)	1	4	41%	153	2.18 (1.02)	1	4	69%
Normen und Standards der Nachhaltigkeit (z.B. deutscher Nachhaltig-keitskodex)	296	1.66 (0.83)	1	4	41%	155	2.06 (0.98)	1	4	69%
Allgemeingesellschaftlicher Nachhaltigkeitsdiskurs (GeNa)		1.72 (1.37)					2.16 (1.16)			
Ökologischer Fußabdruck	299	1.6 (0.86)	1	4	40%	154	2.0 (1.00)	1	4	69%
Erneuerbare Energien	293	1.7 (0.90)	1	4	41%	153	2.1 (1.00)	1	4	69%
Nachhaltigkeitssiegel (z. B. Bio, Fair-Trade, MSC, blauer Engel,)	298	1.6 (0.86)	1	4	40%	156	2.2 (1.13)	1	4	69%
Klimawandel	298	1.9 (0.93)	1	4	40%	151	2.2 (1.03)	1	4	69%
Betriebliches Nachhaltigkeits-management (BetNa)		1.97 (1.69)					2.36 (1.26)			
Corporate Social Responsibility	295	1.97 (0.99)	1	4	40%	154	2.45 (1.03)	1	4	69%
Nachhaltigkeitsorientierte SWOT-Analyse	292	2.03 (1.05)	1	4	41%	154	2.24 (0.97)	1	4	69%
Nachhaltigkeitsorientierte Strategien	293	1.90 (0.92)	1	4	41%	155	2.43 (1.01)	1	4	69%
Externe Kosten (Kosten die Dritte tragen müssen, wie z.B. Schadstoffemissionen)	293	2.30 (1.00)	1	4	41%	154	2.61 (0.94)	1	4	69%
Ökobilanz	294	1.59 (0.84)	1	4	41%	155	2.06 (1.05)	1	4	69%

Anmerkungen. 4-stufige Likert-Skala mit den Ausprägungen 1 = *Trifft gar nicht zu*, 2 = *Trifft eher nicht zu*, 3 = *Trifft eher zu*, 4 = *Trifft voll zu*. FETT gedruckt sind die Skalennamen. *N* = Antworthäufigkeit. Fehlende Werte sind prozentual zur gesamten Längsschnittstichprobe angegeben (N_{Ges} = 499). Zum zweiten Messzeitpunkt liegen nur noch Daten von 189 Personen vor. Darauf bezogen (N_2 = 189) weisen die selbsteingeschätzten Lerngelegenheiten zum zweiten Messzeitpunkt einen Anteil von 1.8% fehlender Werte auf.

Tabelle 6 Skalen und Items der selbsteingeschätzten außerhochschulischen Lerngelegenheiten sowie deskriptive Statistiken auf Itemebene je Messzeitpunkt

	MZP 1					MZP 2				
Wie oft tätigen Sie nachfolgende Handlungen zu den angegeben Themen?	*N*	*M* (*SD*)	*Min*	*Max*	fehlende Werte in %	*N*	*M* (*SD*)	*Min*	*Max*	fehlende Werte in %
Themenbereich Natur und Umwelt (NaUm)		2.48 (1.69)					2.52 (1.24)			
Dokumentationen gucken	286	2.82 (0.96)	1	4	43%	154	2.84 (0.97)	1	4	69%
Bücher lesen	287	1.84 (0.91	1	4	42%	153	1.94 (0.86)	1	4	69%
Berichte in Tageszeitungen lesen	287	2.67 (0.90)	1	4	42%	154	2.75 (0.93)	1	4	69%
Berichte in Zeitschriften lesen	286	2.53 (0.92)	1	4	43%	153	2.56 (0.91)	1	4	69%
Im Internet gezielt nach Informationen suchen	287	2.56 (0.99)	1	4	42%	154	2.47 (1.06)	1	4	69%
Gesellschaftliche Themen wie Probleme und Arbeitsbedingungen in Entwicklungsländern (InGe)		2.36 (1.64)					2.31 (1.16)			
Dokumentationen gucken	288	2.57 (0.95)	1	4	42%	154	2.5 (0.90)	1	4	69%
Bücher lesen	286	1.82 (0.86)	1	4	43%	153	1.84 (0.80)	1	4	69%
Berichte in Tageszeitungen lesen	287	2.52 (0.92	1	4	42%	154	2.55 (0.92)	1	4	69%
Berichte in Zeitschriften lesen	286	2.40 (0.91)	1	4	43%	153	2.37 (0.93)	1	4	69%
Im Internet gezielt nach Informationen suchen	287	2.46 (1.00)	1	4	42%	154	2.31 (0.95)	1	4	69%
Informationshäufigkeit zu nachhaltigkeitsrelevanten Themen (NaIn)		2.47 (1.47)					2.52 (1.05)			
Wirtschaftliche Probleme in Entwicklungsländern	287	2.34 (0.80)	1	4	42%	155	2.41 (0.82)	1	4	69%
Ökologische Probleme in Entwicklungsländern (z. B. Auswirkungen von Monokulturen)	287	2.21 (0.83)	1	4	42%	155	2.28 (0.86)	1	4	69%
Soziale Probleme in Entwicklungsländern (z. B. Arbeitsbedingungen)	285	2.47 (0.85)	1	4	43%	155	2.50 (0.82)	1	4	69%
Klimawandel	286	2.67 (0.86)	1	4	43%	155	2.72 (0.79)	1	4	69%
Durch menschliches Handeln ausgelöste Naturkatastrophen	287	2.64 (0.90)	1	4	42%	155	2.67 (0.86)	1	4	69%

Anmerkungen. 4-stufige Likert-Skala mit den Ausprägungen 1 = *Trifft gar nicht zu*, 2 = *Trifft eher nicht zu*, 3 = *Trifft eher zu*, 4 = *Trifft voll zu*. **FETT** gedruckt sind die Skalennamen. Fehlende Werte sind prozentual zur gesamten Längsschnittstichprobe angegeben (N_{Ges} = 499). Zum zweiten Messzeitpunkt liegen nur noch Daten von 189 Personen vor. Darauf bezogen (N_2 = 189) weisen die selbsteingeschätzten Lerngelegenheiten des zweiten Messzeitpunkts einen Anteil von 1.8% fehlenden Werten auf.

Tabelle 7 Unterschiede zwischen Schwerpunkt- und Kontrollgruppe in der mittleren Ausprägung hochschulischer und außerhochschulischer Lerngelegenheiten je Messzeitpunkt

		Schwerpunktgruppe		Kontrollgruppe	
	Skala	*M*	*(SD)*	*M*	*(SD)*
	Messzeitpunkt 1				
Hochschulische Lerngelegenheiten	NoNa	1.96	(1.34)	1.78	(1.43)
	GeNa	1.94	(1.39)	1.62	(1.33)
	BetNa	1.96	(1.58)	1.97	(1.76)
Außerhochschulische Lerngelegenheiten	NaIn	2.73	(1.28)	2.36	(1.53)
	NaUm	2.72	(1.47)	2.39	(1.78)
	InGe	2.55	(1.40)	2.28	(1.75)
	Messzeitpunkt 2				
Hochschulische Lerngelegenheiten	NoNa	2.94	(1.05)	1.76	(0.91)
	GeNa	2.81	(1.09)	1.73	(0.93)
	BetNa	2.74	(1.27)	2.11	(1.17)
Außerhochschulische Lerngelegenheiten	NaIn	2.72	(1.11)	2.38	(0.99)
	NaUm	2.71	(1.26)	2.39	(1.20)
	InGe	2.53	(1.15)	2.16	(1.13)

Anmerkungen. Angegeben sind Skalenmittelwerte und –standardabweichungen je Gruppe und Messzeitpunkt.

5.4 Testinstrumente

Für jede der kognitiven Aspekte im Kompetenzmodell zu Nachhaltigkeitsmanagement (Seeber et al., 2016) wurde eine Testkomponente entwickelt. Diese basieren zum Teil auf Entwicklungen aus einem vorherigen Projekt, mussten aber auf die Zielgruppe der Studierenden angepasst werden[28].

Die Testentwicklung wird detailliert im Rahmen einer Dissertation eines Projektkollegen dargestellt, der für die Testentwicklung verantwortlich war. Die Autorin dieser Dissertation unterstützte die Entwicklung der Testinstrumente bei der Datenerhebung und durch die Analysen der Pilotierungs- und Haupterhebungsdaten.

[28] Die Instrumente zur Erfassung deklarativen Wissens von Nachhaltigkeit aus gesamtgesellschaftlicher Perspektive und betriebswirtschaftlichem Wissen waren für den Einsatz bei Auszubildenden in kaufmännischen Berufen entwickelt worden (Projekt „KONWIKA - Entwicklung und Prüfung eines Kompetenzmodells für ein nachhaltiges Wirtschaften kaufmännischer Auszubildender“). In einer ersten Erprobung der bestehenden Instrumente wurde festgestellt, dass ein Großteil der Items aus dem betriebswirtschaftlichen Test für Studierende zu leicht war.

Zunächst werden die Tests inhaltlich beschrieben. In den Unterkapiteln folgen Informationen zur Testadministration in der Pilotierung und der Haupterhebung. Anschließend werden IRT-basierte Gütekriterien aus der Pilotierung und Haupterhebung berichtet.

Deklaratives betriebswirtschaftliches Wissen

Der Test zu deklarativem betriebswirtschaftlichem Wissen (BWL) umfasst die Unterbereiche 1) Absatz und Marketing, 2) Beschaffung und Logistik, 3) Finanzwirtschaft, 4) Personalwirtschaft, 5) Planung, 6) Produktion, 7) Rechnungswesen, 8) Unternehmen und Management. Diese Unterbereiche wurden von Inhaltsexperten nach Durchsicht von Lehrbüchern, die an unterschiedlichen Hochschulen im Grundstudium BWL verwendet werden[29]. Der Test besteht aus Single-Choice Items mit jeweils einer richtigen Lösung und drei Distraktoren. Die Testkomponente erfasst Wissen, welches die Grundlage für langfristig erfolgreiches Unternehmertum darstellt (Seeber et al., 2019). Der Test zu deklarativem Wissen über Betriebswirtschaftslehre ist daher der ökonomischen Dimension im Kompetenzmodell zu Nachhaltigkeitsmanagement zuzuordnen.

Deklaratives Wissen zu Nachhaltigkeit aus gesamtgesellschaftlicher Perspektive

Der Test zu deklarativem Wissen zu Nachhaltigkeit aus gesamtgesellschaftlicher Perspektive (NagP) basiert auf Vorarbeiten aus dem Bereich der beruflichen Bildung (Michaelis, 2017). Das im NagP-Test erfasste Wissen kann zum einen durch Verfolgen gesellschaftlicher und politischer Debatten erworben werden. Dazu zählen Prinzipien von Nachhaltigkeit und Ziele internationaler Vereinbarungen zu nachhaltiger Entwicklung, Vor- und Nachteile verschiedener Energieträger und globale Wirtschaftszusammenhänge. Zum anderen wird auch Faktenwissen erfasst, welches nur durch nähere Beschäftigung mit dem jeweiligen Thema erworben werden kann. Beispielsweise sei hier das Wissen über den CO^2-Abdruck verschiedener Heizsysteme oder der Beifang in Kilogramm in der Fischerei genannt. Der Test besteht aus Single-Choice Items mit jeweils einer richtigen Lösung und drei Distraktoren. Der NagP-Test kann den Dimensionen Ökologie und Soziales im Kompetenzmodell von Seeber et al. (2016) zugeordnet werden.

[29] Quelle: Persönliches Gespräch mit D. Siepelmeyer, wissenschaftlicher Mitarbeiter am Lehrstuhl von Prof. Dierkes. Prof. Dierkes und sein Mitarbeiter waren im Ko-NaMa-Projekt mitverantwortlich für die Instrumentenentwicklung der Leistungstests.

Deklaratives Wissen über Nachhaltigkeitsmanagement

Der Test zu deklarativem Wissen über Nachhaltigkeitsmanagement (dNCM) gliedert sich in die zwei inhaltlichen Unterbereiche Controlling und Management. Während klassische betriebswirtschaftliche Ansätze neben legalen Anforderungen nur ökonomische Aspekte berücksichtigen, müssen in den beiden Bereichen auch die ökologische und soziale Zieldimension eines Unternehmens berücksichtigt werden. Im Bereich Controlling wird deklaratives Wissen zu Controllingmethoden erfasst (z.B. die Definition von Ökobilanz, Sachbilanz oder Details zu Materialflussrechnungen). Der zweite Unterbereich erfasst deklaratives Wissen zu Managementmethoden (z.B. die Definition von passivem oder aktivem Nachhaltigkeitsmanagement, Aspekte des Life-Cycle Assessments und zur umweltorientierten Portfolioanalyse).

Der Test besteht aus Single-Choice Items mit jeweils einer richtigen Lösung und drei Distraktoren. Die Items fokussieren Management- und Controllingmethoden zur Berücksichtigung von ökologischen und sozialen Aspekten. Daher ist der dNCM-Test allen Dimensionen (Ökonomie, Ökologie und Soziales) im Kompetenzmodell von Seeber et al. (2016) zuzuordnen.

Strategisches Wissen über Nachhaltigkeitsmanagement

Im Unterschied zu den vorherigen Testkomponenten wird im Test zu strategischem Wissen über Nachhaltigkeitsmanagement (sNCM) nicht nur deklaratives Wissen abgefragt. Studierende sollen in möglichst realen Unternehmenskontexten Entscheidungen treffen und diese begründen. Dabei muss zwischen den konkurrierenden Zieldimensionen Ökonomie, Ökologie und Soziales abgewogen werden. So werden etwa im klassischen Triple-Bottom-Line Ansatz die drei Dimensionen gleichrangig berücksichtigt (Elkington, 1998). Ein modifizierter Triple-Bottom-Line Ansatz gewichtet hingegen die ökonomische Dimension höher, da diese als Grundlage für langjährige erfolgreiche Unternehmungen gesehen wird (Lehmann, 2014). Die Gewichtung dieser drei Zieldimensionen wird im Ko-NaMa Projekt jedoch nicht als fest angenommen. Vielmehr müssen entsprechend der Entscheidungssituation Gewichte festgelegt werden, die den aktuellen Anforderungen an das Unternehmen gerecht werden (Seeber et al., 2019). Gemeinsam ist diesen Ansätzen, dass die Berücksichtigung der ökologischen und sozialen Dimension über gesetzliche Vorgaben hinausgeht. Im sNCM-Test soll Nachhaltigkeitsmanagement entlang der gesamten Wertschöpfungskette eines Unternehmens erfasst werden.

Für die Erfassung des strategischen Wissens zu Nachhaltigkeitsmanagement wurde eine technologie-basierte Simulation in Kooperation mit einem Fahrradhersteller entwickelt. Der Test besteht aus 13 Situationen, die in verschiedenen Unternehmensbereichen spielen. Jede Situation wird durch ein Video eingeführt, in dem eine Problemstellung aufgeworfen wird. Anschließend bearbeiten die Studierenden vier bis sieben Items zu dieser Situation. Für einige Items müssen ergänzende Materialien berücksichtigt werden, welche im Testsystem bereitgestellt werden. Wenn nachfolgende Items auf Lösungen vorhergehender Items aufbauen, wird die von den Testentwicklern als korrekte Lösung eingestufte Antwort im Arbeitsauftrag des nachfolgenden Items präsentiert. Itemformate sind Single- und Multiple-Choice Items, offene Antworten für Begründungen oder Ergebnisse von Berechnungen, Gewichtung von Kriterien und Zuordnungsaufgaben. Diese Antwortformate sind überwiegend in eine Grafik eingebunden, die eine Mail an z.B. Vorgesetzte oder andere Beteiligte im Unternehmen darstellt. Von den 72 in der Haupterhebung verwendeten Items werden 14 Items dichotom bepunktet, die anderen 58 Items weisen ein mehrstufiges Rating auf.

Für den sNCM-Test wurde ein Testheftdesign erstellt, da in der Pilotierung die Bearbeitung einer Situation durchschnittlich zwischen acht und 18 Minuten dauerte und die Testzeit auf 90 Minuten begrenzt war. In 13 Testheften wurden jeweils drei bis vier Situationen zusammengefasst. Kriterien für die Zusammenstellung waren die Variation der kognitiven Anforderungen der Items in den Situationen und die Bearbeitungszeit, die etwa bei 45 Minuten liegen sollte. Dadurch bearbeiteten die Studierenden je nach Testheft nur 17 bis 27 der 72 Items im Test sNCM. Das Testheftdesign mit 13 Testheften für 13 Aufgabenstämme und vier Aufgaben pro Testheft hat normalerweise den Vorteil, dass jede Situation mit jeder anderen Situation einmal in einem Testheft vorgegeben wird. Dabei taucht jede Situation in vier Testheften auf und kann dementsprechend an jeder Position in einem Testheft je einmal gesetzt werden (Youden-Square Design, vgl. Frey, Hartig & Rupp, 2009). Die Situationen 9 und 10, sowie die Situationen 12 und 13, wurden aufgrund ihrer Ähnlichkeit jedoch nicht gemeinsam in einem Testheft präsentiert. In den ersten beiden Situationen muss jeweils eine Nutzwertanalyse durchgeführt werden, die letzten beiden Situationen spielen beide im Bereich Human Resources. Deshalb wurde in einem Testheft die Situation 9 und in einem anderen Testheft die Situation 12 nicht vorgegeben. Für die Situationen 9 liegen daher keine gemeinsamen Beobachtungen mit den Situationen 2, 4 und 10 vor. Für die Situation 12 liegen keine gemeinsamen Beobachtungen mit den Situationen 2, 5 und 13 vor.

Der sNCM-Test ist allen Dimensionen (Ökonomie, Ökologie und Soziales) im Kompetenzmodell von Seeber et al. (2016) zuzuordnen.

5.4.1 Testadministration

Die Pilotierung der neu entwickelten Items fand von Februar 2016 bis Juli 2017 statt. An der GAU und an der Goethe-Universität Frankfurt wurden mittels Aushängen über die Testungen informiert. Angesprochen wurden Studierende aus wirtschaftswissenschaftlichen und wirtschaftsnahen Studiengängen, wie z.B. Wirtschaftspädagogik, Wirtschaftsinformatik oder Volkswirtschaftslehre. Die Testungen fanden in Computerräumen der beiden Universitäten statt. Über die dortigen Computerarbeitsplätze loggten sich die Studierenden mit zufällig generierten Zugangsdaten auf dem Ko-NaMa-Testsystem ein. Innerhalb von 90 Minuten bearbeiteten sie randomisiert präsentierte Items. Für die Erprobung der situativen Items lagen zusätzlich Taschenrechner, Kopfhörer sowie Stift und Papier bereit. Die Studierenden erhielten als Dank für die Bearbeitung einen Gutschein für Amazon oder einen Kinobesuch im Wert von 10 Euro.

Insgesamt wurden 1080 Studierende in der Pilotierung erreicht. Nach Abschluss der Pilotierung liegen für die deklarativen Tests pro Item zwischen 222 und 420 Antworten vor, für den situativen Test 25 bis 185 Antworten.

Haupterhebung

Die Haupterhebung wurde zwischen Oktober 2017 und Juli 2018 durchgeführt. Die Stichprobe umfasst 20 Lehrveranstaltungen an unterschiedlichen deutschen Hochschulstandorten. Die Rekrutierung erfolgte durch persönliche Kontakte der Projektbeteiligten, es handelt sich also um eine Gelegenheitsstichprobe. Jede Lehrveranstaltung wurde in einem wirtschaftswissenschaftlichen oder wirtschaftsnahen Studiengang angeboten. Lehrveranstaltungen aus der Schwerpunkt- und Kontrollgruppe wurden jeweils einmal zu Beginn des Semesters getestet, in dem die relevante Lehrveranstaltung stattfand, und einmal zwei bis drei Wochen vor Ende des Semesters. Dieser Zeitpunkt wurde gewählt, um möglichst wenige Studierende wegen anstehender Klausuren zu verlieren. Zusätzlich wurden Studierende aus zwölf Lehrveranstaltungen ohne Bezug zu Nachhaltigkeitsthemen einmalig getestet (*Surveygruppe*). Das Studiendesign ist in Abbildung 5 dargestellt.

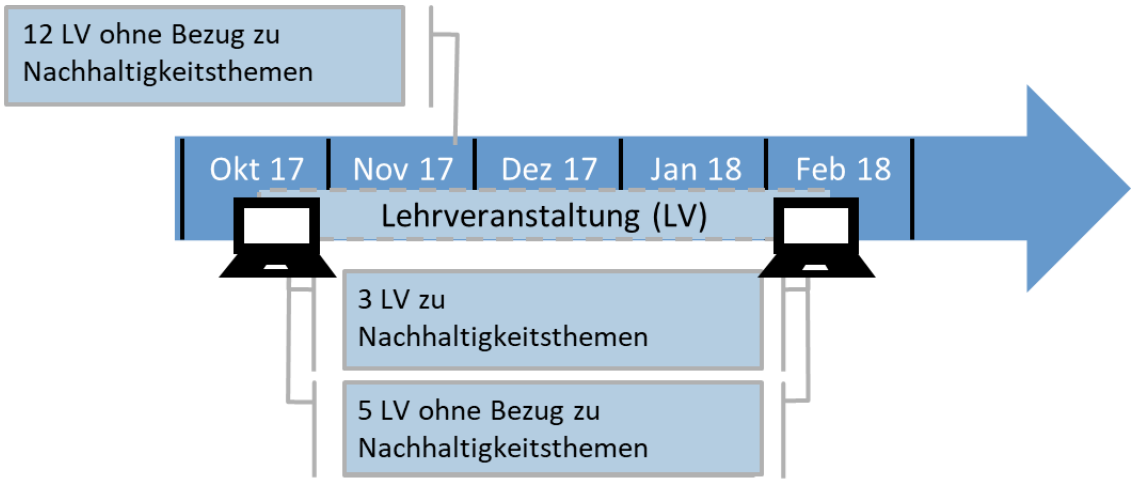


Abbildung 5 Studiendesign der Haupterhebung im Ko-NaMa-Projekt

Die Testungen fanden während der regulären Vorlesungszeit der Lehrveranstaltungen statt und waren auf 90 Minuten begrenzt. Die Testbearbeitung erfolgte auf vom Ko-NaMa-Projekt bereitgestellten Laptops, auf denen zu Beginn der Testung die Startseite des Testsystems bereits geöffnet war. An jedem Arbeitsplatz lag ein zufällig generierter Zugang zum Testsystem. Zusätzlich standen für die Testbearbeitung Kopfhörer, Taschenrechner, Stifte und Papier zur Verfügung. Zu Beginn jeder Testung wurden die Studierenden über die Freiwilligkeit ihrer Teilnahme und bestehende Datenschutzregelungen informiert. Es wurde darauf hingewiesen, dass die Möglichkeit zur nachträglichen Löschung von Daten besteht. Dazu sollten die Studierenden ihren Testzugang aufbewahren, da andernfalls keine Zuordnung von Daten zu einer bestimmten Person möglich gewesen wäre.

Zunächst wurden die Studierenden gebeten, sich in das Testsystem einzuloggen. Anschließend erläuterte die Testleitung die Funktionsweise des Testsystems und machte auf die Möglichkeit aufmerksam, bei technischen Schwierigkeiten die Testleitung anzusprechen. Anschließend begann die Testung, in der die Studierenden die Tests in der vom System präsentierten Reihenfolge bearbeiten sollten. Zuerst sollten die vier Leistungstests BWL, NagP, dNCM und sNCM bearbeitet werden. Die Studierenden hatten für die ersten drei Tests jeweils zehn Minuten Bearbeitungszeit, in denen die Items randomisiert vorgegeben wurden. Die Studierenden wurden im Testsystem über die verbleibende Bearbeitungszeit informiert. Der Test zu sNCM war zeitlich nicht beschränkt. Im sNCM-Test bearbeitete jede Person ein Testheft. Nach der Bearbeitung der Leistungstests wurden die Studierenden um ihre Einschätzung gebeten, in wie weit die Testumgebung geeignet ist, berufliche Situationen handlungsnah

abzubilden. Anschließend wurden sie zu ihren demographischen Angaben, Details ihrer hoch-/schulischen und beruflichen Ausbildung sowie ihren Interessen, Einstellungen und Handlungsmotivation bezüglich Nachhaltigkeitsthemen befragt. Zehn Minuten vor Ende der Testung wurden die Studierenden über die verbleibende Zeit informiert. Zum Dank für die Teilnahme an der Datenerhebung erhielt jede Testperson einen Kino-Gutschein.

Umgang mit nicht bearbeiteten Items

Im Testsystem konnten Items übersprungen werden. In diesen Fällen wurde das Item als *falsch* bewertet. Wurde ein Item während der Testzeit nicht vom Testsystem vorgegeben, wurde ein *fehlender Wert* vergeben.

5.4.2 IRT-Skalierung der Tests

Die Skalierung der Tests erfolgte in R (R Core Team, 2020) mithilfe des TAM Pakets (Robitzsch, Kiefer & Wu, 2020).

Aufgrund der relativ hohen Itemzahlen pro Test im Vergleich zur Testzeit, war absehbar, dass pro Item keine hohen Antworthäufigkeiten realisiert werden können. In der Haupterhebung wurden zum ersten Messzeitpunkt 872 Studierende erreicht. Für die einzelnen Items liegen im BWL-Test durchschnittlich 293 (*Min* = 249, *Max* = 312), im NagP-Test 401 (*Min* = 353, *Max* = 431), im dNCM-Test 395 (*Min* = 369, *Max* = 414) und im sNCM-Test 224 (*Min* = 157, *Max* = 274) Antworten pro Item vor. Die Tests BWL, NagP und dNCM sollten deshalb mit einem 1pl-Modell skaliert werden, wofür mindestens 100 bis 200 Beobachtungen empfohlen werden (DeMars, 2010). Der Test sNCM enthält mehrstufige Items und sollte mit einem Partial Credit Modell (PCM) skaliert werden. Bei dreistufigen Items werden im PCM mindestens 250 Antworten empfohlen (DeMars, 2010).[30]

Das 1pl-Modell stellt eines der einfachsten IRT-Modelle dar. Zusätzlich zur Personenfähigkeit θ schätzt dieses Modell nur einen Itemparameter, die Itemschwierigkeit β eines Items. Die Itemschwierigkeit für ein Item in einem 1pl-Modell gibt den Punkt an, in dem eine Person mit

[30] Im Vergleich zu 1pl-Modellen bzw. PCM werden für die Schätzung von 2pl-Modellen bzw. Generalized Partial Credit Modell (GPCM) mindestens 500 Beobachtungen gefordert (DeMars, 2010). In diesen Modellen wird zusätzlich zur Itemschwierigkeit für jedes Item ein Diskriminationsparameter im 2pl-Modell geschätzt bzw. für jede Kategorie im GPCM.

der gleichen Fähigkeitsausprägung mit 50%iger Wahrscheinlichkeit eine richtige bzw. eine falsche Antwort gibt.

$$P(X_{i,j} = 1 \,|\theta_i, \beta_j) = \frac{\exp(\theta_i - \beta_j)}{1+\exp(\theta_i - \beta_j)} \qquad (1)$$

Gleichung 1 stellt die Modellgleichung eines 1pl-Modells dar für die Wahrscheinlichkeit, dass ein Proband *i* Item *j* löst, gegeben seiner Fähigkeit θ_i und der Itemschwierigkeit β_j.

Das PCM (Masters, 1982) stellt eine Erweiterung des 1pl-Modells auf Items mit mehr als zwei Antwortkategorien dar. Dieses modelliert für jede Kategorie die Wahrscheinlichkeit, bei gegebener Fähigkeit diese Kategorie zu erreichen.

Die Wahrscheinlichkeit einer Person i Kategorie k in Item j zu erreichen ist im PCM definiert als

$$P(X_{ij} = k \,|\theta_i, \beta_{kj}) = \frac{\exp(\theta_i - \beta_{kj})}{1+\exp(\theta_i - \beta_{kj})} \;, \qquad (2)$$

wobei β_{kj} die „Schwierigkeit" ausdrückt, bei Item *j* mit den Kategorien k = 0, 1, 2, … m_j die Kategorie k anstatt k-1 zu wählen. Bis auf das kategorienspezifische Subskript ist Gleichung (2) mit Gleichung (1) für das 1pl-Modell identisch. Das bedeutet, dass in jeder Kategorie k = 0, 1, 2, … m_j das 1pl-Modell gilt. Die Formel des Gesamtmodells (siehe Gleichung 3) ergibt sich aus der zusätzlichen Bedingung, dass die Summe der Wahrscheinlichkeiten aller Kategorien = 1 ist

$$P(X_{ij} = x \,|\theta_i, \beta_j) = \frac{\sum_{k=0}^{x} \exp(\theta_i - \beta_{jk})}{1 + \sum_{m=1}^{M} \exp \sum_{k=1}^{m} (\theta_i - \beta_{jk})} \qquad (3)$$

Die Itemparameter der in dieser Arbeit verwendeten Software (TAM Paket in R, Robitzsch et al., 2020; R Core Team, 2020) geben im PCM die Schnittpunkte der Informationskurven von zwei nebeneinanderliegenden Punktkategorien an (*Delta Parameter* nach Masters). An diesem Schnittpunkt liegt die Antwort einer Person, die diese Fähigkeitsausprägung besitzt, mit gleich hoher Wahrscheinlichkeit in einer der Kategorien. Zur einfacheren Interpretation der Itemparameter im PCM können Schwellenparameter (*Thurstonian Thresholds*) berechnet werden. Thurstonian Thresholds sind definiert als die Fähigkeitsausprägung, bei der die Wahrscheinlichkeit diese oder eine höhere Punktzahl zu erzielen, bei 50% liegt.

5.4.2.1 Modellannahmen

Die Modelle sind zur Schätzung einer Dimension spezifiziert, die Daten müssen also eindimensional sein. Wenn die Daten eindimensional sind, hängt die Korrelation zwischen Items nur von dem interessierenden Konstrukt ab. Unter Kontrolle der Fähigkeit ist die Antwort auf ein Item somit unabhängig von der Beantwortung aller anderen Items (lokale Unabhängigkeit). Die Residuen der Items sollten also nicht korreliert sein. Weiterhin sollte das Modell korrekt spezifiziert sein, was über eine Passung der Daten zum Modell überprüft werden kann.

Eindimensionalität der Testinstrumente

Die Eindimensionalität der einzelnen Instrumente wurde im Rahmen einer Publikation des Ko-NaMa-Projekts (Seeber et al., 2019) über konfirmatorische Faktorenanalysen geprüft. Die Ergebnisse sprechen für die Eindimensionalität der Tests, mit Einschränkung für den BWL-Test. In allen Tests mussten Items mit negativer oder nicht signifikanter Faktorladung ausgeschlossen und lokale Itemabhängigkeiten modelliert werden. Nach Anpassung der Tests zeigte sich für jeden Test ein akzeptabler bis guter Modellfit. Zu beachten ist jedoch, dass die Analysen für den sNCM-Test nicht auf Ebene der Items sondern für die durchschnittlich erreichten Punkte in einer Situation durchgeführt wurden. Die Ergebnisse werden detaillierter in Kapitel 6.2.2.1 dieser Dissertation als Evidenzen für eine Testwertinterpretation beschrieben.

Model-Data-Fit

Beide verwendeten Modelle unterstellen allen Items gleiche Diskrimination. Deshalb werden Items ausgeschlossen, deren Diskrimination von der modellimplizierten Diskrimination abweichen. Als Kriterium wurde der Infit und Outfit eines Items, bzw. einzelner Punktkategorien im PCM betrachtet. Infit und Outfit basieren auf den standardisierten mittleren quadrierten Residuen aller Antworten, also auf den Abweichungen von tatsächlich gegebenen Antworten zu den vom Modell vorhergesagten Antworten. Während beim Outfit alle Residuen gleich gewichtet werden, werden im Infit die Residuen nach ihrer Nähe zur Itemschwierigkeit gewichtet. Für beide gilt als Faustregel, dass die *t*-Werte zwischen -2 und 2 liegen sollten und damit in etwa ein 95% Konfidenzintervall abdecken (Ames & Penfield, 2015). Liegen die *t*-Werte außerhalb dieses Bereiches, trennen die Items signifikant schlechter (bei $t > 2$) bzw. signifikant besser (bei $t < -2$) zwischen Personen mit hohen und niedrigen Fähigkeitsparametern als die anderen Items im Test.

Zusätzlich wurde die Korrelation von Itemantwort zu Testwert berechnet. Im 1pl-Modell wurden Items mit negativen Werten und Werten < .2 ausgeschlossen (vgl. Kuhn, 2014), im PCM wurde entsprechend die Korrelation der höchsten Punktkategorie eines Items mit dem Testwert betrachtet. Zusätzlich wurde die korrigierte Trennschärfe berechnet als Korrelation einer Itemantwort mit der durchschnittlich erreichten Punktzahl pro Item in einem Test, ohne Berücksichtigung des betreffenden Items selbst. Negative Werte der korrigierten Trennschärfe bedeuten, dass Personen die in anderen Items eher die richtige Antwort wählen bei diesem Item falsch antworten und umgekehrt. Im PCM wurde die korrigierte Trennschärfe für das Gesamtitem berechnet. Bei auffälligen Werten wurde hier die Korrelation der einzelnen Punktkategorien zum Testwert betrachtet. Diese sollten mit aufsteigender Kategorie innerhalb eines Items größer werden. Items, bei denen höhere Punktkategorien eine niedrigere Korrelation zum Testwert aufwiesen als niedrigere Punktkategorien, wurden ausgeschlossen.

Für Tests, bei denen vorherige empirische Befunde Hinweise auf andere als eindimensionale Strukturen lieferten, wurden Modellvergleiche angestellt. Als Entscheidungskriterien wurden der Likelihood-Ratio-Test (LRT) und Informationskriterien betrachtet. Der LRT favorisiert komplexere Modelle, da die Abweichung von Daten zum Modell kleiner wird bei zusätzlichen Parametern. Außerdem ist der LRT stichprobenabhängig, bei konstanter Parameterzahl und steigendem Stichprobenumfang wird der Chi-Quadrat Wert größer (Schermelleh-Engel, Moosbrugger & Müller, 2003). Daher wird zusätzlich zum LRT die Anzahl der gewonnen Freiheitsgrade mit der Differenz des Chi-Quadrat-Wertes verglichen. Dabei gelten Werte von $\Delta\chi^2 \geq 3*\Delta df$ als kritisch (Schermelleh-Engel, Moosbrugger & Müller, 2003). Informationskriterien treffen keine Aussagen über signifikante Unterschiede in der Modellpassung. Sie erlauben Aussagen darüber, welches von zwei (oder mehr) miteinander zu vergleichenden Modellen besser auf die Daten passt. Modelle mit niedrigeren Werten sind vorzuziehen. Informationskriterien verwenden die (negative, zweifache) Loglikelihood der geschätzten Modelle und einen additiven Term, in dem die Anzahl der zu schätzenden Parameter als Strafe enthalten sind. In dieser Arbeit wird das *Bayesian Information Criterion* (BIC, Schwarz, 1978) als Informationskriterien verwendet (vgl. für dichotome Items Kang & Cohen, 2007; für polytome Items Kang, Cohen & Sung, 2009). Das BIC favorisiert eher sparsame Modelle, da hier der Strafterm die Anzahl der geschätzten Parameter *p* mit dem natürlichen Logarithmus der Stichprobengröße *N* multipliziert.

Außerdem werden die Itemschwierigkeiten bzw. Thurstonian Thresholds für das PCM im sNCM-Test, und Personenfähigkeiten aus dem finalen Modell auf einer gemeinsamen Skala dargestellt

(*Wright Map*). Diese sollten sich möglichst überlappen, da ein Item im Bereich des Schwierigkeitsparameters die höchste Informationsdichte aufweist, hier also besonders gut die Fähigkeit von Personen unterscheiden kann. Die WrightMap gibt Aufschluss darüber, wie nützlich eine Itemmenge für die Unterscheidung von fähigen und weniger fähigen Personen ist.

Lokale Unabhängigkeit

In einem 1pl-Modell bzw. PCM wird lokale Unabhängigkeit der Items vorausgesetzt, daher sollten die Residuen der Items nicht korreliert sein. Um diese Modellannahme zu überprüfen, werden die Q3-Statistiken betrachtet. Als Faustregel gilt dabei, dass die Q3-Statistiken einen Wert von |0.2| nicht überschreiten sollten (Yen, 1993). Diese sind jedoch unter anderem von der Anzahl der verwendeten Items und der Stichprobengröße abhängig (Chen & Thissen, 1997; Christensen, Makransky & Horton, 2017). Beide Artikel kommen zum Schluss, die optimalen Cut-off Werte für Q3-Statistiken für bestimmte Daten durch eine Simulation unter Annahme lokaler Unabhängigkeit zu bestimmen. Dies wurde im Ko-NaMa-Projekt nicht realisiert. Chen und Thissen (1997) zeigen in einer Simulationsstudie, dass bei Tests mit 40 bis 80 Items, wie sie in den Ko-NaMa-Tests vorliegen, und bei Stichproben mit 1000 Personen ein Cut-off von |0.2| die alpha-Fehlerrate von .05 unterschreitet. Allerdings wurden diese Simulationen mit einem 2pl-Modell durchgeführt, welches für die Ko-NaMa-Tests nicht verwendet wird. Christensen et al. (2017) untersuchen den Einfluss von Testlänge und Stichprobengröße auf die Q3-Statistiken in einem 1pl-Modell und einem PCM. Allerdings simulieren die Autoren nur Tests bis zu einer Länge von 20 Items. Darüber hinaus schlagen Christensen et al. (2017) vor, die Q3-Werte relativ zur durchschnittlichen Korrelation der Residualkorrelationen der Items zu betrachten (adjustierte Q3-Statistiken), da die Q3-Werte bei lokaler Unabhängigkeit einen negativen Erwartungswert haben.

Für die Ko-NaMa-Tests liegen nach der Haupterhebung im Median 57 bis 230 Antwortpaare pro Test vor. Bei einer Stichprobengröße von 200 Personen kommen Christensen et al. (2017) für einen Test mit 20 Items im 1pl-Modell und PCM zu einem 95% Perzentil bei .24 für die adjustierten Q3-Werte (aQ3).[31] Da die Tests im Ko-NaMa-Projekt zwischen 51 und 80 Items enthalten (bzw. zwischen 42 und 65 Items nach Ausschluss von nicht passenden Items), wird im

[31] Alle Simulationen des Artikels können auf folgender Website durchgespielt werden: http://publicifsv.sund.ku.dk/~kach/Q3/critical_values_Yens_Q3.html; zuletzt abgerufen am 25.03.2021.

Folgenden ein höherer Cut-Off von |0.3| für aQ3-Werte gewählt (vgl. Aryadoust, Ng & Sayama, 2021; Røe, Damsgård, Fors & Anke, 2014; La Porta et al., 2011).

5.4.3 Test zur Erfassung deklarativen Wissens über Betriebswirtschaftslehre

Pilotierung

In der Pilotierung wurden 129 Items erprobt. Für diese liegen zwischen 221 und 382 Antworten pro Item vor. 15 Items wurden aufgrund der oben definierten Kriterien als nicht zum Modell passend klassifiziert. Da die Testzeit in der Haupterhebung auf 90 Minuten beschränkt war, sollte die Testzeit für den BWL-Test nur 10 Minuten betragen. Um die Antwortzahlen pro Item zu erhöhen, wurden zusätzlich 34 Items ausgeschlossen. Dieser Itemausschluss erfolgte von den Testentwicklern zunächst nach inhaltlichen Kriterien. Von Items mit Bezug zu ähnlichen Inhalten oder Methoden wurde nur eines behalten. Die acht Unterbereiche sollten weiterhin möglichst gleichmäßig vertreten sein. Gleichzeitig wurde darauf geachtet, dass die Itemschwierigkeiten der verbleibenden Items die geschätzten Fähigkeitsparameter der Stichprobe möglichst gut abdecken. Insgesamt wurden 80 Items für die Haupterhebung ausgewählt.

Haupterhebung

Nach der Haupterhebung liegen zwischen 249 und 312 Antworten pro Item vor. Die Studierenden bearbeiteten im Median 18 Items.

Model-Data-Fit. Weder Infit noch Outfit zeigen auffällige *t*-Werte an, kein Item liegt außerhalb des Bereiches [-2; 2]. Jedoch zeigten 15 der Items eine korrigierte Trennschärfe von unter .2. Nach Ausschluss dieser 15 Items liegen die Itemschwierigkeiten aus einer erneuten 1pl Skalierung zwischen -1.10 und 1.68 (M = 0.35, SD = 0.65). Die Varianz der Testwerte beträgt 0.29 und die EAP-Reliabilität liegt bei .53. Die Items scheinen insgesamt etwas schwerer zu sein, als die Personen fähig sind (siehe Abbildung 6). Die Itemschwierigkeiten decken den Bereich der mittleren geschätzten Personenfähigkeiten gut ab. Der Test scheint jedoch weniger geeignet, um zwischen Personen mit Fähigkeitsparametern < -1 zu differenzieren.

Arbeiten zu bisherigen Tests zu betriebswirtschaftlichem Wissen liefern Evidenzen für eindimensionale Modelle, prüfen jedoch häufig auch die Passung eines zwei-dimensionalen Modells, in dem die Unterbereiche Finanzwirtschaft und Rechnungswesen eine separate Dimension bilden (vgl. Lauterbach, 2015; Jähnig, 2014; Seeber, 2008). Der LRT und der Vergleich von gewonnen Freiheitsgraden zur Chi-Quadrat Differenz sprechen für das zwei-dimensionale

Modell, $\chi^2(2) = 12.39$, $p = .002$. Das BIC spricht für das ein-dimensionale Modell, $BIC_{1DIM} = 24233$, $BIC_{2DIM} = 24234$. Da im zwei-dimensionalen Modell die beiden Dimensionen mit $r = .854$ korrelieren, wird das sparsamere Modell beibehalten.

Lokale Unabhängigkeit. Für den BWL-Test liegen die aQ3 Statistiken zwischen -.28 und .30. Die Residualkorrelationen basieren auf 101 bis 162 Antwortpaaren. Der höchste aQ3 Wert liegt bei .30 und damit gerade nicht über dem Cut-off von |.3|. Damit scheint bei den Items im BWL-Test kein nennenswertes Ausmaß an lokaler Abhängigkeit zu bestehen.

Fazit

Insgesamt wird die Passung der Daten des BWL-Tests zum 1pl-Modell als gut eingeschätzt. Dafür mussten jedoch auch nach der Haupterhebung noch Items ausgeschlossen werden. Die Reliabilität der Testwerte ist als niedrig einzuschätzen und nicht für den Einsatz von Individualdiagnostik ausreichend. Die Bearbeitungszeit der deklarativen Tests betrug nur 10 Minuten, in denen die Studierenden im Median 18 Items beantworteten. Durch eine Verlängerung der Bearbeitungszeit könnte die Reliabilität gesteigert werden (nach der Spearman-Brown-Formel, Gäde et al., 2020; zur Diskussion der Steigerung der Reliabilität: Seeber et al., 2019).

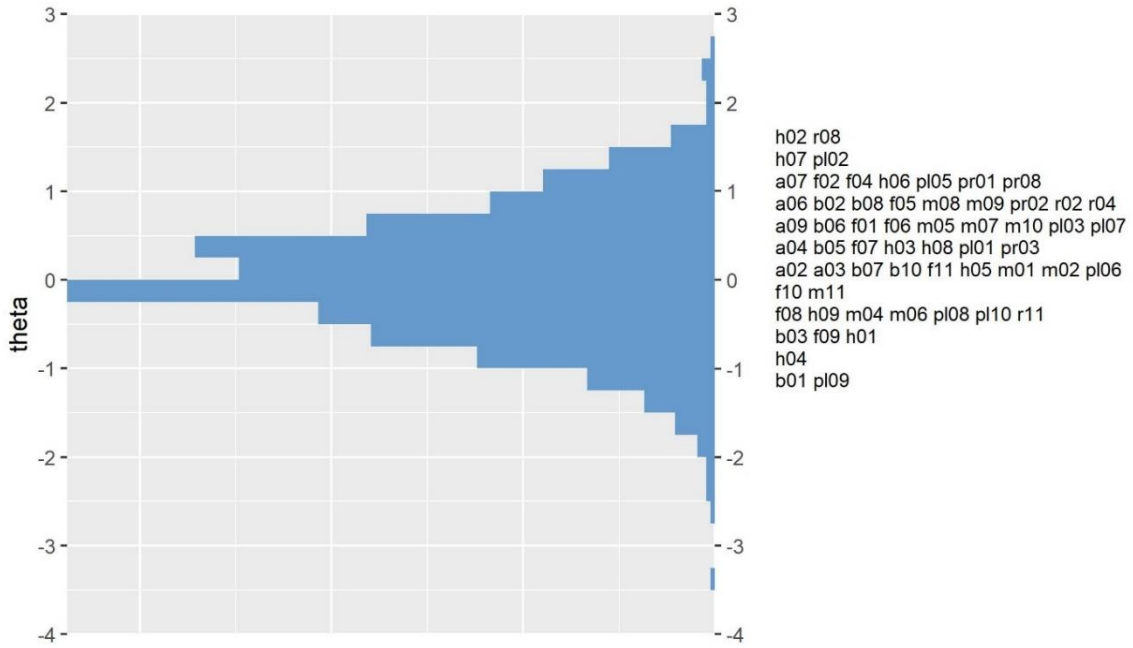


Abbildung 6 WrightMap für den BWL-Test basierend auf den Daten des ersten Messzeitpunkts

5.4.4 Test zur Erfassung deklarativen Wissens über Nachhaltigkeit aus gesamtgesellschaftlicher Perspektive

Pilotierung

In der Pilotierung wurden 66 Items erprobt. Für diese liegen zwischen 212 und 319 Antworten pro Item vor. Drei Items zeigten einen *t*-Wert von Infit und/ oder Outfit kleiner -2, zwei weitere Items zeigen einen Wert größer 2. Diese fünf Items wurden ausgeschlossen. Da die Testzeit in der Haupterhebung auf 90 Minuten beschränkt war, sollte die Testzeit für den NagP-Test nur 10 Minuten betragen. Um die Antwortzahlen pro Item zu erhöhen, wurden zusätzlich 8 Items ausgeschlossen. Dieser Itemausschluss erfolgte von den Testentwicklern zunächst nach inhaltlichen Kriterien. Von Items mit Bezug zu ähnlichen Inhalten wurde nur eines behalten. Gleichzeitig wurde darauf geachtet, dass die Itemschwierigkeiten der verbleibenden Items die geschätzten Fähigkeitsparameter der Stichprobe möglichst gut abdecken. Insgesamt wurden 53 Items für die Haupterhebung ausgewählt.

Haupterhebung

Für die Items des NagP-Tests liegen zwischen 353 und 431 Antworten vor. Die Studierenden bearbeiteten im Median 20 Items.

Model-Data-Fit. Insgesamt sechs Items weisen einen kritischen *t*-Wert im Infit und/ oder Outfit auf. Bei vier Items liegt der *t*-Wert über 2, bei zwei weiteren Items unter -2. Fünf weitere Items wurden ausgeschlossen, da deren Korrelation von Itemantwort zu Testwert unter .2 lagen. Nach Ausschluss dieser elf Items liegen die Itemschwierigkeiten aus einer erneuten 1pl Skalierung zwischen -1.15 und 1.58 (M = -0.00, SD = 0.89). Die Varianz der Testwerte beträgt 0.37 und die EAP-Reliabilität liegt bei .60. Die Items decken einen großen Bereich der geschätzten Personenfähigkeiten ab (siehe Abbildung 7). Lediglich bei Fähigkeitsparametern kleiner -2 sind keine Items vorhanden. Der Test scheint also weniger gut geeignet zur Differenzierung von Personen mit sehr niedrigen Fähigkeitsparametern.

Lokale Unabhängigkeit. Für den NagP-Test liegen die aQ3 Statistiken zwischen -.22 und .23. Die Residualkorrelationen basieren auf 175 bis 254 Antwortpaaren. Der höchste aQ3-Wert liegt bei .23 und überschreitet nicht den Cut-off von |.3|. Damit scheint zwischen den Items im NagP-Test kein nennenswertes Ausmaß an lokaler Abhängigkeit zu bestehen.

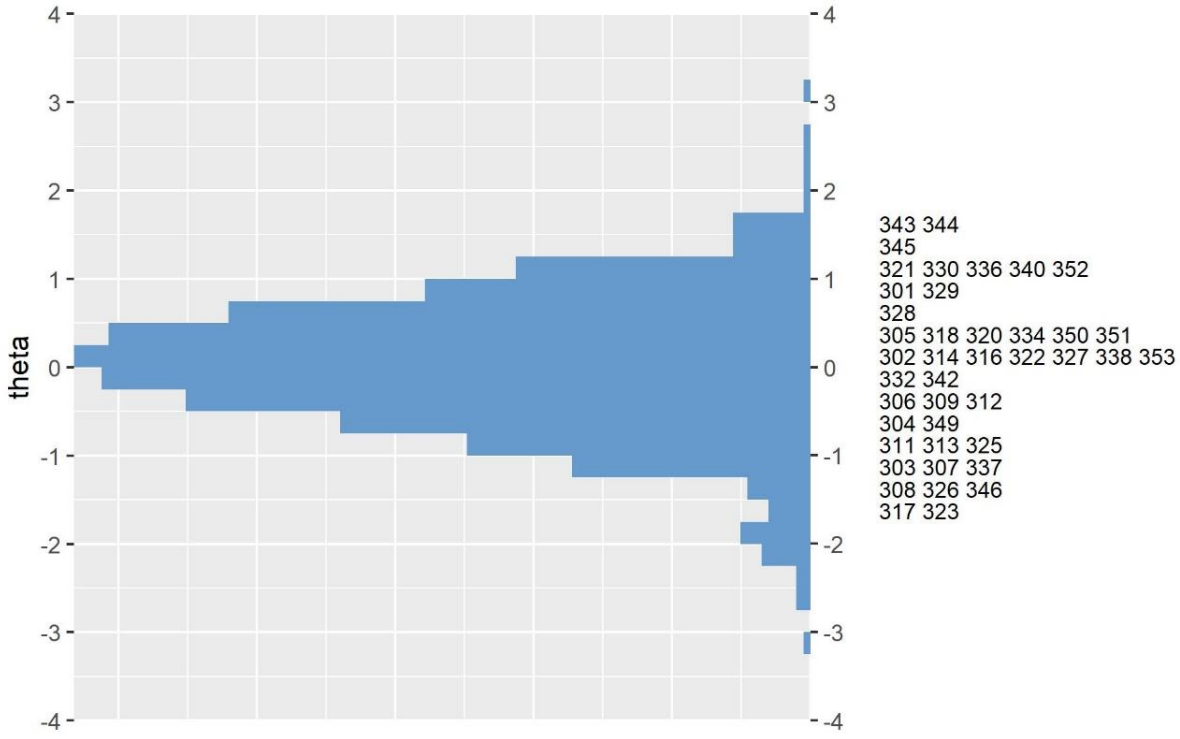


Abbildung 7 WrightMap für den NagP-Test basierend auf den Daten des ersten Messzeitpunkts

Fazit

Insgesamt wird die Passung der Daten des NagP-Tests zum 1pl-Modell als gut eingeschätzt. Dafür mussten jedoch auch nach der Haupterhebung noch Items ausgeschlossen werden. Die Reliabilität der Testwerte ist als niedrig einzuschätzen und nicht für den Einsatz von Individualdiagnostik ausreichend. Wie für den BWL-Test auch, lässt sich jedoch eine Erhöhung der Reliabilität annehmen, wenn die Testbearbeitungszeit verlängert wird.

5.4.5 Test zur Erfassung deklarativen Wissens über Nachhaltigkeitsmanagement

Pilotierung

In der Pilotierung wurden 64 Items erprobt. Für diese liegen zwischen 255 und 305 Antworten pro Item vor. Zwei Items zeigten einen *t*-Wert von Infit und Outfit kleiner -2. Zusätzlich wurden elf Items ausgeschlossen, deren Korrelation von Itemantwort zu Testwert unter .2 lag. Insgesamt wurden 51 Items für die Haupterhebung ausgewählt.

Haupterhebung

Für die Items des dNCM-Tests liegen zwischen 369 und 414 Antworten vor. Die Studierenden bearbeiteten im Median 21 Items.

Model-Data-Fit. Bei jeweils einem Item liegt der *t*-Wert des Infit und/ oder des Outfits bei über 2 bzw. unter -2. Fünf weitere Items wurden ausgeschlossen, da deren Korrelation von Itemantwort zu Testwert unter .2 lagen. Bei den verbleibenden 44 Items liegen die Itemschwierigkeiten aus einer erneuten 1pl Skalierung zwischen -1.39 und 1.76 (*M* = 0.42, *SD* = 0.76). Die Varianz der Testwerte beträgt 0.31 und die EAP-Reliabilität liegt bei .56. Wie der von Null positiv abweichende Mittelwert zeigt und in Abbildung 8 ersichtlich, sind die Itemschwierigkeiten etwas von der Personenfähigkeit verschoben. Die Items scheinen etwas schwerer zu sein, als die getesteten Personen fähig sind. Die Items decken den größten Bereich der geschätzten Personenfähigkeiten ab. Es gibt jedoch nur wenige Items, die ihre höchste Informationsdichte bei Werten unter Null aufweisen. Der Test scheint also weniger gut geeignet zur Schätzung von eher niedrigen Fähigkeitsparametern.

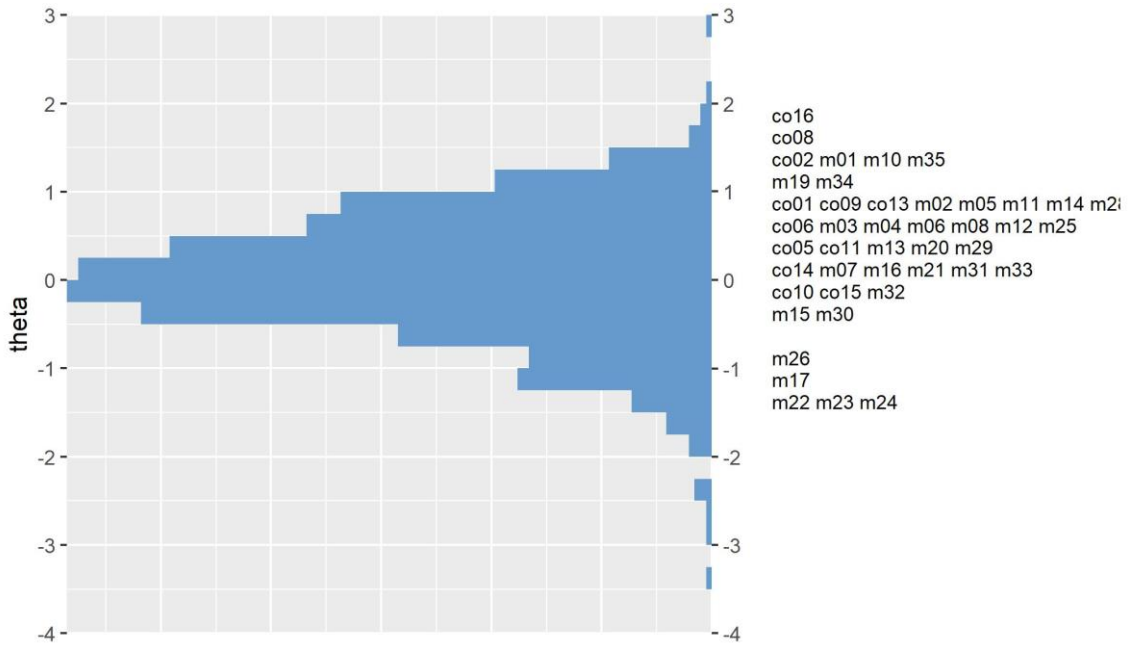


Abbildung 8 WrightMap für den dNCM-Test basierend auf den Daten des ersten Messzeitpunkts

Der dNCM-Test enthält die zwei Unterbereiche „Management“ und „Controlling“. Daher wird das ein-dimensionale 1pl Modell gegen ein zwei-dimensionales Modell getestet, in dem jeder Unterbereich als eine Dimension spezifiziert wird. Der Modellvergleich und der Vergleich von gewonnen Freiheitsgraden zur Differenz im Chi-Quadrat Wert spricht für das zwei-dimensionale Modell, $\chi^2(2) = 12.33$, $p = .002$. Das BIC spricht jedoch für das ein-dimensionale Modell ($BIC_{1DIM} = 21412$, $BIC_{2DIM} = 21413$). Die Dimensionen im zwei-dimensionalen Modell korrelieren hoch mit $r = .86$. Daher wird das sparsamere ein-dimensionale Modell beibehalten.

Lokale Unabhängigkeit. Für den dNCM-Test liegen die aQ3 Statistiken zwischen -.23 und .20. Die Residualkorrelationen basieren auf 203 bis 262 Antwortpaaren. Der maximale aQ3-Wert beträgt .23 und liegt damit unter dem Cut-off von |.3|. Damit scheint zwischen den Items im dNCM-Test kein nennenswertes Ausmaß an lokaler Abhängigkeit zu bestehen.

Fazit

Insgesamt wird die Passung der Daten des dNCM-Tests zum 1pl-Modell als gut eingeschätzt. Dafür mussten jedoch auch nach der Haupterhebung noch Items ausgeschlossen werden. Die Reliabilität der Testwerte ist als niedrig einzuschätzen und nicht für den Einsatz von Individualdiagnostik ausreichend. Wie für den BWL-Test auch, lässt sich jedoch eine Erhöhung der Reliabilität annehmen, wenn die Testbearbeitungszeit verlängert wird.

5.4.6 Test zur Erfassung strategischen Wissens zu Nachhaltigkeitsmanagement

Pilotierung

In der Pilotierung konnten für die Items im sNCM-Test nur zwischen 51 und 152 Antworten pro Item realisiert werden. Insbesondere bei mehrstufigen Items waren die höchsten Punktkategorien gering belegt (*Min* = 1). Die Pilotierung wurde deshalb nicht für die Itemselektion nach den oben definierten Kriterien genutzt. Stattdessen wurde das Scoring Manual getestet und überarbeitet. Dies betraf einerseits die Kodieranweisungen für Antworten auf offenen Items. Andererseits wurden bei mehrstufigen Items mit 4- oder 5-Stufen geprüft, ob Punktkategorien zusammengelegt werden können.

Haupterhebung

Es wurden insgesamt 72 Items in 13 Situationen verwendet. Für die Items liegen zwischen 157 und 274 Antworten vor. Die Studierenden bearbeiteten im Median 21 Items.

Model-Data-Fit. Insgesamt wurden 13 Items aufgrund der oben definierten Kriterien ausgeschlossen. Bei einem drei-stufigen Item liegt der *t*-Wert des Outfits in zwei Punktkategorien über 2, bei einem weiteren drei-stufigen Item liegen die *t*-Werte von Infit und Outfit in einer Punktkategorie über 2. Bei diesem Item liegt außerdem die Korrelation der höchsten Punktkategorie unter .2, deshalb wurde das Item ebenfalls ausgeschlossen. Die korrigierte Trennschärfe ist für vier Items negativ. Bei zwei dieser Items ist steigen die Korrelationen von Punktkategorie zu Testwert nicht entsprechend der Punktkategorien an. Diese Items wurden deshalb ausgeschlossen. Bei einem weiteren Item liegt die Korrelation der höchsten Punktkategorie unter .2, daher wurde auch dieses Item ausgeschlossen. Sieben weitere Items wurden ausgeschlossen weil deren höchste Punktkategorie mit dem Testwert kleiner als .2 korrelierte.

Bei den verbleibenden 59 Items liegen die Itemschwierigkeiten aus einer erneuten Skalierung mit dem PCM zwischen -1.58 und 1.84 (*M* = 0.47, *SD* = 0.83). Die Varianz der Testwerte ist 0.21 und die EAP-Reliabilität liegt bei .58. Für die WrightMap im situativen Test wurden Thurstonian Thresholds verwendet. Wie in Abbildung 9 ersichtlich, liegen viele Punktkategorien im oberen Fähigkeitsbereich, in dem keine der getesteten Personen liegt. Die Items scheinen insgesamt schwerer, als die getesteten Personen fähig. Gleichzeitig gibt es einige wenige Personen mit einer geschätzten Fähigkeit von unter -2. In diesem Bereich liegen kaum Items, bzw. Punktkategorien vor. Der Test scheint weniger gut geeignet zur Schätzung von sehr niedrigen Fähigkeitsparametern. Dafür ist der Test potentiell gut geeignet um Lernfortschritte an der getesteten Stichprobe zu erfassen, da er auch im höheren Fähigkeitsbereich noch ausreichend Items zur Unterscheidung von fähigen und weniger fähigen Personen enthält.

Lokale Unabhängigkeit. Am höchsten sind die aQ3 Statistiken für den sNCM-Test, dort liegen sie zwischen -.51 und 43. Der maximale aQ3-Wert liegt bei .51 und damit deutlich über dem definierten Cut-Off von |.3|. Hier ist jedoch zu berücksichtigen, dass zum Teil nur 16 gemeinsame Beobachtungen vorliegen, im Median 57 Antwortpaare (*Max* = 274). Außerdem liegen für 114 (von 1174) Itemkombinationen keine gemeinsamen Antworten vor, da in zwei Testheften nur drei statt vier Situationen eingesetzt wurden.

Die hohen lokalen Abhängigkeiten könnten durch die Clusterung von Items in Situationen auftreten. Die Korrelationsmatrix der aQ3-Statistiken zeigt jedoch, dass die Itemabhängigkeiten nicht gehäuft innerhalb einer Situation auftreten, da dann besonders hohe Werte nahe der Diagonalen auftreten müssten (siehe Anhang G).

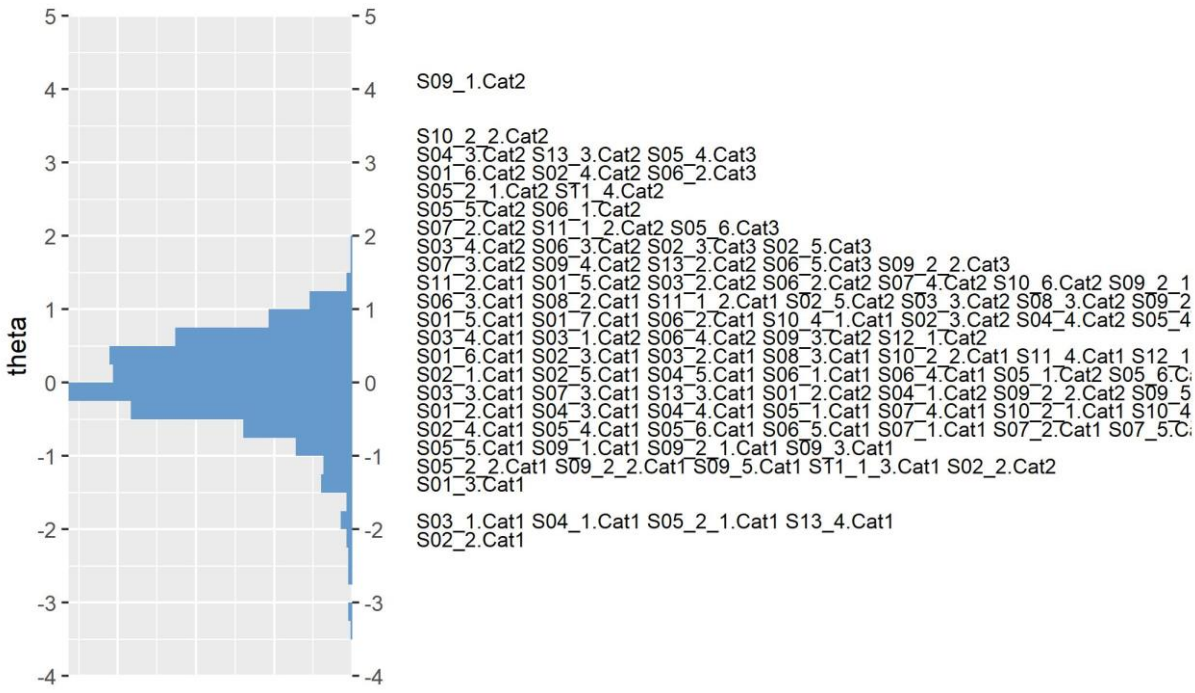


Abbildung 9 WrightMap für den sNCM-Test basierend auf den Daten des ersten Messzeitpunkts

Fazit

Insgesamt wird die Passung der Daten des sNCM-Tests zum PCM als akzeptabel eingeschätzt. Dafür mussten jedoch auch nach der Haupterhebung noch Items ausgeschlossen werden. Die Reliabilität der Testwerte ist als niedrig einzuschätzen und nicht für den Einsatz von Individualdiagnostik ausreichend.

5.4.7 Linking der Messzeitpunkte

Für die Stichprobe mit der Messwiederholung sollten die Testwerte aus beiden Messzeitpunkten vergleichbar sein. Dazu wurden für jeden Test die Rangkorrelation der Itemschwierigkeiten aus für Messzeitpunkte separate Skalierungen verglichen (*freie Skalierung*). Zusätzlich wurde ein Modellvergleich angestellt zwischen der freien Skalierung und einer Skalierung, in der die Itemschwierigkeiten des zweiten Messzeitpunkts auf die des ersten Messzeitpunkts fixiert wurden. Für den Modellvergleich werden wieder LRT, die Differenz der Freiheitsgrade im Verhältnis zur Differenz des Chi-Quadrat-Wertes (Schermelleh-Engel et al., 2003) und das BIC betrachtet.

Für den BWL-Test liegt die Rangkorrelation von Itemschwierigkeiten von erstem und zweiten Messzeitpunkt bei $r_{BWL} = .84$. Beim Modellvergleich zwischen einer freien Skalierung und einer mit fixierten Itemschwierigkeiten für den zweiten Messzeitpunkt, spricht der LRT für eine frei Skalierung, $\chi^2(64) = 125.75$, $p < .001$. Das BIC und $\Delta\chi^2 \geq 3*\Delta df$ favorisieren das sparsamere Modell mit fixierten Itemschwierigkeiten, $BIC_{free} = 5549$, $BIC_{fixed} = 5339$. Daher werden die Itemschwierigkeiten des zweiten Messzeitpunkts auf die des ersten Messzeitpunkts fixiert.

Für den NagP-Test liegt die Rangkorrelation von Itemschwierigkeiten von erstem und zweiten Messzeitpunkt bei $r_{NagP} = .89$. Beim Modellvergleich zwischen einer freien Skalierung und einer mit fixierten Itemschwierigkeiten für den zweiten Messzeitpunkt spricht der LRT für eine frei Skalierung, $\chi^2(41) = 114.23$, $p < .001$. Das BIC und $\Delta\chi^2 \geq 3*\Delta df$ favorisieren das sparsamere Modell mit fixierten Itemschwierigkeiten, $BIC_{free} = 5407$, $BIC_{fixed} = 5306$. Daher werden die Itemschwierigkeiten des zweiten Messzeitpunkts auf die des ersten Messzeitpunkts fixiert.

Für den dNCM-Test liegt die Rangkorrelation von Itemschwierigkeiten von erstem und zweiten Messzeitpunkt bei $r_{dNCM} = .92$. Beim Modellvergleich zwischen einer freien Skalierung und einer mit fixierten Itemschwierigkeiten für den zweiten Messzeitpunkt spricht der LRT für eine frei Skalierung, $\chi^2(43) = 77.26$, $p = .001$. Das BIC und $\Delta\chi^2 \geq 3*\Delta df$ favorisieren das sparsamere Modell mit fixierten Itemschwierigkeiten, $BIC_{free} = 5920$, $BIC_{fixed} = 5772$. Daher werden die Itemschwierigkeiten des zweiten Messzeitpunkts auf die des ersten Messzeitpunkts fixiert.

Für den sNCM-Test liegt die Rangkorrelation von Itemschwierigkeiten von erstem und zweiten Messzeitpunkt bei $r_{sNCM} = .83$. Beim Modellvergleich zwischen einer freien Skalierung und einer mit fixierten Itemschwierigkeiten für den zweiten Messzeitpunkt spricht der LRT für eine frei Skalierung, $\chi^2(114) = 208.49$, $p < .001$. Das BIC und $\Delta\chi^2 \geq 3*\Delta df$ favorisieren das sparsamere Modell mit fixierten Itemschwierigkeiten, $BIC_{free} = 6097$, $BIC_{fixed} = 5710$. Daher werden die Itemschwierigkeiten des zweiten Messzeitpunkts auf die des ersten Messzeitpunkts fixiert.

Pro Test und je Messzeitpunkt wurden wle-basierte Personenparameter geschätzt (Warm, 1989). Diese wurden auf eine Skala mit einem Mittelwert von 50 und einer Standardabweichung von 10 für den ersten Messzeitpunkt transformiert. In Abbildung 10 bis Abbildung 13 sind die mittleren Personenparameter pro Erhebungsgruppe und Messzeitpunkt unter Angabe der Standardfehler abgebildet.

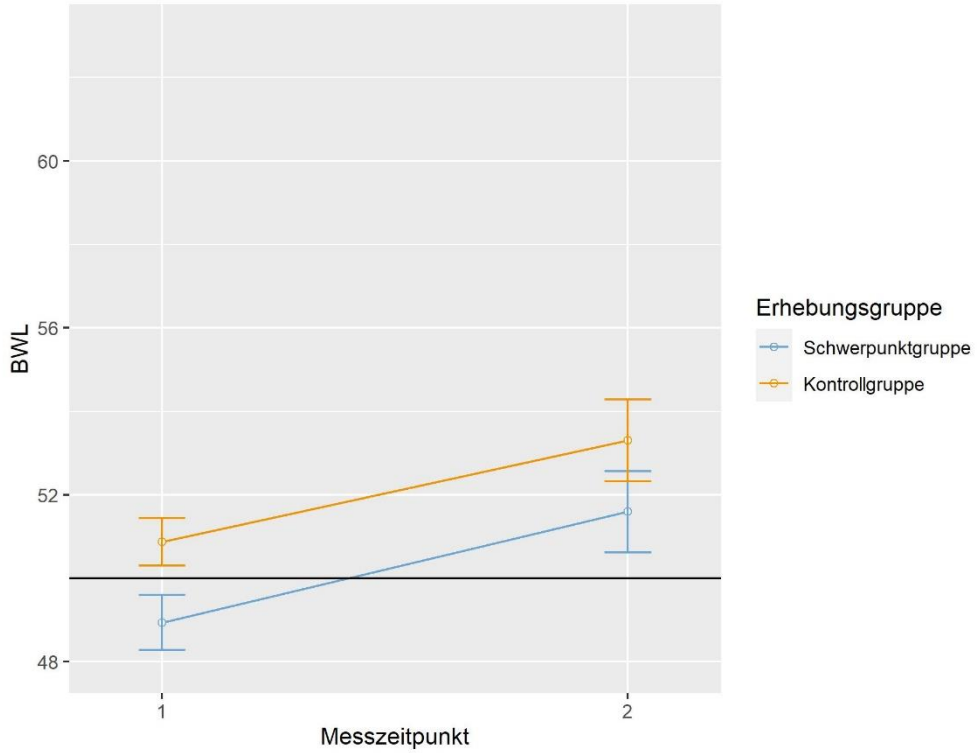


Abbildung 10 Mittlere wle-basierte BWL-Testwerte für die Schwerpunkt- und Kontrollgruppe je Messzeitpunkt unter Angabe des Standardfehlers

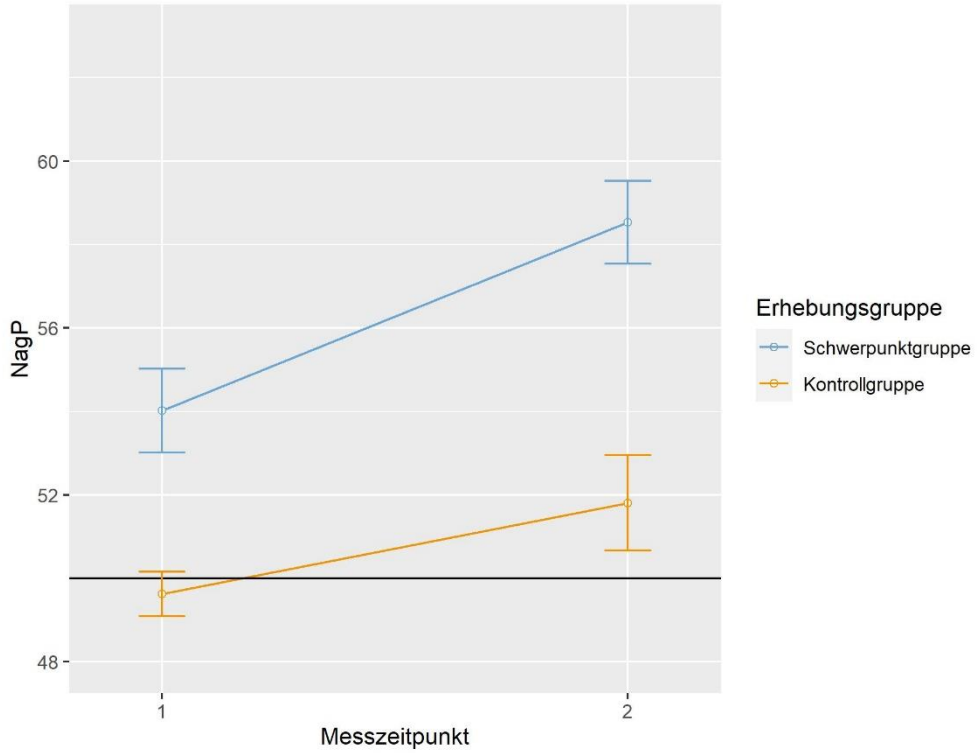


Abbildung 11 Mittlere wle-basierte NagP-Testwerte für Schwerpunkt- und Kontrollgruppe je Messzeitpunkt unter Angabe des Standardfehlers

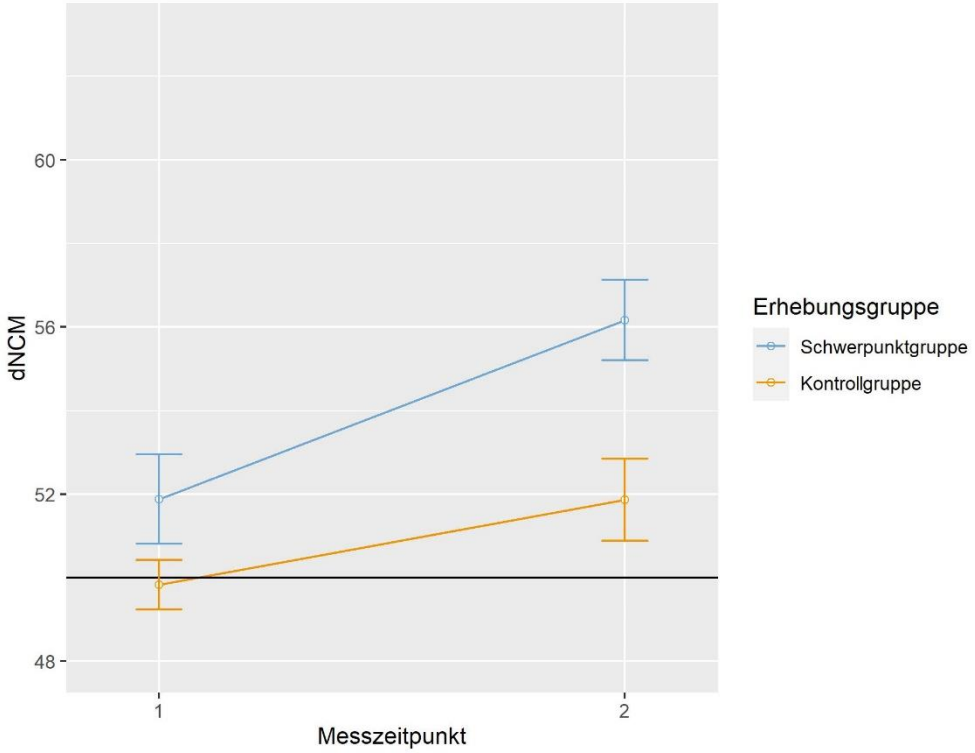


Abbildung 12 Mittlere wle-basierte dNCM-Testwerte für Schwerpunkt- und Kontrollgruppe je Messzeitpunkt unter Angabe des Standardfehlers

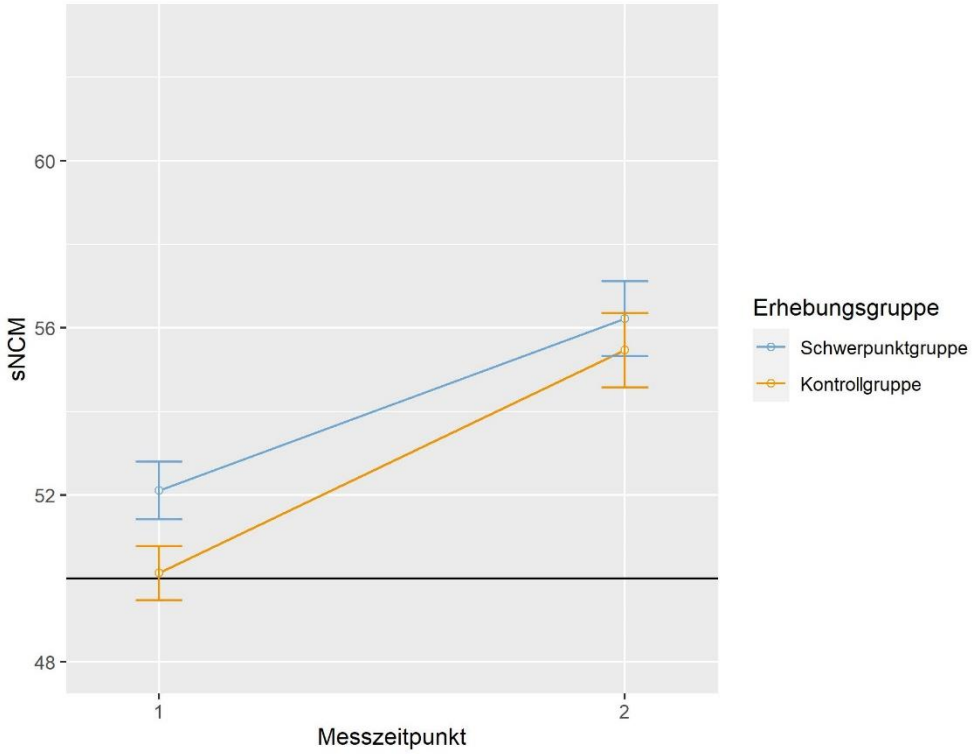


Abbildung 13 Mittlere wle-basierte sNCM-Testwerte für Schwerpunkt- und Kontrollgruppe je Messzeitpunkt unter Angabe des Standardfehlers

6 Testwerte als Indikatoren für die Kompetenz Nachhaltigkeitsmanagement

Testwertinterpretation 1 lautet, dass die Testwerte als Indikatoren für die Kompetenz Nachhaltigkeitsmanagement interpretiert werden können. Dies stellt die wichtigste Testwertinterpretation im Ko-NaMa-Projekt dar, da es hier darum geht, ob der Test überhaupt das intendierte Konstrukt erfasst. Diese Testwertinterpretation bezieht sich auf den Zusammenhang von Test und Konstrukt (siehe Abbildung 14). Die Validierung dieser Testwertinterpretationen wird schwerpunktmäßig von Projektpartnern bearbeitet und ist in dieser Arbeit gekürzt widergegeben. Sofern Evidenzen für diese Testwertinterpretationen schon veröffentlicht sind, wird auf diese Arbeiten verwiesen.

Ergebnisse in Kapitel 4 zeigen, dass Evidenzen zur internen Struktur vorwiegend für diese Art von Testwertinterpretation verwendet werden. Es werden jedoch auch die anderen von den *Standards* vorgeschlagenen Evidenzquellen genutzt.

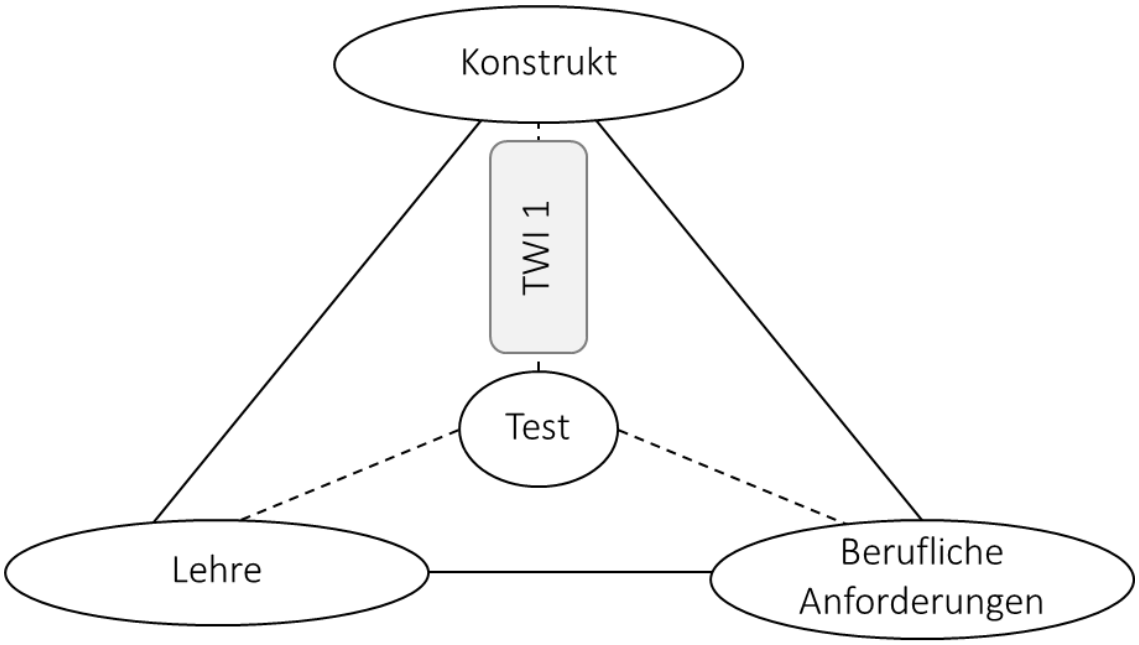


Abbildung 14 Einordnung der Testwertinterpretation 1 in das Validierungsschema

Anmerkungen. *TWI* = Testwertinterpretation.

6.1 Grundannahmen der Testwertinterpretation 1

1. **Grundannahme:** Die Testinhalte stimmen mit den theoretisch definierten Inhalten von Nachhaltigkeitsmanagement überein.

2. **Grundannahme:** Die interne Struktur der Testdaten passt zum theoretischen Modell.

Die vier kognitiven Dimensionen des Kompetenzmodells sind als jeweils eindimensional konzeptualisiert.

a) Die Daten der Leistungstests passen jeweils zu einem eindimensionalen Modell.

Die deklarativen Wissensdimensionen BWL, NagP und dNCM werden als Bedingungsfaktoren für sNCM angesehen. Die Interkorrelationen der vier Faktoren sollten daher signifikant positiv sein.

b) Die Leistungstests lassen sich in vier separate Dimensionen trennen. Die Dimensionen korrelieren positiv.

3. **Grundannahme:** Motivational-affektive Dispositionen wie Einstellungen gegenüber Nachhaltigkeit beeinflussen die Nachhaltigkeitsmanagementkompetenz.

Das Kompetenzmodell für Nachhaltigkeitsmanagement beinhaltet neben kognitiven Aspekten auch motivational-affektive Dispositionen. Dritte Grundannahme ist daher, dass neben den kognitiven Grundfähigkeiten auch motivational-affektive Dispositionen einen Einfluss darauf haben, ob ökonomische, soziale und ökologische Faktoren bei der unternehmerischen Entscheidungsfindung berücksichtigt werden (Seeber et al., 2014; Seeber et al., 2016; Michaelis, 2017).

6.2 Evidenzen für Grundannahmen der Testwertinterpretation 1

Evidenzen für Grundannahme 1 wurden von Projektkollegen gesammelt und sind noch nicht veröffentlicht. Evidenzen für die Grundannahme 2 und 3 bilden Analysen, die bei Seeber et al. (2019) und Michaelis et al. (2020) veröffentlicht sind. In beiden Publikationen war die Verfasserin dieser Arbeit als Koautorin beteiligt.

In Tabelle 8 sind die Grundannahmen von Testwertinterpretation 1 zusammengefasst und die jeweiligen Validitätsevidenzen zugeordnet.

Tabelle 8 Übersicht der Evidenzen für Grundannahmen der Testwertinterpretation 1

Grundannahmen	Evidenz
1: Die Testinhalte stimmen mit den theoretisch definierten Inhalten von Nachhaltigkeitsmanagement überein.	1: Expertenurteile zur inhaltlichen Angemessenheit
2: Die interne Struktur der Testdaten passt zum theoretischen Modell.	2a: Prüfung der Eindimensionalität jedes Tests
	2b: Prüfung eines vierdimensionalen Modells für das Gesamtmodell
	2c: Prüfen der Interkorrelationen der Tests
3: Motivational-affektive Dispositionen beeinflussen die Nachhaltigkeitsmanagementkompetenz.	3: Zusammenhang der Testwerte zu motivational-affektiven Aspekten

6.2.1 Stimmen die Testinhalte mit dem Konstrukt Nachhaltigkeitsmanagement überein?

Evidenzen für die erste Grundannahme kommen aus der Analyse der Testinhalte. Die Itementwicklung erfolgte durch Fachexperten. Die BWL – Items wurden von Fachdidaktikern aus der BWL entwickelt, die Skala Nachhaltigkeitsmanagement aus gesamtgesellschaftlicher Perspektive von Wirtschaftspädagogen mit Schwerpunkt Bildung in Nachhaltiger Entwicklung. Die Skalen Nachhaltigkeitsmanagement deklarativ und strategisch/ Begründungswissen wurden in Zusammenarbeit der Fachexperten aus BWL und Wirtschaftspädagogik entwickelt. Die inhaltliche Angemessenheit der Items wurde in der Pilotphase durch einen wissenschaftlichen Beirat beurteilt und in Zusammenarbeit mit dem wissenschaftlichen Beirat überarbeitet. Abschließend beurteilten Experten, in wie weit die Testinhalte für Nachhaltigkeitsmanagement relevant sind. Die Ergebnisse sind noch nicht veröffentlicht.

6.2.2 Passt die interne Struktur der Daten zum angenommenen Kompetenzmodell?

Für diese Grundannahme kommen Evidenzen aus Analysen zur Eindimensionalität der Skalen und zur Interkorrelation der Skalen in konfirmatorischen Faktorenanalysen.

6.2.2.1 Eindimensionalität der Skalen

Die Analysen bei Seeber et al. (2019) sprechen für die Eindimensionalität der Tests, mit Einschränkung für den BWL-Test. Die ursprünglichen deklarativen Tests mussten jedoch angepasst werden. Items mit negativer oder nicht signifikanter Faktorladung wurden ausgeschlossen und lokale Itemabhängigkeiten modelliert. Im BWL Test verblieben 53 Items, die

acht verschiedenen Unterbereiche waren mit zwei bis neun Items weiterhin alle vertreten. Im dNCM Test verbleiben 43 Items. Im Unterbereich „Management“ verblieben 13 (von 16) Items, im Unterbereich „Controlling“ verblieben 30 (von 35) Items. Im NagP Test, der ohne inhaltliche Unterbereiche konzeptioniert wurde, verblieben 43 Items. Aufgrund der geringen Bearbeitungszahlen auf Itemebene im sNCM Test wurde die eindimensionale konfirmatorische Faktorenanalyse auf Situationsebene durchgeführt. Dabei zeigte eine Situation eine nicht-signifikante Ladung auf den latenten Faktor. Die Handlung einer weiteren Situation spielt im gleichen Unternehmenskontext (Bereich Human Resources, deshalb wurden diese beiden Situationen auch nicht gemeinsam in einem Testheft präsentiert). Daher sollte die intendierte Konstruktbreite nicht durch den Ausschluss der betreffenden Situation beeinträchtigt sein. Nach Anpassung der Tests zeigte sich für jeden Test ein akzeptabler bis guter Modellfit.

6.2.2.2 Sind die Leistungstests separate Dimensionen und korrelieren positiv?

Bei Seeber et al. (2019) wurde die Passung eines vierdimensionalen Modells geprüft, in der jeder Test einen separaten Faktor bildet. Das vier-dimensionale Modell passt signifikant besser auf die Daten als ein eindimensionales Modell. Alle Dimensionen korrelieren positiv. Dies spricht für die Gesamtstruktur des angenommenen Kompetenzmodells.

Die Hypothesen bezüglich der Stärke der Interkorrelationen konnten jedoch nicht bestätigt werden. Die Tests dNCM und BWL erfassen Wissen, welches die Grundlagen für erfolgreiches wirtschaftliches Handeln darstellt. Erst dann können zusätzliche Ziele wie die Einhaltung von sozialen Standards oder Nachhaltigkeit verfolgt werden. Daher wurde erwartet, dass die Testwerte dNCM und BWL stärker korrelieren als die Testwerte dNCM und NagP. Zusätzlich sollte unter den deklarativen Wissenstests der dNCM-Test am stärksten mit sNCM korrelieren, da dieser die Wissensgrundlage für die Anwendung von Nachhaltigkeitsmanagement im Unternehmenskontext bildet. Differenztests ergaben, dass die Korrelation von dNCM und BWL nicht größer als die Korrelation von dNCM und NagP ist. Ebenso ist die Korrelation zwischen sNCM und dNCM nicht größer als die Korrelation von sNCM und den anderen beiden deklarativen Wissenstests.

Weitere Einschränkungen bezüglich der Dimensionalität des Gesamtmodells ist die Analyse auf Päckchenebene. Das vier-dimensionale Modell wurde mit je drei Itempäckchen pro Test geschätzt, da aufgrund des Testheftdesigns, hoher Itemzahlen pro Test und einer vergleichsweise kurzen Testzeit, geringe Kontingenzen auftraten.

6.2.3 Haben motivational-affektive Dispositionen einen Einfluss auf Nachhaltigkeitsmanagement?

Michaelis et al. (2020) untersuchen den Einfluss motivational-affektiver Dispositionen auf Nachhaltigkeitsmanagementkompetenz. Im Artikel wird der Einfluss von Interesse an und Einstellung zu Nachhaltigkeitsthemen, sowie der Motivation, sich mit Nachhaltigkeitsthemen zu beschäftigen, auf den sNCM-Test untersucht. Dabei zeigt sich Aversion gegenüber Nachhaltigkeit als einziger affektiv-motivationaler (negativer) Prädiktor von strategischem und Begründungswissen in Nachhaltigkeitsmanagement. Keinen Einfluss auf sNCM haben 1) Interesse an sozialen Aspekten von Nachhaltigkeit, 2) Interesse an ökologischen Aspekten von Nachhaltigkeit, 3) Motivation nachhaltig zu handeln und 4) Einstellung zu unternehmerischer Verantwortung für Nachhaltigkeit.

6.3 Plausibilität der Testwertinterpretation 1

Die Ergebnisse der konfirmatorischen Faktorenanalysen zur internen Struktur der Daten stützen weitgehend die theoretischen Annahmen und die Interpretation, dass die Testwerte Indikatoren für Kompetenz im Nachhaltigkeitsmanagement sind. Angesichts der Testanpassungen aufgrund von nicht-signifikant ladenden Items und Itemabhängigkeiten sowie der Notwendigkeit, das vier-dimensionale Modell mit Itempäckchen zu schätzen, empfehlen die Autoren jedoch, das Gesamtmodell zur Struktur der Nachhaltigkeitsmanagementkompetenz in einer größeren Stichprobe zu überprüfen. Die Interkorrelationen der Faktoren sind wie angenommen positiv, die Stärke des Zusammenhangs der einzelnen Faktoren widerspricht jedoch den theoretischen Erwartungen. Die Evidenzen zu Grundannahme 3 zeigen, dass affektiv-motivationale Dispositionen einen Einfluss auf Nachhaltigkeitsmanagement haben können. In Abbildung 15 sind die Evidenzen der Grundannahmen den in den Standards aufgeführten Quellen von Validitätsevidenzen zugeordnet.

Die Entwicklung der Ko-NaMa-Tests erfolgte durch Fachexperten in den jeweiligen Domänen und wurde durch einen wissenschaftlichen Beirat und Berufsexperten begleitet. Evidenzen aus inhaltlicher Perspektive sprechen klar für die Testwertinterpretation. Die interne Struktur, genauer die Interkorrelationen der Testkomponenten, widersprechen jedoch den theoretischen Annahmen. Wie oben berichtet, gibt es zur Struktur und Förderung der Kompetenz Nachhaltigkeitsmanagement bislang nur wenig empirische Befunde. Da die theoretischen Annahmen zur internen Struktur sich auf nur wenige empirische Befunde stützen, werden die gegenteiligen Befunde jedoch nicht als Widerspruch zur Testwertinterpretation interpretiert.

Neben der weiteren Überprüfung der internen Struktur der Kompetenz Nachhaltigkeitsmanagement sollte bei zukünftiger Verwendung des Tests auch beachtet werden, dass dieser möglicherweise differentiell auf motivational-affektive Dispositionen der Testteilnehmer*innen reagiert. (siehe Diskussion bei Seeber et al., 2019).

Die Testwertinterpretation 1, dass die Testwerte Indikatoren für die Kompetenz Nachhaltigkeitsmanagement sind, wird vorläufig als plausibel angesehen.

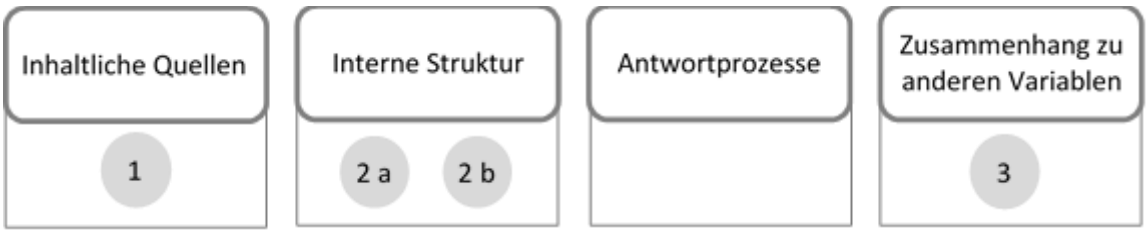


Abbildung 15 Quellen der Validitätsevidenzen für Testwertinterpretation 1

7 Testwerte als Indikatoren beruflich relevanter Kompetenz

In der zweiten Testwertinterpretation sollen die Testwerte als Indikatoren für beruflich relevante Kompetenzen interpretiert werden. Diese Testwertinterpretation fokussiert den Zusammenhang von Test zu beruflichen Anforderungen (Abbildung 16). Die stärkste Evidenzquelle für diese Art von Testwertinterpretation sind Zusammenhangsmaße von Testwerten und zukünftigem beruflichem Erfolg. Eine entsprechende Datenlage kann meist nicht erreicht werden, die wenigen Studien aus Kapitel 4, die sich mit dieser Art von Testwertinterpretation befassen, stammen aus dem Bereich des Lehramts. In den Publikationen werden dann Korrelationen zwischen Testwerten und Abschlussnoten nach praktischen Lehrphasen als Evidenzquellen genutzt. Weitere Evidenz wird aus Experteneinschätzungen zur Angemessenheit des Testinhalts bezüglich beruflicher Anforderungen und erfolgreicher Handlungsweisen geliefert. Das zeigt, dass in den untersuchten Publikationen nicht nur berufliches Handeln vorhergesagt werden soll, sondern auch die dieses Handeln hervorrufenden Bedingungsfaktoren, also die Testinhalte, interessieren. Dies geschieht vor dem Hintergrund, dass die durch den Test erfassten Kompetenzen im Studium vermittelt und entwickelt werden sollen (z.B. Lehrerbildung, Stichwort Lehrerprofessionalisierung). Die in den Standards vorgeschlagenen Evidenzquellen Maße zur internen Struktur und Antwortprozesse kommen in dieser Art von Testwertinterpretationen nicht in den untersuchten Publikationen vor.

Das Ko-NaMa-Projekt hatte die Entwicklung von Tests zum Ziel, die es erlauben Nachhaltigkeitsmanagement bei Studierenden in wirtschaftswissenschaftlichen Studiengängen zu erfassen. Diesen Studierenden wird aufgrund ihrer hohen Studierendenzahlen[32] und üblichen Berufswegen eine hohe Bedeutung bei der Integration von sozialen und ökologischen Aspekten in unternehmerischen Entscheidungen zugemessen (Seeber et al., 2016). Daher ist für Testwertinterpretation 2 auch relevant, dass Testwertinterpretation 1 als plausibel gilt (vgl. auch Diskussion zur Abgrenzung von Testwertinterpretationen in Kapitel 4.1).

[32] Im Wintersemester 2016/17 bildeten Studierende wirtschaftswissenschaftlicher Studiengänge in Deutschland die größte Fächergruppe nach Angaben des Statistischen Bundesamtes (Destatis), abgerufen am 13.06.2018 / 16:22:35 unter https://www-genesis.destatis.de/genesis/online/data;jsessionid=1EF6F0C47D57269AEF6CD0B1033DFF39.tomcat_GO_1_2?operation=abruftabelleBearbeiten&levelindex=2&levelid=1528899705034&auswahloperation=abruftabelleAuspraegungAuswaehlen&auswahlverzeichnis=ordnungsstruktur&auswahlziel=werteabruf&selectionname=21311-0003&auswahltext

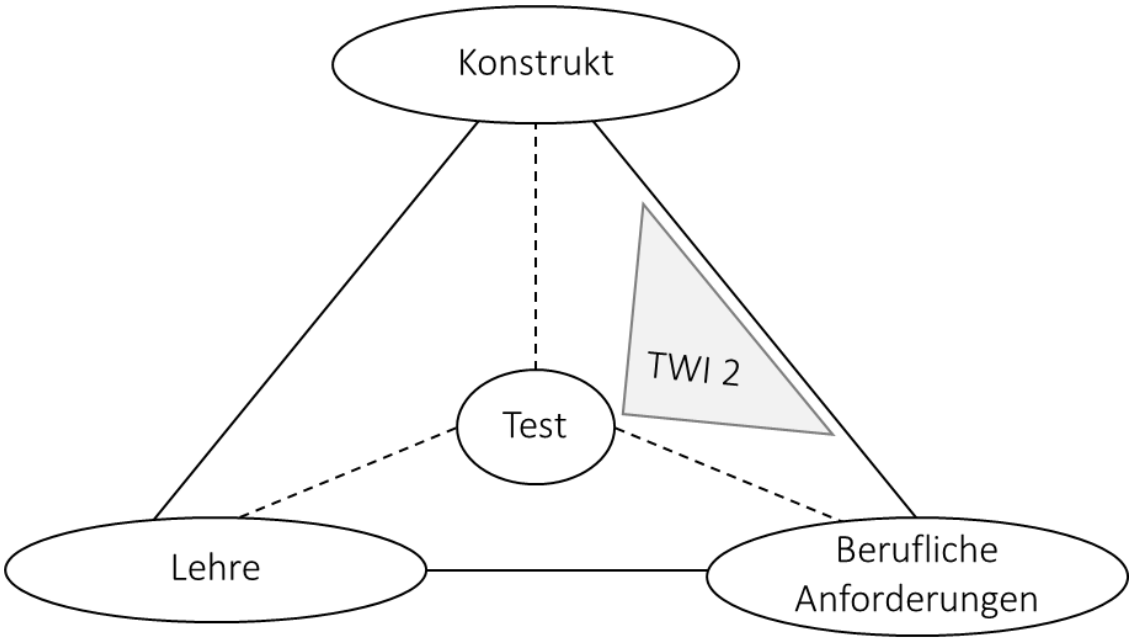


Abbildung 16 Einordnung der Testwertinterpretation 2 in das Validierungsschema

Anmerkungen. *TWI* = Testwertinterpretation.

7.1 Grundannahmen der Testwertinterpretation 2

1. **Grundannahme:** Der sNCM-Test ist geeignet, handlungsnahe Kompetenzen zu erfassen.

Um die Testwertinterpretation zu stützen, dass die Testwerte Indikatoren für beruflich relevante Kompetenzen im Bereich Nachhaltigkeitsmanagement sind, müssen die Tests zunächst in der Lage sein, handlungsnahe Kompetenzen zu erfassen.

2. **Grundannahme:** Der sNCM-Test bildet inhaltlich beruflich relevante Situationen ab, in denen Aspekte von Nachhaltigkeitsmanagement relevant sind.

Außerdem wird davon ausgegangen, dass die 13 Szenarien im sNCM-Test realistische berufliche Situationen darstellen, in denen Aspekte von Nachhaltigkeitsmanagement relevant sind.

3. **Grundannahme:** Personen mit höheren Testwerten im sNCM-Test sind erfolgreicher in beruflichen Situationen mit Anforderungen in Nachhaltigkeitsmanagement.

Wenn die Tests in der Lage sind handlungsnahe Kompetenzen zu erfassen und inhaltlich beruflich relevante Situationen widerspiegeln, sollten Personen mit höheren sNCM-Testwerten in beruflichen Situationen mit Bezug zu Nachhaltigkeitsmanagement erfolgreicher handeln.

7.2 Evidenzen für Grundannahmen der Testwertinterpretation 2

Evidenzen zu den Grundannahmen 1 und 2 wurden von Projektpartnern verantwortet und sind noch nicht veröffentlicht. Zur Grundannahme 3 wurden im Projekt keine Evidenzen gesammelt. Tabelle 9 zeigt eine Übersicht der Grundannahmen der Testwertinterpretation 2 und den jeweiligen Validitätsevidenzen.

Tabelle 9 Übersicht der Evidenzen für Grundannahmen der Testwertinterpretation 2

Grundannahmen	Evidenz
1: Der sNCM-Test ist geeignet, handlungsnahe Kompetenzen zu erfassen.	1a: Einschätzung der Studierenden 1b: Experteneinschätzung zur Realitätsnähe des Instruments
2: Der sNCM-Test erfasst Kompetenzen, die in Berufen mit Anforderungen im Nachhaltigkeitsmanagement erfolgsrelevant sind.	2a: Expertenrating zur Angemessenheit der Testinhalte 2b: Qualitative Analyse von Stellenanzeigen im Bereich Nachhaltigkeitsmanagement und den darin geforderten Kompetenzen
3: Personen mit höheren Testwerten in sNCM sind erfolgreicher in beruflichen Situationen mit Anforderungen in Nachhaltigkeitsmanagement.	-

7.2.1 Erfasst der sNCM Test handlungsnahe Kompetenzen?

Durch die Entwicklung einer simulativen Testkomponente sollten die Testteilnehmer*innen möglichst nah in einen realen Unternehmenskontext versetzt werden. Dazu wurden die Items in Kooperation mit einem Fahrradhersteller entworfen. Es wurden 13 Situationen in Form von Case Studies aufgebaut. Jede Situation spielt in einem anderen Geschäftsbereich und wird durch kurze Videosequenzen (z.B. Teammeeting mit anschließendem Arbeitsauftrag durch Vorgesetzte) eingeführt. Zur Bearbeitung müssen verschiedene Dokumente zu Rate gezogen werden. Die Antwortformate umfassen neben Single Choice Items realitätsnahe Antwortformate wie etwa das Verfassen einer E-Mail, in der Transportwege vorgeschlagen werden, unterschiedliche Zielgrößen zur Auswahl von Produktionsmaschinen priorisiert werden oder eine Entscheidungsvorlage für Vorgesetzte angefertigt werden müssen.

Zusätzlich wurden die Studierenden am Ende der Testung befragt, in wie weit sie den Test und die Testumgebung als geeignet einschätzten, handlungsnahe Kompetenzen zu erfassen. Die Befragung wurde von Projektpartnern verantwortet und ist noch nicht veröffentlicht.

7.2.2 Decken die Testinhalte berufliche Anforderungen im Nachhaltigkeitsmanagement ab?

Berufsexperten schätzten in der Pilotphase des Ko-NaMa-Projektes ein, ob die Testinhalte berufliche Anforderungen abbilden und die Umsetzung der simulativen Testkomponente berufliche Alltagssituationen widerspiegelt, in denen Nachhaltigkeitsmanagement eine Rolle spielt. Die Interviews mit Berufsexperten wurden von Projektpartnern durchgeführt und flossen in die Überarbeitung der Items vor der Haupterhebung ein. Die Ergebnisse der Expertenurteile sind jedoch nicht veröffentlicht.

Zudem untersuchten Projektpartner Stellenausschreibungen, in wie fern Kompetenzen im Nachhaltigkeitsmanagement gefordert sind. Die Ergebnisse hierzu sind ebenfalls nicht veröffentlicht.

7.2.3 Hängen Testwerte positiv mit Berufserfolg zusammen?

Für diese Grundannahme wurden im Rahmen des Projekts keine Evidenzen untersucht.

7.3 Plausibilität der Testwertinterpretation 2

Das theoriegeleitete Vorgehen in der Testentwicklung spricht für die Erfassung handlungsnaher Kompetenzen und die Expertenurteile schätzen die Testinhalte als passend zu beruflichen Anforderungen ein. Dies spricht für die Testwertinterpretation. Abbildung 17 zeigt die Zuordnung der gesammelten Evidenzen für Testwertinterpretation 2 in die Evidenzquellen der *Standards*.

Da die empirischen Evidenzen zu den Grundannahmen 1 und 2 noch nicht veröffentlicht wurden und es keine empirischen Evidenzen für Grundannahme 3 gibt, wird Testwertinterpretation 2 zum jetzigen Zeitpunkt nicht als plausibel angesehen. Sollte der Test zu einem Zweck mit stärkerer beruflicher Ausrichtung genutzt werden, z.B. als Entscheidungshilfe bei der Bewerberauswahl oder als Zertifizierung über bestimmte berufliche Kompetenzen, müsste diese Testwertinterpretation stärker untermauert werden. Evidenzen könnten über einen Vergleich von Testwerten mit Berufserfolg in Situationen mit Anforderungen im Nachhaltigkeitsmanagement kommen (Evidenzquelle „Zusammenhang zu anderen Variablen“).

Abbildung 17 Quellen der Validitätsevidenzen für Testwertinterpretation 2

8 Testwerte als Indikatoren hochschulisch vermittelter Kompetenz

Testwertinterpretation 3 lautet, dass die Testwerte als Ergebnis hochschulischer Lerngelegenheiten interpretiert werden und bezieht sich damit auf den Zusammenhang von Test zu Lehre. Wichtig ist, dass die Testwertinterpretation nicht nur davon ausgeht, dass eine Kompetenz im Hochschulkontext vermittelt wird, sondern expliziert diese Kompetenz auch. Daher kann diese Testwertinterpretation nur gestützt werden, wenn zuvor die Testwertinterpretation 1 validiert wurde und als plausibel gilt (Abbildung 18, zur Abgrenzung von Testwertinterpretationen siehe auch Diskussion in Kapitel 4.1).

Für diese Art von Testwertinterpretation lassen sich unterschiedliche Evidenzquellen nutzen. Die in Kapitel 4 untersuchten Publikationen nutzen vor allem den Bereich des Zusammenhangs zu anderen Variablen, hier insbesondere zur Anzahl oder Qualität formaler Lerngelegenheiten. Weitere Evidenzen werden aus inhaltlicher Sicht über Expertenurteile zur curricularen Passung und aus dem Bereich der Antwortprozesse über kognitive Interviews berichtet. Maße zur internen Struktur spielen eine untergeordnete Rolle bei dieser Art von Testwertinterpretation.

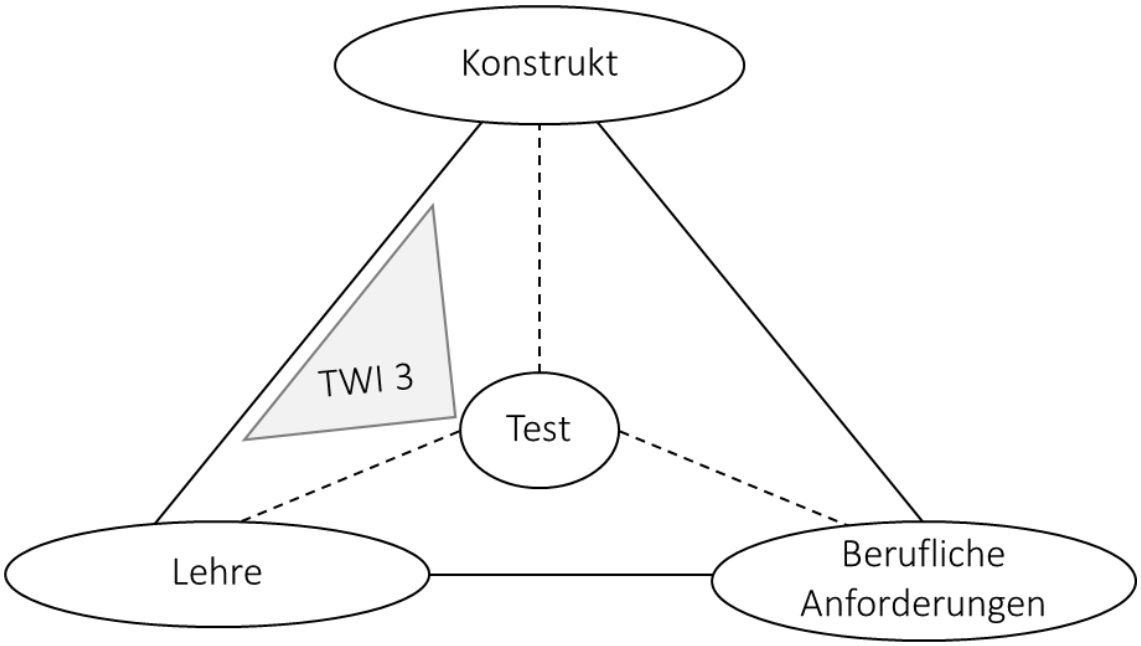


Abbildung 18 Einordnung der Testwertinterpretation 3 in das Validierungsschema

Anmerkungen. *TWI* = Testwertinterpretation.

8.1 Grundannahmen der Testwertinterpretation 3

1. **Grundannahme:** Die Testinhalte der nachhaltigkeitsbezogenen Tests stimmen mit den in den Lehrveranstaltungen der Schwerpunktgruppe behandelten Themen zu Nachhaltigkeitsmanagement überein.

Wenn der Test eingesetzt werden soll um Lernergebnisse zu erfassen, müssen die Testinhalte mit Inhalten aus der jeweiligen Lehrveranstaltung übereinstimmen (Pellegrino et al., 2001).

2. **Grundannahme:** Die Tests sind sensitiv gegenüber Veränderungen.

Im Ko-NaMa-Projekt wurde ein Messwiederholungsdesign realisiert, um Lernfortschritte zu erfassen. Um in Messwiederholungen überhaupt Lernfortschritte sehen zu können, muss der Test Veränderungen abbilden können. Zeigt sich im Längsschnitt keine Veränderung in den Testwerten, ist sonst fraglich, ob es keinen Lernzuwachs gibt oder der Test nicht veränderungssensitiv ist (Naumann et al., 2019).

3. **Grundannahme:** Kompetenzen im Nachhaltigkeitsmanagement werden vorwiegend in hochschulischen Lehrveranstaltungen vermittelt.

Bislang gab es keine Testinstrumente zur Erfassung studentischer Kompetenzen in Nachhaltigkeitsmanagement. Daher lassen sich Annahmen über Faktoren zur Förderung von Kompetenz in Nachhaltigkeitsmanagement nur bedingt aus empirischen Arbeiten ableiten. Auf Ebene der einzelnen Testkomponenten ist belegt, dass wirtschaftswissenschaftliches und betriebswirtschaftliches Wissen in fortgeschrittenem Studium höher ist als zu Beginn des Studiums (z.B. Jähnig, 2014) bzw. im Laufe des Studiums zunimmt (z.B. Schmidt, Zlatkin-Troitschanskaia & Fox, 2016). Für die Tests mit Bezug zu Nachhaltigkeit zeigen Forschungsergebnisse, dass nicht-formale Lerngelegenheiten an Hochschulen (außerhalb von Lehrveranstaltungen) Wissen in nachhaltigkeitsbezogenen Themen fördern (Hopkins, Hughes & Layer, 2008). Eine Untersuchung aus dem Bereich der beruflichen Bildung zeigt, dass Gespräche mit Freunden mit höherem Wissen über Nachhaltigkeitsthemen zusammenhängen (Michaelis, 2017). Beide Forschungsergebnisse deuten darauf hin, dass Lernfortschritte im NagP-Test auch auf Lerngelegenheiten außerhalb hochschulischer Lehrveranstaltungen zurückgeführt werden können. Insgesamt wird jedoch davon ausgegangen, dass Lernfortschritte in den nachhaltigkeitsbezogenen Tests vorwiegend durch hochschulische Lerngelegenheiten vermittelt werden.

8.2 Evidenzen für Grundannahmen der Testwertinterpretation 3

Evidenzen zur Grundannahme 1 wurde von Projektkollegen verantwortet, sind jedoch bislang nicht veröffentlicht. Der Schwerpunkt dieser Arbeit bilden Evidenzen für die dritte Grundannahme, die gleichzeitig Evidenz für Grundannahme 2 liefern. Eine Kurzfassung der Analysen für Evidenzen zur Grundannahme 3 wurde von der Autorin dieser Dissertation bei der Zeitschrift *Studies in Higher Education* eingereicht (Aichele, Hartig, Michaelis. Assessing Learning Progress in Sustainability Management Competence). Tabelle 10 fasst die Grundannahmen der Testwertinterpretation 3 zusammen und bildet die dazugehörigen Validitätsevidenzen ab.

Tabelle 10 Übersicht der Evidenzen für Grundannahmen der Testwertinterpretation 3

Grundannahmen	Evidenz
1: Testinhalte stimmen mit hochschulischen Inhalten zu Nachhaltigkeitsmanagement überein.	1a: Inhaltliche Analyse der Modulhandbücher
	1b: Einschätzung der Lehrperson zur Passung von Testinhalten zu behandelten Themen in der Veranstaltung
	1c: Einschätzung der Studierenden zur Thematisierung ausgewählter Inhalte während des Studiums
2: Tests sind veränderungssensitiv	2: siehe Evidenz 3 a und 3 b
3: Kompetenzen im Nachhaltigkeitsmanagement werden vorwiegend im Hochschulkontext vermittelt	3a-e: Strukturgleichungsmodelle prüfen, ob Gruppen mit mehr hochschulischen Lerngelegenheiten höhere Testwerte aufweisen als Gruppen mit weniger Lerngelegenheiten
	3f-i: Strukturgleichungsmodelle prüfen, ob Gruppen mit mehr außerhochschulischen Lerngelegenheiten höhere Testwerte aufweisen als Gruppen mit weniger Lerngelegenheiten

8.2.1 Stimmen die Testinhalte mit behandelten Themen in den Schwerpunktgruppen überein?

Im hochschulischen Bildungsbereich gibt es wenige Studiengänge deren Curricula über Hochschulstandorte hinweg vergleichbar sind. Im Falle von Nachhaltigkeitsmanagement kommt hinzu, dass dieses Themengebiet noch sehr wenig in betriebswirtschaftliche Studiengänge integriert ist und kein breiter Konsens über zu behandelnde Inhalte und Kompetenzen herrscht (Seeber et al., 2016). Da es kein allgemeines Curriculum für Nachhaltigkeitsmanagement gibt, können Evidenzen für Grundannahme 1 nur in Abstimmung mit den in der Datenerhebung beteiligten Lehrveranstaltungen gewonnen werden.

Evidenzen zu Grundannahme 1a kommen aus inhaltlichen Analysen von Modulhandbüchern der an der Erhebung teilnehmenden Lehrveranstaltungen. Die Ergebnisse sind noch nicht veröffentlicht. Diese Evidenzen liefern Erkenntnisse zum Zusammenhang von Test und intendiertem Curriculum.

Evidenzen zu den Grundannahmen 1b und 1c liefern darüber hinaus Erkenntnisse zum Zusammenhang von Test und dem implementierten Curriculum (Naumann et al., 2019). Während der Datenerhebung wurden die anwesenden Lehrpersonen gebeten einzuschätzen, in wie weit die Testinhalte mit den in der Veranstaltung behandelten Themen übereinstimmen. Die Ergebnisse sind noch nicht veröffentlicht. Die Studierenden wurden im Rahmen der Datenerhebung befragt, in wie fern sie bisher hochschulische Lerngelegenheiten zu Nachhaltigkeitsthemen hatten. Studierende in Lehrveranstaltungen mit Bezug zu Nachhaltigkeitsthemen (*Schwerpunktgruppe*), sollten zum zweiten Messzeitpunkt bedeutsam mehr Lerngelegenheiten berichten als noch zum ersten Messzeitpunkt. Gleichzeitig sollten Studierende ohne Lerngelegenheiten zu Nachhaltigkeitsthemen (*Kontrollgruppe*) keine Veränderung vom ersten zum zweiten Messzeitpunkt berichten.

Die Fragebogenskalen wurden in Kapitel 5.3.1 beschrieben. Wie dort gezeigt, bestätigen sich beide Annahmen. Studierende aus der Schwerpunktgruppe berichten zum zweiten Messzeitpunkt signifikant mehr bisherige hochschulische Lerngelegenheiten als zum ersten Messzeitpunkt, $t(38) = 9.33$, $p < .001$. Studierende der Kontrollgruppe berichten zum zweiten Messzeitpunkt nicht mehr hochschulische Lerngelegenheiten als zum ersten Messzeitpunkt, $t(71) = 0.82$, $p = .42$.

8.2.2 Sind die Tests veränderungssensitiv?

Evidenzen zur Veränderungssensitivität werden typischerweise durch Modelle zur Analyse von Lernzuwächsen gewonnen. Testpersonen werden vor und nach einer systematischen Lerngelegenheit getestet, von der man ausgeht, dass sie zu Lernerfolg führt. Daher müsste zur Untersuchung von Veränderungssensitivität eines Tests schon in der Pilotierung eines Tests eine Messwiederholung durchgeführt werden. Da die Pilotierungsstichprobe im Ko-NaMa-Projekt keine Analyse der Veränderungssensitivität zuließ, kann für diese Grundannahme keine direkte Evidenz geliefert werden. Indirekt lässt sich diese Grundannahme prüfen, in dem die theoretisch angenommenen Lernfortschritte analysiert werden (siehe Evidenz zu Grundannahme 3). Bleiben

die angenommenen Lernfortschritte aus, ist allerdings unklar, ob der Test nicht veränderungssensitiv ist oder kein Lernfortschritt stattgefunden hat (Naumann et al., 2019).

8.2.3 Werden Kompetenzen im Nachhaltigkeitsmanagement vorwiegend hochschulisch vermittelt?

Um Grundannahme 3 mit Evidenzen zu stützen bzw. zu widerlegen, werden Lernfortschritte der Studierenden auf hochschulische und außerhochschulische Lerngelegenheiten zurückgeführt. Die Evidenzen zu Grundannahme 3 bauen inhaltlich und in der gewählten Analyse aufeinander auf. Daher werden im folgenden Kapitel zunächst die jeweiligen Evidenzen beschrieben. Die verwendeten Methoden, Ergebnisse und die Diskussion werden anschließend gemeinsam dargestellt.

Hochschulische Lerngelegenheiten haben einen Erklärungswert für Testwerte in den nachhaltigkeitsbezogenen Tests.

Für die Datenerhebung im Ko-NaMa-Projekt wurden Studierende aus Lehrveranstaltungen rekrutiert, die Themen und Methoden des Nachhaltigkeitsmanagements und –controllings behandeln. Diese bilden die „Schwerpunktgruppe“. Demgegenüber stehen Studierende aus betriebs- und wirtschaftswissenschaftlichen Lehrveranstaltungen ohne Bezug zu Nachhaltigkeitsmanagement, welche die „Kontrollgruppe“ bilden. Die Zuweisung in die Schwerpunkt- bzw. Kontrollgruppe einer Lehrveranstaltung erfolgte durch das Ko-NaMa-Projektteam. Dazu wurden die verwendeten oder geplanten Lehrmaterialien gesichtet und Rücksprache mit den verantwortlichen Lehrpersonen gehalten. Unter der Bedingung, dass Kompetenzen erlernbar sind (Hartig & Klieme, 2006), wird angenommen, dass Studierende aus der Schwerpunktgruppe größeren Lernfortschritt in den Tests NagP, dNCM und sNCM zeigen, als Studierende der Kontrollgruppe.

> 3a: Unter Kontrolle des Ausgangniveaus haben Studierende der Schwerpunktgruppe zum zweiten Messzeitpunkt höhere Testwerte für NagP, dNCM und sNCM als Studierende der Kontrollgruppe.

Der beschriebene Zusammenhang von Lerngelegenheiten und Gruppenzugehörigkeit bezieht sich nur auf die Tests NagP, dNCM und sNCM. Sowohl Schwerpunkt- als auch Kontrollgruppe wurden aus betriebs- und wirtschaftswissenschaftlichen Studiengängen rekrutiert. Zudem erfasst der BWL-Test Wissen, welches in grundständigen betriebswirtschaftlichen Veranstaltungen vermittelt wird (Seeber et al., 2019). Daher werden keine systematischen

Unterschiede in den Lerngelegenheiten zu betriebswirtschaftlichem Wissen erwartet und es sollte keine differentiellen Lernfortschritte zwischen den Gruppen auftreten.

> 3b: Zum ersten Messzeitpunkt gibt es keine Unterschiede in den BWL-Testwerten zwischen Schwerpunkt- und Kontrollgruppe.
>
> 3c: Unter Kontrolle des Ausgangsniveaus gibt es zum zweiten Messzeitpunkt keine Unterschiede in den BWL-Testwerten zwischen Schwerpunkt- und Kontrollgruppe.

Die Einteilung in Schwerpunkt- bzw. Kontrollgruppe basiert auf vorab erstellten Lehrmaterialien und vor der ersten Messung durchgeführten Gesprächen mit der verantwortlichen Lehrperson und bildet somit eine Einteilung auf Basis des intendierten Curriculums ab (Naumann et al., 2019). Da das intendierte Curriculum abweichen kann von dem, was tatsächlich in der Lehre umgesetzt wird (implementiertes Curriculum; Naumann et al., 2019), wurden die Studierenden gebeten, ihre bisherigen hochschulischen Lerngelegenheiten zu Themen aus Nachhaltigkeit und Nachhaltigkeitsmanagement einzuschätzen. Während die Einteilung von Schwerpunkt- und Kontrollgruppe lediglich Informationen zu Lerngelegenheiten in einer, für die Testung relevante Lehrveranstaltung bietet, liefern die Selbsteinschätzungen Informationen zu Lerngelegenheiten aus anderen Lehrveranstaltungen und auch aus vorherigen Semestern.

> 3d: Studierende die zum ersten Messzeitpunkt mehr bisherige hochschulische Lerngelegenheiten berichten, haben höhere Testwerte in NagP, dNCM und sNCM.
>
> 3e: Unter Kontrolle des Ausgangsniveaus haben Studierende, die zum zweiten Messzeitpunkt mehr bisherige hochschulische Lerngelegenheiten berichten, höhere Testwerte in NagP, dNCM und sNCM.

Außerhochschulische Lerngelegenheiten haben nur für den nachhaltigkeitsbezogenen Test aus gesamtgesellschaftlicher Perspektive einen Erklärungswert.

Weitere Evidenz für oder gegen unsere Testwertinterpretation liefern Analysen, die Lernfortschritte auf außerhochschulische Lerngelegenheiten zurückführen. Forschungsergebnisse aus dem Bereich der beruflichen Bildung zeigen, dass Gespräche mit Freunden mit höherem Wissen über Nachhaltigkeitsthemen aus gesamtgesellschaftlicher Perspektive zusammenhängen (Michaelis, 2017). Außerhochschulische Lerngelegenheiten zu Nachhaltigkeitsthemen können durch persönliches Engagement in entsprechenden NGOs oder

durch das Verfolgen von Themen in den Medien gegeben sein. In den Medien breit aufgegriffene Themen, die zu den Testinhalten passen, sind exemplarisch die Diskussionen über Bedingungen in der Textilproduktion nach dem Einsturz einer Fabrik in Bangladesch 2013, der Abgasskandal in Deutschland (ein in den Medien seit 2015 aufgegriffenes Thema) oder die Überschreitung von Luftreinhaltegrenzen und Diskussion um Fahrverbote in Innenstädten. Aktuellere Beispiele sind Volksbegehren für mehr Artenschutz in Bayern (2019 von der Landesregierung umgesetzt) und für ein Gesetz, das Unternehmen zur Überprüfung sozialer Standards auch bei Lieferanten verpflichtet (welches 2021 als „Lieferkettengesetz“ in Deutschland verabschiedet wurde, wenn auch unter großer Kritik einiger NGOs). Daher sind Zusammenhänge von außerhochschulischen Lerngelegenheiten zum Test NagP denkbar.

> 3f: Studierende die zum ersten Messzeitpunkt mehr bisherige außerhochschulische Lerngelegenheiten berichten, haben einen höheren Testwert im NagP-Test.
>
> 3g: Unter Kontrolle des Ausgangsniveaus haben Studierende, die zum zweiten Messzeitpunkt mehr bisherige außerhochschulische Lerngelegenheiten berichten, einen höheren Testwert in NagP-Test.

Die Argumentation für Evidenz 3f und 3g wird nicht für die Tests dNCM und sNCM fortgeführt. Spezifische Controlling- und Managementmethoden zu Nachhaltigkeit werden außerhalb hochschulischer Lehrveranstaltungen vermutlich weniger häufig thematisiert. Am ehesten kommt man mit Nachhaltigkeitsmanagement durch Werbekampagnen in Berührung, oder wenn es um externe Kosten der Unternehmen geht (also solche Kosten die zwar anfallen, die aber nicht das Unternehmen selbst zu bezahlen hat), die staatlich reguliert werden. Beispielhaft seien hier CO2-Zertifikate oder die EEG-Umlage genannt. Spezifische Controlling- und Managementpraktiken in Unternehmen sowie die Anwendung dieser Methoden werden jedoch selten thematisiert. Daher sollten auftretende Lernfortschritte in den Tests dNCM und sNCM nicht auf außerhochschulische Lerngelegenheiten zurückzuführen sein.

> 3h: Zum ersten Messzeitpunkt sagen die selbstberichteten außerhochschulischen Lerngelegenheiten nicht die Testwerte in dNCM und sNCM vorher.
>
> 3i: Unter Kontrolle des Ausgangsniveaus sagen die selbstberichteten außerhochschulischen Lerngelegenheiten nicht die Testwerte in dNCM und sNCM vorher.

8.2.3.1 Methode

Die Analysen wurden in R (Version 4.0.2; R Core Team, 2020) mit Hilfe des lavaan Pakets (Rosseel, 2012) durchgeführt. Sofern nicht anders angegeben, werden die beschriebenen Modelle für alle drei Leistungstests identisch spezifiziert. Deskriptive Statistiken wurden ebenfalls in R mit Hilfe des stats Paketes berechnet (R Core Team, 2020).

Stichprobe

Für die Analysen wurde ein Teildatensatz des Ko-NaMa-Projekts verwendet (Seeber et al., 2020). Die folgenden Analysen werden nur mit Studierenden durchgeführt, für die im Studiendesign eine Messwiederholung vorgesehen war. Der Teildatensatz umfasst 499 Studierende aus acht Lehrveranstaltungen unterschiedlicher Hochschulen. Die Testpersonen waren im Median 23 Jahre alt und befanden sich etwa je zur Hälfte im Bachelorstudium und im Masterstudium. Tabelle 11 zeigt neben den Angaben für die gesamten verwendeten Daten auch die deskriptiven Statistiken separat für die Schwerpunkt- und Kontrollgruppe.

Tabelle 11 Stichprobencharakteristika für die Gesamtstichprobe und separat für Schwerpunkt- und Kontrollgruppe

Gesamt SG KG	***N* = 499** *n* = 188 *n* = 311	*N* weiblich (%)	Median (*SD*)	*Min*	*Max*	*N* in Bachelor-studiengang (%)	Fehlende Werte (%)
	Geschlecht	**213 (43%)**					**108 (22%)**
	SG	84 (45%)					45 (24%)
	KG	129 (41%)					63 (20%)
	Alter		**23 (2.68)**	**19**	**35**		**119 (24%)**
	SG		23 (2.41)	19	35		46 (24%)
	KG		23 (2.83)	19	32		73 (23%)
	Studienfortschritt					**274 (55%)**	**110 (22%)**
	SG					91 (48%)	47 (25%)
	KG					183 (58%)	63 (20%)

Anmerkungen. **FETT**: Angaben für den Teil des Ko-NaMa-Datensatzes (Seeber et al., 2020), für den eine Messwiederholung vorgesehen war. SG = Schwerpunktgruppe. KG = Kontrollgruppe.

Die Studierenden wurden in die Schwerpunktgruppe eingeteilt, wenn die Lehrveranstaltung, in der sie getestet wurden, geplante Lerngelegenheiten in Bezug auf Nachhaltigkeitsthemen enthielt. Studierende aus Lehrveranstaltungen ohne solch einen Bezug zu Nachhaltigkeitsthemen bilden die Kontrollgruppe. Die Zuordnung erfolgte durch die Studienautoren nach Rücksprache mit den verantwortlichen Lehrpersonen sowie nach Durchsicht der verwendeten Lehrmaterialien (vgl. Kapitel 5.3.1).

Instrumente

Die Instrumentenentwicklung wurde von Projektkolleg*innen verantwortet. Die Verfasserin der Arbeit unterstützte die Instrumentenentwicklung im Ko-NaMa-Projekt durch die Aufbereitung der Datensätze und die Bereitstellung psychometrischer Kennwerte zu den Items und Skalen in der Pilotierungsphase.

Leistungstests. Die Leistungstests sind detailliert in Kapitel 5.3 beschrieben. Daher werden hier nur kurz die thematischen Schwerpunkte der einzelnen Tests genannt. In Anhang F sind Beispielitems für jeden der Tests abgebildet.

Aus vorherigen Analysen war bekannt, dass die Tests nur bedingt eindimensional sind (vgl. Ergebnisse der IRT-Skalierung in Kapitel 5.4). Für die folgenden Analysen werden daher nicht alle in der Haupterhebung eingesetzten Items verwendet, sondern nur die zu einem eindimensionalen Modell passenden Items. Dafür wurden pro Test eindimensionale Modelle spezifiziert, in denen die Faktorvarianz auf 1 und der Faktormittelwert auf Null fixiert wurden. Alle nicht-signifikant auf den Faktor ladenden Items wurden ausgeschlossen. Für den sNCM-Test erfolgten die Analysen auf Ebene der Situationen, da hier nicht beliebig einzelne Items ausgeschlossen werden konnten. Das Vorgehen ist detailliert bei Seeber et al. (2019) beschrieben. Im sNCM-Test wurden lediglich die Items einer Situation ausgeschlossen, die in einem Unternehmensbereich spielt der durch eine weitere Situation abgedeckt ist. In den Tests BWL, NagP und dNCM verblieb der Großteil der Items im Test. Insgesamt wurde daher keine Einschränkung in der Konstruktbedeutung erwartet.

BWL-Test

Es werden 53 der 80 Items aus der Haupterhebung verwendet. Die BWL-Items decken die Unterbereiche 1) Absatz und Marketing, 2) Beschaffung und Logistik, 3) Finanzwirtschaft, 4) Personalwirtschaft, 5) Planung, 6) Produktion, 7) Rechnungswesen, 8) Unternehmen und Management. Diese Unterbereiche wurden von Inhaltsexperten nach Durchsicht gängiger

Standard-BWL-Lehrbücher an unterschiedlichen Hochschulstandorten identifiziert. Die Items bestehen aus einem Itemstamm gefolgt von vier Antwortoptionen im Single-Choice Format.

NagP-Test

Es werden 43 Items der 53 Items aus der Haupterhebung verwendet. Der Test erfasst zum einen Wissen, das durch Verfolgen gesellschaftlicher und politischer Debatten erworben werden kann. Dazu zählen Prinzipien von Nachhaltigkeit und Ziele internationaler Vereinbarungen zu nachhaltiger Entwicklung, Vor- und Nachteile verschiedener Energieträger und globale Wirtschaftszusammenhänge. Zum anderen wird auch Faktenwissen erfasst, welches nur durch nähere Beschäftigung mit dem jeweiligen Thema erworben werden kann. Dazu zählen die Definitionen von Fachbegriffen, Wissen über den CO^2-Abdruck verschiedener Heizsysteme oder der Beifang in Kilogramm in der Fischerei. Die Items bestehen aus einem Itemstamm gefolgt von vier Antwortoptionen im Single-Choice Format.

dNCM-Test

Es werden 43 der 51 Items aus der Haupterhebung verwendet. Inhaltlich erfasst der Test zum einen Fachwissen zu und Anwendung von Controllingmethoden unter Berücksichtigung von Nachhaltigkeitskriterien. Zum anderen erfasst der Test Fachwissen zu Managementmethoden, in denen Nachhaltigkeitsaspekte berücksichtigt werden. Die Items bestehen aus einem Itemstamm gefolgt von vier Antwortoptionen im Single-Choice Format.

sNCM-Test

Es werden die Items von 12 der 13 Situationen aus der Haupterhebung eingesetzt. Der Test bildet die gleiche inhaltliche Komponente wie der dNCM-Test ab. Er unterscheidet sich von diesem in der Repräsentationsform des Wissens. Im sNCM-Test müssen Unternehmenssituationen aus ökonomischer, ökologischer und sozialer Perspektive bewertet werden und eine Entscheidung getroffen werden, welche die jeweiligen Kriterien angemessen berücksichtigt.

Fragebogenskalen. Die Studierenden schätzten auf vierstufigen Likert-Skalen von 1 (*Trifft gar nicht zu*) bis 4 (*Trifft voll zu*) ihre bisherigen Lerngelegenheiten zu Nachhaltigkeitsthemen ein. Sowohl für die hochschulischen als auch die außerhochschulischen Lerngelegenheiten wurden jeweils drei Skalen verwendet. Diese Skalen sind ausführlich in Kapitel 5.3 beschrieben.

Die Skalen werden formativ betrachtet. Die verwendeten Items beziehen sich auf Kernaspekte des jeweiligen Themengebietes und sind nicht beliebig austauschbar, ohne dass sich die Bedeutung des Konstrukts verändert. Daher wurden die Antworten der Studierenden zu einem manifesten Mittelwert pro Skala zusammengefasst.

Zum ersten Messzeitpunkt korrelieren die drei Skalen für hochschulische Lerngelegenheiten zwischen r = .61 - .75 und zum zweiten Messzeitpunkt zwischen r = .64 - .76 (siehe Anhang H). Die drei Skalen für außerhochschulische Lerngelegenheiten korrelierten zum ersten Messzeitpunkt zwischen r = .52 - .62 und zum zweiten Messzeitpunkt zwischen r = .53 - .65 (siehe Anhang H). Um Multikollinearität bei der Regression des Lernzuwachses auf Lerngelegenheiten zu vermeiden, wurden die jeweiligen Skalen zusammengefasst. Der Mittelwert aller selbstberichteten hochschulischen Lerngelegenheiten wird als eine manifeste Prädiktorvariable pro Messzeitpunkt im Strukturgleichungsmodell verwendet. Gleiches gilt für die außerhochschulischen Lerngelegenheiten.

Umgang mit fehlenden Werten. In den Leistungstests wurden nicht erreichte Items als fehlende Werte, übersprungene Items als falsch codiert. Die Daten wurden anschließend einer Bereinigung unterzogen. Antworten von Personen, welche nicht mindestens fünf Items in den deklarativen Tests bearbeiteten, wurden in dem betreffenden Test ausgeschlossen. Für den sNCM-Test wurden die Daten situationsspezifisch betrachtet. Hier wurden Antworten von Personen ausgeschlossen, wenn diese nicht mindestens die Hälfte der Items innerhalb einer Situation bearbeitetet hatten.

Messmodell und Strukturgleichungsmodelle

Um Evidenzen für die dritte Grundannahme zu erhalten, wurde ein kovarianz-analytischer Ansatz gewählt (z.B. Köhler, Hartig & Schmid, 2020), um Lernfortschritt in den Tests zu modellieren[33]. Dieser soll anschließend durch unterschiedliche Lerngelegenheiten erklärt werden.

[33] Für die Fragestellung kann auch ein Latent-Change Modell gewählt werden, in dem der Lernfortschritt als Differenz von zweitem und erstem Messzeitpunkt modelliert wird. Latent-Change und kovarianz-analytische Modelle können zur Beantwortung der Fragestellung nach Lernfortschritt genutzt werden, kommen aber in manchen Fällen zu unterschiedlichen Ergebnissen (auch bekannt als Lord‘s Paradox; Lord, 1967). Für die geplanten Analysen wurde das kovarianz-analytische Modell als vorteilhaft bewertet, da 1) die Varianz zeitpunktspezifischer Komponenten in den Testwerten des ersten Messzeitpunktes als gering eingeschätzt wird und gleichzeitig 2) ein Zusammenhang zwischen Ausgangsniveau und der Anzahl an

In einem ersten Schritt wurde ein Messwiederholungsmodell spezifiziert, in dem jeder Faktor einen Messzeitpunkt darstellt und die beiden Faktoren kovariieren. Beide Messzeitpunkte wurden eindimensional spezifiziert und die Faktorladung des ersten Indikators jeweils auf 1 gesetzt. Indikatorspezifische Residualkorrelationen über Messzeitpunkte wurden zunächst zugelassen. Um möglichst sparsame Modelle zu schätzen, wurden nicht signifikanten Residualkorrelationen wieder restringiert. Der Mittelwert des ersten Messzeitpunkts ist auf Null fixiert und der Mittelwert des zweiten Messzeitpunkts wird frei geschätzt. Der Mittelwert des zweiten Messzeitpunkts wird, unter Kontrolle des Ausgangswertes, als Lernfortschritt in einem Leistungstest interpretiert. Die Regression des zweiten auf den ersten Messzeitpunkt führte in einigen Modellen zu einer negativ geschätzten latenten Varianz für den zweiten Messzeitpunkt. Dann wurde ein 95% bootstrap-Konfidenzintervall mit 1000 Replikationen für die Varianz geschätzt. Da dieses in allen Fällen die Null einschloss (siehe Anhang J), wurden die Varianzen auf einen positiven Wertebereich fixiert (Kolenikov & Bollen, 2012). Die betreffenden Modelle werden im Ergebnisteil gekennzeichnet.

Um der Clusterung der Daten Rechnung zu tragen, wurden die Lehrveranstaltungen als feste Effekte modelliert. In einem Feste Effekte Modell wird für jede Lehrveranstaltung ein Effekt in Relation zu einer Referenzgruppe geschätzt. Das Regressionsgewicht der Referenzgruppe (in den folgenden Analysen Lehrveranstaltung *Göttingen C*, eine der Lehrveranstaltungen aus der Kontrollgruppe) wird auf Null fixiert. Die Regressionsgewichte der anderen Lehrveranstaltungen können als Abweichung im Interzept zur Referenzgruppe interpretiert werden.

Um die durchschnittlichen Unterschiede zwischen Lehrveranstaltungen der Schwerpunktgruppe und denen der Kontrollgruppe abzubilden, wurden lineare Kontraste gebildet (McNeish & Stapleton, 2016). Dazu wurden die Regressionsgewichte der Standorte innerhalb der Schwerpunktgruppe (bzw. Kontrollgruppe) mit ihrer prozentualen Größe innerhalb der Schwerpunktgruppe (bzw. Kontrollgruppe) multipliziert und dann addiert. Anschließend wurde das durchschnittliche Regressionsgewicht der Kontrollgruppe vom durchschnittlichen Regressionsgewicht der Schwerpunktgruppe subtrahiert. Positive Werte im linearen Kontrast bedeuten daher höhere Werte für die Standorte der Schwerpunktgruppe. Das Analysemodell

Lerngelegenheiten und 3) ein Zusammenhang des Ausgangsniveaus und des Lernfortschritts nicht ausgeschlossen werden kann, vgl. Köhler, Hartig und Schmid (2020).

für Evidenz 3a bis 3c wird im folgenden Modell 1 genannt (siehe grau hinterlegter Bereich in Abbildung 19).

Modell 2 liefert Ergebnisse zu den Evidenzen 3d und 3e. Hier wird die selbsteingeschätzte Anzahl bisheriger *hochschulischer Lerngelegenheiten* als Prädiktorvariable von Lernfortschritt genutzt (siehe Abbildung 19). Die Prädiktorvariable zu Messzeitpunkt 1 sagt den Ausgangswert des Testwerts vorher, die Prädiktorvariable zu Messzeitpunkt 2 sagt den Lernfortschritt vorher. Da die selbsteingeschätzten hochschulischen Lerngelegenheiten von der Lehrveranstaltung abhängen sollten, in der die Studierenden befragt werden, wird die Prädiktorvariable zusätzlich durch die Standortvariablen vorhergesagt.

In Modell 3 wird der Lernfortschritt durch bisherige *außerhochschulische Lerngelegenheiten* vorhergesagt. Da die außerhochschulischen Lerngelegenheiten nicht von der besuchten Lehrveranstaltung abhängen sollten, wird diese Prädiktorvariable nicht durch die festen Effekte der Standorte vorhergesagt. Modell 3 liefert Ergebnisse zu den Evidenzen 3f bis 3i.

Zu berücksichtigen ist, dass in den beschriebenen Modellen 2 und 3 die Effekte von selbstberichteten Lerngelegenheiten nur unter Kontrolle der festen Effekte analysiert werden. Dies ist dem Umgang mit der hierarchischen Datenstruktur durch feste Effekte geschuldet. Im Ergebnisteil werden für die Modelle 2 und 3 daher *Alternativmodelle* präsentiert, in denen der Clusterung der Daten durch eine post-hoc Korrektur der Standardfehler begegnet wird. Allerdings zeigen McNeish und Stapleton (2016), dass die Nicht-Berücksichtigung der komplexen Datenstruktur in der Analyse mit späterer Korrektur der Standardfehler bei nur wenigen Clustern (simuliert mit vier und acht Clustern) zur Unterschätzung von Standardfehlern führen kann. Gleichzeitig konnte ein Modell mit festen Effekten bei acht Clustern angemessen breite Konfidenzintervalle reproduzieren. Des Weiteren sollte die Standardfehlerkorrektur nur vorgenommen werden, wenn die Anzahl der vorhandenen Cluster kleiner der zu schätzenden Parameter ist, da sonst die Varianz-Kovarianzmatrix nicht positiv definit ist (Rosseel, 2012). In den beschriebenen Alternativmodellen 2 und 3 werden jeweils mehr Parameter geschätzt, als Cluster vorhanden sind. Daher sollten die Ergebnisse der *Alternativmodelle* vorsichtig interpretiert werden, da die Standardfehler möglicherweise zu klein sind und dementsprechend häufiger Fehler 1. Art auftreten können.

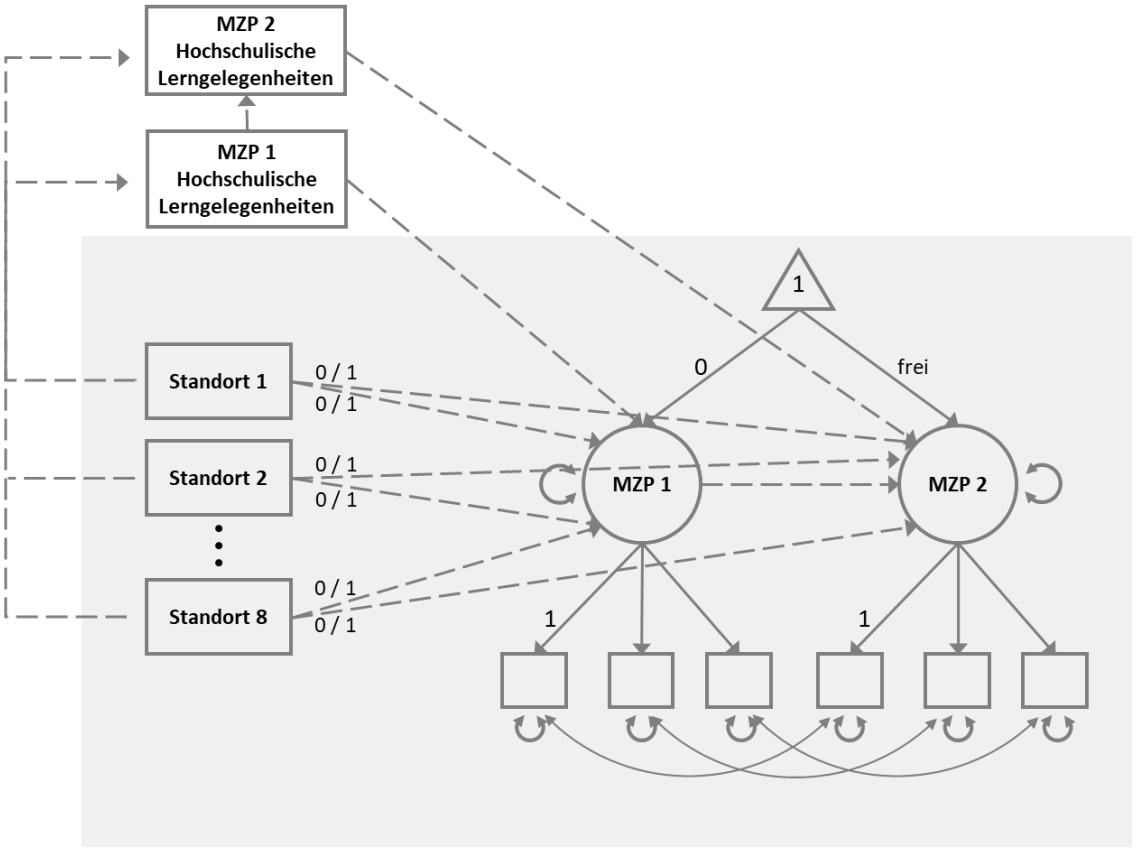


Abbildung 19 Analysemodell 2 mit selbstberichteten hochschulischen Lerngelegenheiten als Prädiktoren von Lernfortschritt

Anmerkungen. Modell 1 ist grau hinterlegt. Die durchgehenden Pfeile stellen das Messmodell dar, die Residualkorrelationen werden nur für signifikante Korrelationen modelliert. Gestrichelte Pfeile stellen das Strukturgleichungsmodell dar.

Untersuchung auf Päckchen-Ebene

Die folgenden Analysen werden nicht auf Itemebene, sondern auf Päckchenebene durchgeführt. Die vier Leistungstests im Ko-NaMa-Projekt bestehen aus 51 bis 80 Items bzw. 43 bis 79 Items nach Ausschluss von Items um die Tests eindimensional zu modellieren. Die Bearbeitungszeit der deklarativen Testteile war auf jeweils zehn Minuten beschränkt, weshalb die Studierenden im Median nur 23 bis 26 Items in den Tests BWL, NagP und dNCM bearbeiteten. Im sNCM-Test bearbeiteten die Studierenden im Median 21 Items. Bedingt durch das Testheftdesign waren nur zwischen 17 und 27 Items zu bearbeiten. Insgesamt ist die Kovarianzabdeckung daher gering und die Kontingenztabellen enthalten vereinzelt leere Zellen. Dies betrifft insbesondere mehrstufige Items mit hoher Schwierigkeit im sNCM-Test. Um Probleme bei der Modellschätzung zu vermeiden, wurden für die Analysen drei Päckchen pro Test gebildet. Jedes

Päckchen enthält die durchschnittlich erreichte Punktzahl in den jeweiligen Items für die deklarativen Tests und die durchschnittlich erreichte Punktzahl für die jeweiligen Situationen im sNCM-Test. Die Päckchen wurden unter Berücksichtigung der Faktorladungen (Little, Rhemtulla, Gibson & Schoemann, 2013) sowie der Itemschwierigkeiten erstellt (siehe Tabelle 12).

Bei den deklarativen Tests mit inhaltlichen Unterbereichen wurde darauf geachtet, dass jedes Päckchen alle inhaltlichen Bereiche der Skala abdeckt. Im sNCM-Test wurde darauf geachtet, dass in einem Päckchen nicht zwei Situationen aus einem Arbeitsbereich stammen und dass die Situationen unterschiedliche Unternehmensprozesse[34] repräsentieren. Ziel dieses Vorgehen war es, möglichst parallele und domänenrepräsentative Indikatoren für das jeweilige Konstrukt zu bilden (vgl. Little et al., 2013).

Da die Päckchenbildung zu kontinuierlichen Indikatoren führt, werden Maximum-Likelihood Schätzer verwendet. Mit fehlenden Werten wird mit dem im lavaan-Package implementierten zweischrittigen Verfahren umgegangen (Rosseel, 2012).

Tabelle 12 Zusammenstellung der Päckchen für die Leistungstests

Test	Päckchen	*N* Items	*M* Faktorladung	*M* Itemschwierigkeit
BWL	1	18	.117	0.304
	2	17	.129	0.306
	3	18	.133	0.257
NagP	1	15	.127	0.025
	2	14	.114	0.007
	3	14	.126	-0.004
dNCM	1	15	.118	0.481
	2	14	.117	0.489
	3	14	.116	0.432
sNCM	1	22 (Situation 1, 8, 9 und 13)	.165	0.480
	2	26 (Situation 3, 5, 6 und 10)	.152	0.862
	3	21 (Situation 2, 4, 7 und 11)	.162	0.486

Anmerkungen. M Faktorladung = mittlere Faktorladung der Items eines Päckchens. Für die Berechnung wurden die Daten des ersten Messzeitpunkts genutzt. Für jeden Test wurden eindimensionale Faktorenmodelle mit allen Items spezifiziert. Im Modell sind die Faktorvarianz auf 1 und der Faktormittelwert auf Null fixiert. *M* Itemschwierigkeit = mittlere Itemschwierigkeit der Items eines Päckchens aus einer 1pl-IRT Skalierung aller Items eines Tests.

[34] Diese sind im Kompetenzmodell von Nachhaltigkeitsmanagement als Kern-, Management- und Supportprozesse definiert.

Messinvarianz

Um Testwerte von unterschiedlichen Zeitpunkten und in verschiedenen Gruppen vergleichen zu können, müssen die interessierenden latenten Variablen messinvariant sein. Da die Mittelwerte der latenten Variablen verglichen werden sollen, ist *skalare* Messinvarianz (MI) notwendig. Das bedeutet, dass neben der gleichen Faktorstruktur in den interessierenden Gruppen (*konfigurale* MI) auch die Faktorladungen (*metrische* MI) und die Interzepte der Indikatoren gleichgesetzt werden können, ohne dass diese Restriktionen zu einer signifikant schlechteren Modellpassung der Daten führt (zum Vorgehen vgl. Kline, 2011, Kapitel 9).

Um konfigurale längsschnittliche MI zu etablieren, muss zunächst jeder Messzeitpunkt einen guten Modellfit zum angenommenen Modell aufweisen. Da im vorliegenden Fall die Tests eindimensional konzeptioniert wurden und die Analysen mit je drei Päckchen durchgeführt werden, ist das eindimensionale Modell genau identifiziert. Dadurch passen die Daten perfekt zum Modell. Als erster Schritt wird deshalb hier ein zweifaktorielles Modell mit korrelierten Faktoren präsentiert. Jeder Faktor steht für einen Messzeitpunkt, jeweils die Faktorladung des ersten Indikatorpäckchens ist auf 1 fixiert. Zusätzlich werden Korrelationen zwischen den zusammengehörenden Päckchen über die Messzeitpunkte zugelassen. Um möglichst sparsame Modelle zu schätzen, werden diese längsschnittliche Indikatorkorrelationen wieder auf Null fixiert, wenn die Korrelationen nicht signifikant sind. Dieses Modell stellt das Modell der konfiguralen MI dar. Im nächsten Schritt werden die Faktorladungen über Messzeitpunkte gleichgesetzt und die Veränderung der Modellgüte betrachtet. Sprechen die Gütekriterien für das Modell, wird von metrischer Invarianz ausgegangen. Anschließend werden die Interzepte der Indikatoren über Messzeitpunkte gleichgesetzt und der Faktormittelwert für den ersten Messzeitpunkts auf Null fixiert. Der Faktormittelwert des zweiten Messzeitpunkts wird frei geschätzt. Kann dieses Modell beibehalten werden, sind die Mittelwerte der Testwerte der beiden Messzeitpunkte vergleichbar.

Zur Beurteilung der Modellgüte wurde der χ^2-Likelihood Test herangezogen. Zusätzlich werden CFI, TLI, RMSEA und SRMR berichtet. Dabei werden Werte von CFI > .95, RMSEA < .06 und SRMR < .08 als Indikatoren für guten Modellfit betrachtet (Hu & Bentler, 1999). Ob eingeführte Parameterrestriktionen zu einem signifikant schlechteren Modellfit führen, wird über den skalierten χ^2-Differenztest (Satorra & Bentler, 2001) geprüft. Zusätzlich wird die Veränderung des Chi-Quadrat Werts im Verhältnis zur Veränderung der Freiheitsgrade, und die Veränderungen in CFI, RMSEA und SRMR betrachtet. Dabei werden Werte von $\Delta\chi^2 \geq 3*\Delta df$

(Schermelleh-Engel, Moosbrugger & Müller, 2003), ΔCFI ≥ -.01, ΔRMSEA ≥ .015 und ΔSRMR ≥ .03 (bzw. ≥ .01 für die Invarianz von Interzepten) als Indikatoren für die Verletzung von Invarianz betrachtet (Chen, 2007). Weist eine Stufe der Messinvarianz einen absoluten oder relativ schlechten Modellfit auf, wird das Modell der vorherigen invarianten Stufe inspiziert. Dadurch sollen Parameter identifiziert werden, deren Gleichsetzen zu einem schlechten Modellfit führt. Wenn durch Freisetzen einzelner Parameter über Zeitpunkte ein akzeptables Modell geschätzt wird, das nicht schlechter zu den Daten passt als das der vorherigen MI-Stufe, wird partielle Invarianz angenommen. Zu beachten ist, dass die mit einem partiell invarianten Modell durchgeführten Analysen nur hinsichtlich der weiterhin invarianten Modellparameter interpretiert werden dürfen.

In Tabelle 13 sind die Gütekriterien für die Modelle unterschiedlicher Stufen von MI aufgeführt. Für die Tests BWL und dNCM sprechen alle Kriterien der Modellgüte für längsschnittliche skalare MI. Beim Test NagP sprechen die Kriterien der globalen Modellgüte für eine gut Passung des Modells der skalaren MI an die Daten. Der Modellfit ist jedoch signifikant schlechter als der des Modells der metrischen MI. Nach Inspektion des metrischen Invarianzmodells wurde ein Interzept für die Messzeitpunkte frei geschätzt. Das resultierende partielle skalare Invarianzmodell erreicht einen guten Modellfit und das Modell passt nicht signifikant schlechter auf die Daten als das Modell der metrischen MI.

Im sNCM-Test spricht der CFI gegen die Passung des Modells der metrischen Invarianz auf die Daten. Zudem passt das Modell signifikant schlechter auf die Daten als das konfigurale Invarianzmodell. Eine Inspektion der Faktorladungen im Modell der konfiguralen Invarianz zeigt, dass die Ladungen für das zweite Päckchen vermutlich der Grund für die schlechte Modellpassung sind. Ein partielles metrisches Invarianzmodell, mit nicht restringierten Faktorladungen für das zweite Päckchen, zeigt eine gute Modellpassung und passt nicht signifikant schlechter auf die Daten als das konfigurale Invarianzmodell. Für ein partielles skalares Invarianzmodell wurden die Interzepte für die weiterhin invarianten Päckchen gleichgesetzt. Für das Päckchen mit varianten Faktorladungen wurde das Interzept frei geschätzt. In diesem Modell wird die Residualvarianz für das Päckchen mit nicht restringierten Faktorladungen negativ geschätzt. Da das 95% bootstrap-Konfidenzintervall die Null einschließt, 95% KI_{boot} [-0.27; 0.06], basierend auf 1000 Replikationen, liegen keine Hinweise für eine Fehlspezifikation des Modells vor (Kollenikov & Bollen, 2012). Daher wird die Residualvarianz auf einen positiven Wertebereich fixiert. Das partielle skalare Invarianzmodell zeigt eine gute Passung auf die Daten. Die Restriktionen führen zu keinem signifikant schlechteren Modellfit,

die Veränderung im SRMR und im Chi-Quadrat Wert ist kleiner als die definierten Cut-offs. Die Veränderung des CFI und RMSEA sprechen gegen die Invarianz des Modells über Messzeitpunkte. Da die Mehrheit der Kriterien für die Passung des Modells zu den Daten spricht, wird ein partielles skalares längsschnittliches Invarianzmodell für den sNCM-Test angenommen.

Fazit zur Vergleichbarkeit der Testwerte über Messzeitpunkte hinweg. Zusammenfassend lässt sich sagen, dass für die Tests BWL und dNCM skalare längsschnittliche Invarianz und für die Tests NagP und sNCM partielle skalare längsschnittliche Invarianz angenommen wird. Die Mittelwerte der beiden Messzeitpunkte können also für alle Tests verglichen werden. Bei den Tests NagP und sNCM beruhen die Mittelwertvergleiche jedoch nur auf den zwei invarianten Indikatoren-Päckchen.

Der NagP-Test wurde ohne inhaltliche Unterbereiche konzipiert. Von den 43 für die Analysen verwendeten Items sind noch 29 Items vergleichbar. Daher wird keine Einschränkung der Konstruktbedeutung für den NagP Test erwartet.

Im sNCM-Test enthalten die zwei invarianten Päckchen acht Situationen. Nicht vergleichbar sind Situationen die in den Unternehmensbereichen Produktentwicklung (Situation 3), Marketing (Situation 5), Reporting (Situation 6) und Transport (Situation 10) angesiedelt sind.

Situation 3 besteht aus je zwei offen und teil-offen konstruierten Items. Im Fokus steht ausschließlich strategisches Wissen zu Produktentwicklung, welches argumentativ demonstriert werden muss. Die Anforderungen und inhaltliche Thematik von Situation 3 kann durch Situation 4 aufgefangen werden, welche in einem invarianten Päckchen steckt. Situation 4 spielt ebenfalls in der Produktentwicklung. Je zwei Items sind offen, teil- offen und geschlossen konstruiert, fünf der Items zielen auf strategisches Wissen. Daher deckt Situation 4 nicht nur Inhalte, sondern auch Itemformate und die intendierte Wissensstufen (nach der Klassifikation von Shavelson et al., 2005) von Situation 3 ab. Es wird deshalb keine Einschränkung der Konstruktbedeutung durch die Nicht-Vergleichbarkeit von Situation 3 erwartet.

Situation 5 spielt im Unternehmensbereich Marketing, der nicht in anderen Situationen aufgegriffen wird. Für die folgenden Analysen bedeutet das, dass Lernfortschritte und die Erklärung dieser Lernfortschritte nicht durch Items aus dem Inhaltsbereich Marketing gestützt werden. Die Ergebnisse können möglicherweise nicht auf diesen Bereich generalisiert werden. Marketing ist den Supportprozessen im Unternehmensablauf zugeordnet, ebenso wie Personalentwicklung (Situation 13) oder Bilanzierung (Situation 11). Daher werden neben der

inhaltlichen Einschränkung keine Konsequenzen hinsichtlich der Bedeutung der Ergebnisse für Supportprozesse im Unternehmen erwartet.

Situation 6 stellt den Unternehmensbereich Reporting dar und besteht aus sechs Items die mehrheitlich geschlossen konstruiert sind (5 Items). Diese erfassen prozedurales und deklaratives Wissen zu *Integrated Reporting*, einer Methode zur Berichtlegung aller wertschöpfenden Prozesse des Unternehmens. Dabei sollen sowohl finanzielle als auch ökologische Auswirkungen der Prozesse berücksichtigt werden. Die drei Lehrveranstaltungen in der Schwerpunktgruppe weisen unterschiedlichen Curricula auf. Die Nicht-Invarianz der Situation 6 könnte daher mit unterschiedlichen Schwerpunkten in den Lehrveranstaltungen und unterschiedlichen Ausprägungen an Lerngelegenheiten zu dieser spezifischen Methode zusammenhängen. Das Reporting lässt sich den Supportprozessen zurechnen, das Analysen und Zahlen für Entscheidungen auf Managementebene bereitstellt. Während Situation 6 deklaratives und prozedurales Wissen über die Methode Integrated Reporting thematisiert, wird in Situation 11 deklaratives und prozedurales Wissen über die Methode *Ökobilanz* erfasst. Hier werden jedoch nur die Auswirkungen von Geschäftsprozessen auf die Umwelt analysiert. Im Gegensatz zu Integrated Reporting umfasst die Ökobilanz also nur die ökologische Perspektive von Nachhaltigkeitsmanagement. Auch Bilanzierung zählt zu den Supportprozessen, welche die Grundlage für Entscheidungen auf Managementebene bilden. Daher werden neben der inhaltlichen Einschränkung in der Konstruktbedeutung keine Konsequenzen hinsichtlich der Bedeutung der Ergebnisse für Supportprozesse im Unternehmen erwartet.

In der nicht invarianten Situation 10 soll ein Spediteur und ein Transportmittel für die Produktauslieferung zu den Kunden ausgewählt werden. In Situation 9 muss für eine ähnliche Aufgabenstellung (Auswahl eines Lieferanten) die gleiche Methode angewandt werden wie in Situation 10, eine *Nutzwertanalyse*. Beide Situationen bilden Kernprozesse in einem Unternehmen ab. Aufgrund der Ähnlichkeit der Situationen wurden diese nicht in einem Testheft präsentiert. Daher wird erwartet, dass die Bedeutung der Faktormittelwerte trotz der Nicht-Invarianz von Situation 10 immer noch hinreichend die intendierte Kompetenz abbilden.

Für den sNCM-Test werden Einschränkungen hinsichtlich der vergleichbaren Konstruktbedeutung erwartet. Da Kompetenz in Nachhaltigkeitsmanagement die Integration von ökologischen, sozialen und ökonomischen Aspekten in alle betrieblichen Entscheidungen umfasst (Seeber et al., 2019), wurden Aufgaben zu allen Unternehmensbereichen entwickelt. Items aus den Unternehmensbereichen Marketing und Reporting sind jedoch nicht invariant

über Messzeitpunkte und können inhaltlich auch nicht durch Items anderer Situationen abgedeckt werden. Für den sNCM-Test können die Analyse, welche auf Basis der invarianten Päckchen stattfinden, nicht auf alle Unternehmensbereiche übertragen werden. Es sind jedoch weiterhin Items aus den Management-, Support und Produktionsprozessen eines Unternehmens vergleichbar über Messzeitpunkte.

Tabelle 13 Gütekriterien und Modellvergleiche für die Stufen der Messinvarianz über Messzeitpunkte für die Leistungstests

	BWL [a]					NagP [a,b]				
	$\chi^2(df)$ [c] *p* =	CFI ΔCFI	RMSEA [90% KI] ΔRMSEA	SRMR ΔSRMR	$\Delta\chi^2(\Delta df)$ [d] *p* =	$\chi^2(df)$ [c] *p* =	CFI ΔCFI	RMSEA [90% KI] ΔRMSEA	SRMR ΔSRMR	$\Delta\chi^2(\Delta df)$ [d] *p* =
konfigural	3.86 (6) *p* = .696	1.00	.000 [.000 - .041]	.028		2.712 (7) *p* = .91	1.00	.000 [.000 - .000]	.02	
metrisch	6.90 (8) *p* = .547	1.00 ΔCFI = .00	.000 [.000 - .046] ΔRMSEA = .00	.039 ΔSRMR = .011	3.784 (2) *p* = .151	4.618 (9) *p* = .866	1.00 ΔCFI = .00	.000 [.000 - .021] ΔRMSEA = .00	.03 ΔSRMR = .01	3.114 (2) *p* = .211
skalar	10.712 (10) *p* = .38	.993 ΔCFI = .007	.012 [.000 - .049] ΔRMSEA = .012	.041 ΔSRMR = .002	3.594 (2) *p* = .166	10.191 (11) *p* = .513	1.00 ΔCFI = .00	.000 [.000 - .044] ΔRMSEA = .00	.045 **ΔSRMR = .015**	**6.815 (2)** ***p* = .033**
partiell skalar						5.018 (10) *p* = .89	1.00 ΔCFI = .00	.000 [.000 - .018] ΔRMSEA = .00	.031 ΔSRMR = .001	0.348 (1) *p* = .555

Fortsetzung auf der nächsten Seite

	dNCM [a]					sNCM [a,e]				
	$\chi^2(df)$ [c] p =	CFI ΔCFI	RMSEA [90% KI] ΔRMSEA	SRMR ΔSRMR	$\Delta\chi^2(\Delta df)$ [d] p =	$\chi^2(df)$ [c] p =	CFI ΔCFI	RMSEA [90% KI] ΔRMSEA	SRMR ΔSRMR	$\Delta\chi^2(\Delta df)$ [d] p =
konfigural	3.726 (7) p = .811	1.00	.000 [.000 - .032]	.03		7.387 (7) p = .390	.995	.011 [.000 - .060]	.052	
metrisch	5.712 (9) p = .768	1.00 ΔCFI = .00	.000 [.000 - .036] ΔRMSEA = .00	.029 ΔSRMR = -.001	2.884 (2) p = .237	14.537 (9) p = .104	**.928** **ΔCFI = -.067**	.037 [.000 - .07] **ΔRMSEA = .026**	.07 ΔSRMR = .018	**6.69 (2)** **p = .035**
partiell metrisch						7.944 (8) p = .439	1.00 ΔCFI = .005	.00 [.000 - .055] ΔRMSEA = -.011	.051 ΔSRMR = -.001	0.6063 (1) p = .436
(partiell) Skalar	7.28 (11) p = .776	1.00 ΔCFI = .00	.000 [.000 - .033] ΔRMSEA = .00	.032 ΔSRMR = .003	1.533 (2) p = .465	13.002 (10) p = .224	.961 **ΔCFI = -.039**	.026 [.000 - .061] **ΔRMSEA = .026**	.06 ΔSRMR = .009	4.893 (2) p = .087

Anmerkungen. Konfigural: zweidimensionales Modell (eine Dimension pro MZP), der erste Indikator eines Faktors ist auf 1 fixiert, sonstige Ladungen und Interzepte werden frei geschätzt, die latenten Faktoren korrelieren. Metrisch: Zusätzlich zum konfiguralen Modell werden Faktorladungen über MZP gleichgesetzt. Skalar: zusätzlich zum metrischen Modell werden Interzepte für Indikatoren über MZP gleichgesetzt. Mittelwert der latenten Variablen zum ersten MZP wird auf Null fixiert und für den zweiten MZP frei geschätzt. **FETT** gedruckt sind die Kennziffern, die auf einen schlechten Modellfit oder auf einen signifikant schlechteren Modellfit bzw. zu großen Veränderungen in den Fit-Werten nach Einführung von Gleichheitsrestriktionen bedeuten.

[a] Modelle beinhalten indikatorspezifische Korrelationen zwischen Residualvarianzen über MZP.

[b] Im Modell für die partielle skalare MI wird das Interzept für ein Päckchen von dreien über die MZP frei geschätzt.

[c] Es werden die robusten χ^2 Statistiken berichtet.

[d] Es wird ein skalierter χ^2 Differenztests berichtet (Satorra & Bentler, 2001).

[e] Im Modell für die partielle metrische Invarianz im sNCM Test wird eine Faktorladung von dreien über MZP frei geschätzt. Im nachfolgenden Modell der partiellen skalaren Invarianz wird das entsprechende Interzept frei geschätzt.

8.2.3.2 Ergebnisse

Im Folgenden werden standardisierte Regressionskoeffizienten berichtet. Die Tabellen geben zweiseitige *p*-Werte an. Die Tabellen enthalten nur die Regressionsgewichte für die interpretierten Ergebnisse. In Anhang K bis N sind Tabellen inklusive der Regressionsgewichte der einzelnen Standorte abgebildet. Im Text werden einseitige *p*-Werte berichtet, wenn die Hypothese gerichtet war.

Zusammenhang von Testwerten und hochschulischen Lerngelegenheiten

Analysemodell 1. Analysen zu den Evidenzen 3a bis 3c

Im BWL-Test zeigt das Modell mit festen Effekten eine gute Passung an die Daten. Im Ausgangsniveau unterscheiden sich die Gruppen nicht signifikant. Unter Kontrolle des Ausgangsniveaus unterscheiden sich die mittleren Lernfortschritte zwischen den zwei Gruppen nicht (Tabelle 14).

Tabelle 14 Lineare Kontraste der Schwerpunkt- und Kontrollgruppen mit den BWL-Testwerten als abhängigen Variablen (Modell 1)

	β	*SE*	*p*
BWL.T1 ~			
Ø Kontrollgruppen	.098	.021	<.001
Ø Schwerpunktgruppen	.026	.053	.629
Kontrast.T1	-.072	.046	.119
BWL.T2 ~			
Ø Kontrollgruppen	-.037	.056	.510
Ø Schwerpunktgruppen	.119	.132	.368
Kontrast.T2	.156	.094	.097
BWL.T1	.760	.184	<.001
R² (erklärte Varianz)	.72		

Anmerkungen. Die Regressionsgewichte der Kontroll- und Schwerpunktgruppen basieren auf den standardisierten Regressionsgewichten der Standorte. „Ø Schwerpunktgruppen“: die mittleren Regressionsgewichte der Standorte der Schwerpunktgruppe, gewichtet nach der jeweiligen Gruppengröße. „Ø Kontrollgruppen“: die mittleren Regressionsgewichte der Standorte der Kontrollgruppe, gewichtet nach der jeweiligen Gruppengröße. „Kontrast“ = Ø Schwerpunktgruppen *minus* Ø Kontrollgruppen.

Modellfit: $\chi^2(38) = 32.40$, $p = .73$, CFI = 1.00, TLI = 1.06, RMSEA = .00, 90% KI [.00 - .02], SRMR = .03.

In Tabelle 15 sind die Ergebnisse zu den nachhaltigkeitsbezogenen Tests NagP, dNCM und sNCM dargestellt. Im NagP-Test gibt es keinen Unterschied im Ausgangswert zwischen Kontroll- und Schwerpunktgruppe. Der Lernfortschritt ist in der Schwerpunktgruppe signifikant höher. Auffällig ist, dass der standardisierte Koeffizient für den autoregressiven Effekt der NagP-Testwerte in Modell 1 größer als 1 ist und im Modell 100% der Varianz der Testwerte zum zweiten Messzeitpunkt erklärt werden. Im dNCM-Test zeigt die Kontrollgruppe ein höheres Ausgangsniveau als die Schwerpunktgruppe. Der Lernfortschritt in der Schwerpunkgruppe ist jedoch signifikant höher als in der Kontrollgruppe. Im sNCM-Test gibt es zwischen Schwerpunkt- und Kontrollgruppe weder im Ausgangswert noch im Lernfortschritt einen Unterschied.

Es zeigt sich eine gute Passung des Modells 1 für NagP und dNCM. Im sNCM-Test zeigt das Modell eine mäßig gute Passung an die Daten, CFI- und TLI-Wert deuten auf eine Fehlspezifikation des Modells hin.

Tabelle 15 Lineare Kontraste der Schwerpunkt- und Kontrollgruppen mit den NagP-, dNCM- und sNCM-Testwerten als abhängigen Variablen (Modell 1)

	NagP [a]			**dNCM**			**sNCM**		
	β	*SE*	*p*	β	*SE*	*p*	β	*SE*	*p*
Testwert.T1 ~									
Ø Kontrollgruppen	.014	.018	.462	.043	.019	.021	.052	.023	.026
Ø Schwerpunktgruppen	.002	.056	.969	-.074	.062	.234	.143	.058	.013
Kontrast.T1	-.011	.051	.822	-.117	.055	.033	.091	.048	.055
Testwert.T2 ~									
Ø Kontrollgruppen	-.012	.042	.781	-.061	.047	.202	.123	.047	.009
Ø Schwerpunktgruppen	.286	.106	.007	.185	.125	.140	.210	.126	.095
Kontrast.T2	.297	.080	<.001	.245	.095	.010	.087	.094	.357
Testwert.T1	1.161	.103	<.001	.791	.235	.001	.781	.186	<.001
R^2 (erklärte Varianz)	1.00			.50			.91		

Anmerkungen. Die Regressionsgewichte der Kontroll- und Schwerpunktgruppen basieren auf den standardisierten Regressionsgewichten der Standorte. „Ø Schwerpunktgruppen“: die mittleren Regressionsgewichte der Standorte der Schwerpunktgruppe, gewichtet nach der jeweiligen Gruppengröße. „Ø Kontrollgruppen“: die mittleren Regressionsgewichte der Standorte der Kontrollgruppe, gewichtet nach der jeweiligen Gruppengröße. „Kontrast“ = Ø Schwerpunktgruppen *minus* Ø Kontrollgruppen.

Modellfit NagP: $\chi^2(38) = 35.22$, $p = .60$, CFI = 1.00, TLI = 1.02, RMSEA = .00, 90% KI [.00 - .03], SRMR = .04. dNCM: $\chi^2(39) = 36.26$, $p = .60$, CFI = 1.00, TLI = 1.03, RMSEA = .00, 90% KI [.00 - .03], SRMR = .04. sNCM: $\chi^2(37) = 50.14$, $p = .07$, CFI = .89, TLI = .83, RMSEA = .03, 90% KI [.00 - .04], SRMR = .05.

[a] Für das Modell wurde eine Varianz auf einen Wertebereich größer gleich Null fixiert.

Analysemodell 2. Analysen zu den Evidenzen 3d und 3e

In Modell 2 werden die selbsteingeschätzten hochschulischen Lerngelegenheiten hinsichtlich Aspekten von Nachhaltigkeit als Prädiktoren der Testwerte zu Messzeitpunkt 1 und 2 verwendet. Die festen Effekte der Standorte bleiben im Modell, um der Clusterung der Daten Rechnung zu tragen. Dadurch lassen sich die Ergebnisse für die selbstberichteten hochschulischen Lerngelegenheiten nur innerhalb der Kurse interpretieren. Die Ergebnisse zu den Modellen 2 und Alternativmodellen 2 sind in Tabelle 16 bis Tabelle 18 abgebildet.

Die selbstberichteten hochschulischen Lerngelegenheiten zu Nachhaltigkeitsmanagement sagen in keinem der Tests den Ausgangswert vorher. Für die Prädiktion des Lernfortschritts zeigen sich unterschiedliche Ergebnisse. Im NagP-Test wird Lernfortschritt nicht vorhergesagt. Im dNCM-Test wird Lernfortschritt bei einseitigem Test knapp nicht signifikant vorhergesagt, $\beta = .22$, $SE(\beta) = .13$, $p = .05$. Der Lernfortschritt im sNCM-Test wird signifikant vorhergesagt, allerdings mit negativem Vorzeichen. Zu beachten ist, dass der standardisierte Koeffizient für den autoregressiven Effekt der NagP-Testwerte in Modell 1 größer als 1 ist und im Modell 100% der Varianz der Testwerte zum zweiten Messzeitpunkt erklärt werden. Für die Tests NagP und dNCM passt das Modell 2 jeweils gut auf die Daten. Für den sNCM-Test passt das Modell akzeptabel bis gut zu den Daten.

Alternativmodell 2

In Alternativmodell 2 sind keine festen Effekte der Kurse modelliert. Die Clusterung der Daten wird durch eine post-hoc Korrektur der Standardfehler berücksichtigt. Die selbstberichteten hochschulischen Lerngelegenheiten sagen in keinem der Tests den Ausgangswert vorher. In allen Tests zeigen die selbstberichteten Prädiktoren einen signifikanten Zusammenhang zum Lernfortschritt. Der Lernfortschritt wird in den Tests NagP und dNCM positiv und im sNCM-Test negativ vorhergesagt. Der Modellfit wird für alle Tests als schlecht eingeschätzt.

Tabelle 16 Standardisierte Regressionsgewichte der selbstberichteten hochschulischen Lerngelegenheiten mit den NagP-Testwerten als abhängigen Variablen unter Kontrolle der festen Effekte (Modell 2) bzw. mit post-hoc Korrektur der Standardfehler (Alternativmodell 2)

	Modell 2 [a]			Alternativmodell 2		
	β	*SE*	*p*	β	*SE*	*p*
NagP.T1 ~						
Hochschulische LG.T1	-.021	.066	.755	.026	.095	.787
NagP.T2 ~						
NagP.T1	1.162	.104	<.001	.943	.150	<.001
Hochschulische LG.T2	.026	.121	.831	.227	.106	.032
Hochschulische LG.T2 ~						
Hochschulische LG.T1	.452	.053	<.001	.451	.102	<.001
R^2 (erklärte Varianz)	1.00			.95		

Anmerkungen. Modellfit NagP, Model 2: $\chi^2 = 41.277(48)$ $p = .743$, CFI = 1.00, TLI = 1.03, RMSEA = .00, 90% KI [.00 - .02], SRMR = 0.04. Alternativmodell 2: $\chi^2 = 120.42(62)$ $p < .001$, CFI = .83, TLI = .79, RMSEA = .04, 90% KI = [.03 - .06], SRMR = .08.

[a] Für das Modell wurde eine Varianz auf einen Wertebereich größer gleich Null fixiert.

Tabelle 17 Standardisierte Regressionsgewichte der selbstberichteten hochschulischen Lerngelegenheiten mit den dNCM-Testwerten als abhängigen Variablen unter Kontrolle der festen Effekte (Modell 2) bzw. mit post-hoc Korrektur der Standardfehler (Alternativmodell 2)

	Modell 2			Alternativmodell 2		
	β	*SE*	*p*	β	*SE*	*p*
dNCM.T1 ~						
Hochschulische LG.T1	-.009	.072	.904	.071	.091	.436
dNCM.T2 ~						
dNCM.T1	.777	.237	.001	.580	.091	<.001
Hochschulische LG.T2	.218	.132	.100	.268	.101	.008
Hochschulische LG.T2 ~						
Hochschulische LG.T1	.449	.053	<.001	.447	.104	<.001
R^2 (erklärte Varianz)	.52			.42		

Anmerkungen. Modellfit dNCM, Modell 2: $\chi^2(49) = 49.30$ $p = .46$, CFI = 1.00, TLI = 1.00, RMSEA = .00, 90 % KI = [.00 - .03], SRMR = .04. Alternativmodell 2: $\chi^2 = 106.42(63)$ $p = .001$, CFI = .84, TLI = .81, RMSEA = .04, 90% KI [.02 - .05], SRMR = .08.

[a] Für das Modell wurde eine Varianz auf einen Wertebereich größer gleich Null fixiert.

Tabelle 18 Standardisierte Regressionsgewichte der selbstberichteten hochschulischen Lerngelegenheiten mit den sNCM-Testwerten als abhängigen Variablen unter Kontrolle der festen Effekte (Modell 2) bzw. mit post-hoc Korrektur der Standardfehler (Alternativmodell 2)

	Modell 2 [a]			Alternativmodell 2 [a]		
	β	*SE*	*p*	β	*SE*	*p*
sNCM.T1 ~						
Hochschulische LG.T1	-.045	.075	.550	.021	.065	.743
sNCM.T2 ~						
sNCM.T1	.765	.111	<.001	.603	.171	<.001
Hochschulische LG.T2	-.402	.154	.009	-.310	.173	.073
Hochschulische LG.T2 ~						
Hochschulische LG.T1	.453	.052	<.001	.452	.097	<.001
R^2 (erklärte Varianz)	1.00			.46		

Anmerkungen. Modellfit sNCM, Modell 2: $\chi^2(47)= 57.89$, $p = .13$, CFI = .95, TLI = .92, RMSEA = .02, 90% KI [.00 - .04], SRMR = .04. Alternativmodell 2: $\chi^2 = 119.83(61)$ $p < .001$, CFI = .75, TLI = .68, RMSEA = .04, 90% KI [.03 - .06], SRMR = .07.

[a] Für das Modell wurde eine Varianz auf einen Wertebereich größer gleich Null fixiert.

Zusammenhang von Testwerten und außerhochschulischen Lerngelegenheiten

Analysemodell 3. Analysen zu den Evidenzen 3e bis 3i

In Modell 3 werden die selbstberichteten außerhochschulischen Lerngelegenheiten hinsichtlich Aspekten von Nachhaltigkeit als Prädiktoren der Testwerte zu Messzeitpunkt 1 und 2 verwendet. Die festen Effekte der Standorte bleiben im Modell, um der Clusterung der Daten Rechnung zu tragen. Dadurch lassen sich die Ergebnisse für die selbstberichteten außerhochschulischen Lerngelegenheiten nur innerhalb der Kurse interpretieren. Die selbstberichteten außerhochschulischen Lerngelegenheiten sagen positiv den Ausgangswert im NagP-Test vorher, jedoch nicht den Lernfortschritt. In den Tests dNCM und sNCM wird weder der Ausgangswert noch der Lernfortschritt signifikant vorhergesagt. Der Modellfit wird für alle Tests als gut beurteilt.

Alternativmodell 3

In Alternativmodell 3 sind keine festen Effekte der Kurse modelliert. Die Clusterung der Daten wird durch eine post-hoc Korrektur der Standardfehler berücksichtigt. Die Alternativmodelle replizieren die Befunde aus den Modellen mit festen Effekten. Der einzige signifikante Zusammenhang besteht zwischen selbstberichteten außerhochschulischen Lerngelegenheiten und dem Ausgangsniveau im NagP-Test. Der Modellfit wird für alle Tests als gut beurteilt. Die Ergebnisse zu den Modellen 3 und Alternativmodellen sind in Tabelle 19 bis Tabelle 21 abgebildet.

Tabelle 19 Standardisierte Regressionsgewichte der selbstberichteten außerhochschulischen Lerngelegenheiten mit den NagP-Testwerten als abhängigen Variablen unter Kontrolle der festen Effekte (Modell 3) bzw. mit post-hoc Korrektur der Standardfehler (Alternativmodell 3)

	Modell 3 [a]			Alternativmodell 3 [a]		
	β	*SE*	*p*	β	*SE*	*p*
NagP.T1 ~						
Außerhochschulische LG.T1	.112	.067	.096	.212	.093	.024
NagP.T2 ~						
NagP.T1	1.165	.103	<.001	.982	.027	<.001
Außerhochschulische LG.T2	.026	.096	.784	.097	.105	.358
Außerhochschulische LG.T2 ~						
Außerhochschulische LG.T1	.658	.048	<.001	.653	.086	<.001
R^2 (erklärte Varianz)	1.00			1.00		

Anmerkungen. Regressionsgewichte der festen Effekte sind in Relation zur Referenzgruppe *Göttingen C* (Kontrollgruppe) angegeben.

Modellfit NagP, Modell 3: $\chi^2(55) = 60.10$, $p = .30$, CFI =.98, TLI = .98, RMSEA = .01, 90% KI [.00 - .03], SRMR = .04. Alternativmodell 3: $\chi^2 = 24.87(20)$ $p = .21$, CFI = .98, TLI = .97, RMSEA = .00, 90% KI = [.00 - .05], SRMR = .05.

[a] Für das Modell wurde eine Varianz auf einen Wertebereich größer gleich Null fixiert.

Tabelle 20 Standardisierte Regressionsgewichte der selbstberichteten außerhochschulischen Lerngelegenheiten mit den dNCM-Testwerten als abhängigen Variablen unter Kontrolle der festen Effekte (Modell 3) bzw. mit post-hoc Korrektur der Standardfehler (Alternativmodell 3)

	Modell 3			Alternativmodell 3		
	β	*SE*	*p*	β	*SE*	*p*
dNCM.T1 ~						
Außerhochschulische LG.T1	.084	.073	.248	.156	.103	.130
dNCM.T2 ~						
dNCM.T1	.770	.232	.001	.618	.077	<.001
Außerhochschulische LG.T2	-.005	.109	.960	.069	.066	.291
Außerhochschulische LG.T2 ~						
Außerhochschulische LG.T1	.655	.048	<.001	.649	.089	<.001
R^2 (erklärte Varianz)	.49			.40		

Anmerkungen. Regressionsgewichte der festen Effekte sind in Relation zur Referenzgruppe *Göttingen C* (Kontrollgruppe) angegeben.

Modellfit dNCM, Modell 3: $\chi^2(56) = 51.98$, $p = .63$, CFI = 1.00, TLI = 1.03, RMSEA = .00, 90% KI [.00 - .02], SRMR = .04. Alternativmodell 3: $\chi^2 = 18.76(21)$ $p = .601$, CFI = 1.00, TLI = 1.02, RMSEA = .00, 90% KI [.00 - .04], SRMR = .04.

Tabelle 21 Standardisierte Regressionsgewichte der selbstberichteten außerhochschulischen Lerngelegenheiten mit den sNCM-Testwerten als abhängigen Variablen unter Kontrolle der festen Effekte (Modell 3) bzw. mit post-hoc Korrektur der Standardfehler (Alternativmodell 3)

	Modell 3			Alternativmodell 3		
	β	*SE*	*p*	β	*SE*	*p*
sNCM.T1 ~						
Außerhochschulische LG.T1	.002	.076	.976	.028	.093	.761
sNCM.T2 ~						
sNCM.T1	.786	.182	<.001	.630	.230	.006
Außerhochschulische LG.T2	-.177	.111	.129	-.191	.141	.174
Außerhochschulische LG.T2 ~						
Außerhochschulische LG.T1	.656	.048	< .001	.650	.087	<.001
R² (erklärte Varianz)		.95			.43	

Anmerkungen. Regressionsgewichte der festen Effekte sind in Relation zur Referenzgruppe *Göttingen C* (Kontrollgruppe) angegeben.

Modellfit sNCM, Modell 3: $\chi^2(54) = 66.27$, $p = .12$, CFI = .93, TLI = .90, RMSEA = .02, 90% KI [.00 - .04], SRMR = .05. Alternativmodell 3: $\chi^2 = 19.74(19)$ $p = .41$, CFI = .99, TLI = .99, RMSEA = .01, 90% KI [.00 - .04], SRMR = .05.

8.2.3.3 Diskussion

Evidenz 3b und 3c

Wie vermutet zeigen die Schwerpunkt- und Kontrollgruppe weder im Ausgangsniveau noch im Lernfortschritt Unterschiede in den BWL-Testwerten.[35] Grundlegendes Wissen über Aspekte unternehmerischen Handelns scheint in beiden Gruppen vorhanden zu sein und verändert sich nicht differentiell während des getesteten Semesters.

Evidenz 3a, 3d und 3e: Hochschulische Lerngelegenheiten haben einen Erklärungswert für Testwerte in den nachhaltigkeitsbezogenen Tests.

Modell 1. Für die Tests NagP und dNCM sind die Lernfortschritte in der Schwerpunktgruppe größer als in der Kontrollgruppe. Studierende in Lehrveranstaltungen mit Bezug zu Nachhaltigkeit und Nachhaltigkeitsmanagement erreichen unter Kontrolle des

[35] Angesichts des fehlenden Lernfortschritts im BWL-Test kann keine bestätigende Evidenz für Grundannahme 2 festgestellt werden, dass der Test instruktionssensitiv ist.

Ausgangsniveaus höhere Testwerte zum zweiten Messzeitpunkt als Studierende, die nicht solche Kurse besuchen.[36]

Für den sNCM-Test zeigt sich zwischen den Gruppen kein Unterschied im Lernfortschritt. Dieser Test simuliert realitätsnahe Berufssituationen, in denen zwischen ökonomischen, sozialen und ökologischen Faktoren abgewogen werden muss. Dies steht teilweise in Widerspruch zu üblichen betriebswirtschaftlichen Modellen (z.B. *Homo Oeconomicus*; *Gewinnmaximierung*; *Kostenreduzierung*). Unter der Annahme, dass der sNCM-Test prinzipiell sensitiv für Veränderungen ist (siehe Diskussion zu Modell 2), könnten die Lerngelegenheiten für Kompetenz in Nachhaltigkeitsmanagement in einem Semester nicht ausreichend gewesen sein, um die in anderen betriebswirtschaftlichen Lehrveranstaltungen gelehrten Prinzipien weniger stark zu berücksichtigen. Dafür sprechen auch die hohen standardisierten autoregressiven Effekte der Testwerte. Für alle Tests liegen diese über .78. Bei einer Änderung des Ausgangswertes um eine Standardabweichung ändert sich also der Testwert zum zweiten Messzeitpunkt (unter Kontrolle des Ausgangsniveaus) um mindestens eine dreiviertel Standardabweichung. Insgesamt deutet das auf eine hohe Stabilität der Testwerte über ein Semester.

In den vorgestellten Analysen werden nur die kognitiven Aspekte von Nachhaltigkeitsmanagement berücksichtigt. In bisherigen Arbeiten zeigten diese den größten Einfluss auf Handlungsintentionen zu Nachhaltigkeitsmanagement (Seeber, Fischer, Michaelis & Müller, 2014; Michaelis, 2017). In diesen Arbeiten hängen aber auch nachhaltigkeitsbezogene Einstellungen in kleinerem Maße mit Handlungsintentionen in Nachhaltigkeitsmanagement zusammen. Eine aktuellere Arbeit aus dem Ko-NaMa-Projekt zeigt, dass Aversion gegenüber Nachhaltigkeit signifikant negativ die sNCM-Testwerten vorhersagt (Michaelis et al., 2020). Insbesondere im sNCM-Test, in dem nicht nur die Reproduktion von Wissen verlangt wird, sondern unterschiedliche Zieldimensionen abgewogen und eine Entscheidung begründet werden soll, könnte die reine Vermittlung von Wissen zu kurz greifen, um tatsächlich Lernfortschritte zu sehen. Diese Vermutung könnte durch eine erneute Analyse der Materialien

[36] Evidenzen 3a bis 3c liefern bestätigende Evidenz für Grundannahme 2. Da die Analysen Lernfortschritt in der Schwerpunktgruppe für die Tests NagP und dNCM zeigen, kann für diese Tests Instruktionssensitivität angenommen werden. Für den sNCM-Test zeigt sich Lernfortschritt in beiden Gruppen. Dieser Test ist also veränderungssensitiv. In wie weit sich Lernfortschritt jedoch auf Instruktion zurückführen lässt, ist hier fraglich.

der Lehrveranstaltungen gestützt werden, die über die Identifikation von behandelten Lehrthemen hinausgeht.

Modell 2. Hinsichtlich der selbstberichteten hochschulischen Lerngelegenheiten sind die Ergebnisse nicht einheitlich. In den Tests NagP und dNCM zeigt sich (knapp) kein Einfluss hochschulischer Lerngelegenheiten auf Lernfortschritt. Im sNCM-Test wird Lernfortschritt signifikant vorhergesagt, allerdings mit negativem Vorzeichen. Studierende die zum zweiten Messzeitpunkt mehr hochschulische Lerngelegenheiten zu Nachhaltigkeitsthemen berichten, zeigen einen niedrigeren Lernfortschritt als Studierende, die weniger solche Lerngelegenheiten berichten.

Im NagP-Test wird schon durch die festen Effekte und den autoregressiven Effekt sämtliche Varianz im Lernfortschritt erklärt (Modell 1). Die Hinzunahme von selbstberichteten hochschulischen Lerngelegenheiten kann keinen weiteren Erklärungswert bieten. Für alle Modelle 2 (und 3) gilt darüber hinaus, dass die Effekte der zusätzlichen Prädiktoren nur innerhalb der Kurse interpretiert werden dürfen, da die Modellierung der Standorte als feste Effekte sämtliche Varianz auf Ebene der Kurse erklärt.

In den Alternativmodellen 2 zeigen sich in den Tests NagP und dNCM die erwarteten positiven Vorhersagen von Lernfortschritten durch selbstberichtete hochschulische Lerngelegenheiten. Wie im Methodenteil erläutert, sollten die angegebenen Standardfehler und zugehörigen *p*-Werte jedoch vorsichtig interpretiert werden. Zusätzlich weisen alle Alternativmodelle 2 einen schlechten Modellfit auf, wenn CFI- und TLI-Wert betrachtet werden.

Der negative Zusammenhang im sNCM-Test (der auch in Alternativmodell 2 besteht) könnte mit einer unterschiedlichen Gewichtung der Dimensionen Sozial, Ökologisch und Ökonomisch in den Instrumenten zur Erfassung von Lerngelegenheiten und im sNCM-Test zusammenhängen. Der Indikator selbstberichteter hochschulischer Lerngelegenheiten setzt sich aus drei Skalen zusammen. Zwei davon beziehen sich auf soziale und ökologische Nachhaltigkeitsthemen aus gesellschaftlicher Perspektive (*NoNa* und *GeNa*, insgesamt acht Items). Demgegenüber steht die Skala zu betrieblichen Nachhaltigkeitsthemen (*BetNa*, fünf Items). Da die Antworten für die vorgestellten Analysen gemittelt werden, sind die soziale und ökologische Perspektive im Indikator hochschulischer Lerngelegenheiten stärker gewichtet als die betriebliche Perspektive. Im Test müssen dahingegen die Zieldimensionen entsprechend der jeweiligen Unternehmenssituation gewichtet werden. Eine Entscheidungsfindung die überwiegend auf ökologische und soziale Faktoren fokussiert, führt nur zu einem mittleren Abschneiden im

sNCM-Test. Beispielsweise sollte in einem Item eine Lieferung per Luftfracht erfolgen, wenn nur so eine vertraglich vereinbarte Lieferfrist eingehalten werden kann. Hier müsste die ökonomische Zieldimension höher als die ökologische gewichtet werden, da Nachhaltigkeitsmanagement ein langfristig erfolgreiches Unternehmenshandeln voraussetzt (Seeber et al., 2019).

Evidenz 3f bis 3i: Außerhochschulische Lerngelegenheiten haben nur für den nachhaltigkeitsbezogenen Test aus gesamtgesellschaftlicher Perspektive einen Erklärungswert.

Selbstberichtete außerhochschulische Lerngelegenheiten sagen in keinem der Tests Lernfortschritte vorher. Während das für die Tests dNCM und sNCM erwartet wurde, widerspricht dies den Annahmen für den NagP-Test und Ergebnissen aus dem Bereich der beruflichen Bildung, wo ein Zusammenhang informaler Lerngelegenheiten auf Wissen über Nachhaltigkeit gefunden wurde (Michaelis, 2017). Die Ergebnisse werden von den Alternativmodellen 3 repliziert. Der nicht signifikante Zusammenhang im NagP-Test kann also nicht durch die bereits durch Modell 1 perfekt erklärte Varianz der Testwerte zum zweiten Messzeitpunkt erklärt werden.

In Modell 3 und Alternativmodell 3 sagen jedoch außerhochschulische Lerngelegenheiten das Ausgangsniveau im NagP-Test vorher. Insgesamt deuten die hier präsentierten Ergebnisse und die Vorstudie von Michaelis (2017) darauf hin, dass außerhochschulische Lerngelegenheiten zwar in Zusammenhang mit allgemeinem Wissen über Nachhaltigkeit stehen. Außerhochschulische Lerngelegenheiten scheinen jedoch keine bedeutsame Rolle mehr für Lernfortschritte zu spielen, wenn hochschulische Lerngelegenheiten berücksichtigt werden.

Limitationen

Die Interpretation der Ergebnisse wird durch einige Aspekte eingeschränkt.

Selbstselektion der Studierenden in die Schwerpunkgruppe. Im verwendeten Studiendesign wählten die Studierenden ihre Lehrveranstaltungen bzw. Studiengänge selbst und wurden nicht randomisiert auf Lehrveranstaltungen zugewiesen. Daher ist es plausibel anzunehmen, dass in der Schwerpunktgruppe vermehrt Studierende mit höherem Interesse an Nachhaltigkeitsthemen sind. Diese Selbstselektionsprozesse könnten zu systematischen Unterschieden im Vorwissen führen. Kompetenzunterschiede zum ersten Messzeitpunkt könnten wiederum mit höherem Lernfortschritt in der Schwerpunktgruppe zusammenhängen

(*Matthäus-Effekt*). Bislang gab es kein Instrument zur Erfassung von Kompetenz in Nachhaltigkeitsmanagement bei Studierenden. Daher gibt es bislang keine Befunde zur Abhängigkeit von Lernfortschritt vom Ausgangsniveau in der Kompetenz Nachhaltigkeitsmanagement. Diese bedingt aber, dass sich die Kompetenzen in Nachhaltigkeitsmanagement schon zum ersten Messzeitpunkt systematisch zwischen Studierenden der Schwerpunkt- und der Kontrollgruppe unterscheiden.

Unterschiede im Ausgangsniveau sollten nicht durch selbstberichtete *hochschulische Lerngelegenheiten* hervorgerufen werden. Zwar wählten die Studierenden der Schwerpunktgruppen freiwillig die entsprechende Veranstaltung bzw. einen Studiengang mit Schwerpunkt auf Nachhaltigkeitsthemen. In der Schwerpunktgruppe sollten also vorwiegend Personen sein, die sich für Nachhaltigkeitsmanagement oder Aspekte des Themas Nachhaltigkeit interessieren. Nachhaltigkeitsmanagement ist jedoch bislang in nur wenigen Studiengängen implementiert (Seeber et al., 2016). Deshalb werden hochschulische Lerngelegenheiten außerhalb der von der Stichprobe besuchten Studiengängen und Lehrveranstaltungen kaum vorhanden sein.

In Kapitel 5.3 in Tabelle 7 werden die deskriptiven Unterschiede zwischen Schwerpunkt- und Kontrollgruppe hinsichtlich selbstberichteter hochschulischer Lerngelegenheiten dargestellt. Wie vermutet, berichten Studierende aus der Schwerpunktgruppe zum ersten Messzeitpunkt nicht mehr hochschulische Lerngelegenheiten als Studierende aus den Kontrollgruppen, $t(174.41) = 1.79$, $p = .08$.

In der Schwerpunktgruppe werden jedoch zum ersten Messzeitpunkt mehr *außerhochschulische Lerngelegenheiten* berichtet als in der Kontrollgruppe, $t(149.22) = 4.59$, $p < .001$. Das deutet daraufhin, dass Studierende, die sich mehr für Nachhaltigkeitsthemen interessieren und sich häufiger mit diesen Themen in ihrer Freizeit beschäftigen, eher entsprechende Studiengänge oder Kurse wählen.

Diese systematischen Unterschiede in vorherigen Lerngelegenheiten könnten zu einem höheren Ausgangsniveau der Studierenden in der Schwerpunktgruppe führen und deren höheren Lernfortschritt in den Tests NagP und dNCM erklären. Die Ergebnisse aus Analysemodelle 1 entkräften diese Argumentation jedoch. Signifikante Unterschiede im Vorwissen zwischen Schwerpunkt- und Kontrollgruppe zeigen sich nur im dNCM-Test. Zum ersten Messzeitpunkt haben Studierende der Kontrollgruppe höhere Testwerte als Studierende der Schwerpunktgruppe. Zum zweiten Messzeitpunkt dreht sich dieses Muster jedoch: Studierende

aus der Schwerpunktgruppe zeigen höheren Lernfortschritt als Studierende aus der Kontrollgruppe (siehe Tabelle 15). Dies spricht gegen die Argumentation, dass der höhere Lernfortschritt in der Schwerpunktgruppe auf höhere Testwerte im Ausgangsniveau zurückzuführen sei.

Umgang mit fehlenden Werten. In der deskriptiven Analyse der Daten fiel der hohe Prozentsatz an fehlenden Werten auf. Ein hoher Anteil der Studierenden erschien nicht zur Messwiederholung (N_1 = 499, N_2 = 189). In den Fragebogenskalen fehlen rund ein Viertel aller demografischen Angaben und etwa 40% bei den Fragebogenskalen.

Der hohe Prozentsatz fehlender Werte in den demografischen Angaben und Fragebogenskalen liegt vermutlich am Studiendesign. Die Bearbeitung der simulativen Aufgaben im sNCM-Test war nicht zeitlich beschränkt und im Testablauf vor der Beantwortung der demografischen Angaben und Fragebogenskalen vorgesehen. Letztere wurden zum Abschluss der Testung präsentiert. Die Testzeit war insgesamt jedoch auf 90 Minuten beschränkt. Daher wird vermutet, dass die fehlenden Werte bei den demographischen Angaben und in den Fragebogenskalen hauptsächlich durch Zeitmangel entstanden.

Problematisch ist jedoch der hohe Ausfall zum zweiten Messzeitpunkt. Wenn Studierende mit weniger Interesse an Nachhaltigkeitsthemen oder geringerer Motivation am Test teilzunehmen der Messwiederholung fernbleiben, könnten die Testwerte in der Messwiederholung positiv verzerrt sein. Den fehlenden Werten wurde in dieser Analyse mit Maximum-Likelihood Schätzern begegnet. Diese werden so geschätzt, dass die Wahrscheinlichkeit der vorhandenen Daten maximiert wird. Für jede Person im Datensatz wird eine eigene Likelihood geschätzt, deren Wahrscheinlichkeit maximiert wird gegeben der für diese Person vorhandenen Daten. Dadurch können Fälle mit fehlenden Werten in den Analysen berücksichtigt werden, die fehlenden Werte selbst werden jedoch nicht berücksichtigt (Lüdtke, Robitzsch, Trautwein & Köller, 2007). Das Vorgehen ist dann angemessen, wenn fehlende Werte nicht von der Ausprägung des Wertes selbst abhängen, sondern durch andere beobachtete Variablen erklärt werden können (*Missing at Random* [MAR] nach Rubin, 1976). Ist diese Annahme verletzt, weil z.B. Studierende mit geringeren Kompetenzen in Nachhaltigkeitsmanagement die zweite Messwiederholung eher versäumen oder Items eher überspringen, fehlen die Werte *Not Missing at Random* (NMAR). In diesen Fällen kann die Maximum-Likelihood Schätzung zu verzerrten Schätzern führen (z.B. Lüdtke et al., 2007).

Ein anderes Verfahren zum Umgang mit fehlenden Werten unter der Annahme MAR stellt Multiple Imputation dar (Lüdtke et al., 2007). Bei Multipler Imputation werden fehlende Werte zunächst durch mehrere *plausible Werte* ersetzt, die durch ein Imputationsmodell geschätzt werden. Das Imputationsmodell enthält dabei meist nicht nur die im Analysemodell verwendeten Variablen, sondern enthält meist weitere *Hilfsvariablen*, die einen Erklärungswert für die fehlenden Werte haben können. Jeder der imputierten, nun vollständigen Datensätze, wird separat analysiert. Die Ergebnisse werden anschließend unter Berücksichtigung der Unsicherheit durch die Imputation der Werte zusammengeführt, die Standardfehler also entsprechend angepasst. Studien aus dem medizinische Bereich legen nahe, dass auch in Fällen mit MNAR Multiple Imputationen (Tan, Jolani & Verbeek, 2018) und darauf aufbauenden Sensitivitätsanalysen (Leurent et al., 2018) zuverlässigere Schätzungen liefern als Maximum-Likelihood Schätzungen. Multiple Imputationen wurden im Rahmen des Projektes und dieser Arbeit jedoch nicht durchgeführt.

Negative Varianzschätzung des Lernfortschritts in vier Modellen. Negative Fehlervarianzen können Hinweis auf Fehlspezifikation des Modells sein. In der vorliegenden Arbeit wurden Bootstrap Konfidenzintervalle um die fraglichen Varianzen erzeugt. Nach Kolenikov und Bollen (2012) ist dies eine Möglichkeit zu prüfen, ob negativ geschätzte Varianzen auf eine Fehlspezifikation des Modells hinweisen. In allen Modellen, in denen in dieser Arbeit negative Varianzen geschätzt wurden, schloss das jeweilige Konfidenzintervall die Null ein. Dies stellt nach den Autoren eine notwendige, jedoch keine hinreichende Bedingung dar, um Fehlspezifikationen auszuschließen.

In den drei Modellen des NagP-Tests wird der Lernfortschritt mit einem standardisierten Regressionsgewicht von über 1 durch die Ausgangswerte vorhergesagt. In den Alternativmodellen 2 und 3, in denen der hierarchischen Datenstruktur mit einer post-hoc Standardfehlerkorrektur Rechnung getragen wird, sind die standardisierten Regressionskoeffizienten kleiner 1.[37] Im sNCM-Test wird die Varianz des Lernfortschritts nur in Modell 2 negativ geschätzt. In diesem Modell werden die Lernfortschritte durch zwei Indikatoren hochschulischer Lerngelegenheiten gleichzeitig vorhergesagt. Auch hier wird im Alternativmodell 2 der standardisierte autoregressive Effekt kleiner 1.

[37] Ein Alternativmodell 1, ohne feste Effekte und mit post-hoc Standardfehlerkorrektur führt ebenfalls zu einem standardisierten autoregressiven Effekt kleiner 1 für NagP, $\beta = .97$, $SE(\beta) = .13$, $p < .001$.

Dieses Muster kann durch Multikollinearität der Prädiktoren hervorgerufen werden. Insbesondere bei gleichzeitiger Verwendung zweier Prädiktoren für hochschulische Lerngelegenheiten ist dies zu erwarten. Der *Variance Inflation Factor* (VIF; VIF = der Kehrwert der Toleranz; Toleranz = 1 abzüglich der erklärten Varianz in einem Prädiktor durch alle anderen Prädiktoren im Modell) ist für alle relevanten Prädiktoren kleiner 4 (siehe Anhang O; Hair, 2010). Deshalb sollten die vorgestellten Ergebnisse nicht nennenswert durch Multikollinearität der Prädiktoren verzerrt sein.

Analyse auf Päckchenebene. Aufgrund der Vielzahl an Items in den deklarativen Testteilen und der geringen Kovarianzabdeckung im sNCM-Test wurden die Analysen für Testwertinterpretation 3 mit Itempäckchen analysiert. Die Verwendung von Päckchen wird dafür kritisiert, mögliche Modellfehlspezifikationen zu verschleiern und instabile Ergebnisse in Abhängigkeit von der Zusammenstellungen der Päckchen zu produzieren (z.B. Marsh, Lüdtke, Nagengast, Morin & Davier, 2013). Fehlspezifikationen sind besonders bei mehrdimensionalen Modellen möglich, bei eindimensionalen Faktoren ist diese Gefahr jedoch geringer (Bandalos & Finney, 2001). Da die vier Tests als eindimensionale Skalen konstruiert wurden und deren Eindimensionalität nach Anpassungen angenommen werden kann (Seeber et al., 2019), sollte dieser Kritikpunkt für die verwendeten Daten nicht zutreffen.

Dem stehen jedoch Vorteile der Analyse auf Päckchenebene gegenüber, wenn komplexe Modelle oder Daten mit einem ungünstigen Verhältnis von Variablen zu Personen analysiert werden (Little, Cunningham, Shahar & Widaman, 2002). Im Ko-NaMa-Projekt trifft sowohl das ungünstige Verhältnis von Variablen zu Personen auf, als auch eine komplexe Datenstruktur. Die Studierenden wurden nicht zufällig gezogen und in unterschiedliche Lehrveranstaltungen verteilt.

Für Little, Rhemtulla, Gibson und Schoemann (2013) sind Analysen auf Päckchenebene dann unangemessen, wenn es um Fragen der inneren Struktur eines Konstrukts geht. Unter Berücksichtigung der Konstruktrepräsentativität der Päckchen, können Päckchen aber hilfreich sein, wenn Zusammenhänge des Konstrukts zu anderen Variablen analysiert werden. Das trifft auf die vorgestellten Analysen zu. Es geht nicht um die innere Struktur der Tests, sondern um den Zusammenhang von Testwerten zu unterschiedlichen Indikatoren von Lerngelegenheiten. Aufgrund der Vielzahl an Items in den Tests NagP und dNCM konnten die Päckchen domänenrepräsentativ gebildet werden. Für den sNCM-Test bilden die Päckchen nicht jeweils alle Unternehmensbereiche ab. Hier wurde jedoch darauf geachtet, die verschiedenen

Unternehmensprozesse (Produktions-, Support- und Managementprozess) in jedem Päckchen abzubilden.

Insgesamt werden die Vorteile einer Analyse auf Päckchenebene für Evidenzen zur Testwertinterpretation 3 als überwiegend eingeschätzt. Nichtsdestotrotz wäre eine Analyse auf Itemebene wünschenswert, die eine bessere Datenlage voraussetzt. Durch das Testheftdesign liegen für den sNCM-Test die wenigsten Antworten pro Item vor.[38] Viele Items werden zudem mehrstufig bepunktet. Dadurch ist die Kovarianzabdeckung geringer als in den deklarativen Tests. Gleichzeitig weist dieser Test ein komplexes Design auf. Während das einfaktorielle Modell eher schlecht auf die Daten passt (Modell 1), könnte z.B. eine hierarchische Modellierung von Items, geclustert in Situationen, zu einem besseren Modellfit führen. Die Stichprobengröße schränkt hier weitere Analysemöglichkeiten ein.

Partielle Messinvarianz in zwei der Tests. Für zwei der verwendeten Tests konnte nur partielle skalare Messinvarianz über Messzeitpunkte festgestellt werden. Wie im Methodenteil in Kapitel 8.2.3.1 beschrieben, wird für den NagP-Test keine Einschränkung der Konstruktbedeutung erwartet. Hier sind noch zwei von drei Päckchen mit einer Gesamtzahl von 29 Items weiterhin über Messzeitpunkte invariant. Beim partiell skalaren sNCM-Test ist die Konstruktbedeutung jedoch eingeschränkt. Nachhaltigkeitsmanagement bedeutet, soziale und ökologische Faktoren in jegliche betriebliche Entscheidungsfindung zu integrieren. Die Items sind in Situationen geclustert, von denen jede in einer Unternehmensabteilung spielt. Drei bis vier Situationen wurden zu jeweils einem Indikatorpäckchen zusammengefasst. Daher sind abteilungsspezifische Inhalte und Methoden meist nur durch ein Päckchen abgedeckt. Im sNCM-Test ist ein Päckchen nicht invariant. Dadurch sind zwei Unternehmensabteilungen und spezifische Methoden aus diesen nicht längsschnittlich vergleichbar. Dies stellt für Testwertinterpretation 3, in der die Testwerte als Indikatoren hochschulisch vermittelter Kompetenzen validiert werden, eine geringere Einschränkung dar. Sind jedoch Analysen zum Zusammenhang zu beruflich relevanten Handlungen geplant (vgl. Testwertinterpretation 2 in Kapitel 7), sollte die Bedeutung der partiellen Invarianz im sNCM-Test genauer untersucht werden.

[38] Während in den deklarativen Tests zwischen 249 und 450 Antworten pro Item vorliegen, sind es für den sNCM-Test zwischen 157 und 274 Items (vgl. Kapitel 6.2).

8.3 Plausibilität der Testwertinterpretation 3

Grundannahme 1

Evidenzen für Grundannahme 1 betreffen nur die Tests NagP, dNCM und sNCM. Für den BWL-Test wurden keine Unterschiede in Lerngelegenheiten zwischen den Lehrveranstaltungen erwartet (siehe Kapitel 8.2.1).

Bisher liegen Evidenzen zu Grundannahme 1 nur in eingeschränktem Umfang vor. Die Analyse von Modulhandbüchern und die Einschätzung der bei der Datenerhebung anwesenden Lehrpersonen sind Teil der Dissertation eines Kollegen im Ko-NaMa-Projekt und wurden noch nicht veröffentlicht. Für Grundannahme 1 können daher nur Evidenzen berücksichtigt werden, die sich auf die Abstimmung von Testinhalten und Lehrinhalten aus Studierendensicht stützen. Diese stützen die Grundannahme, dass Inhalte aus Lehrveranstaltungen mit Lerngelegenheiten zu Nachhaltigkeitsthemen mit Testinhalten übereinstimmen. Studierende aus Lehrveranstaltungen, die vom Ko-NaMa-Projekt als Schwerpunktgruppen klassifiziert wurden, berichten zum zweiten Messzeitpunkt mehr hochschulische Lerngelegenheiten als Studierende der Kontrollgruppe.

Grundannahme 2

Evidenzen für Grundannahme 2, dass die Tests veränderungssensitiv sind, können nur aus den Ergebnissen zu Grundannahme 3 abgeleitet werden. Da sich in den Tests NagP und dNCM die erwarteten Lernfortschritte in der Schwerpunktgruppe zeigen, ist für diese Tests Instruktionssensitivität anzunehmen. Angesichts des fehlenden Lernfortschritts im BWL-Test kann keine stützende Evidenz zur Instruktionssensitivität des Tests festgestellt werden. Gleiches gilt für den sNCM-Test. Dieser scheint zwar sensitiv für Veränderungen zu sein, allerdings bleibt der erwartete positive Zusammenhang aus.

Damit liegt bei den Tests BWL und sNCM das grundsätzliche Problem der Analyse von Instruktionssensitivität vor. Wenn die erwarteten Lernfortschritte ausbleiben ist unklar, ob die Tests nicht sensitiv sind oder ob die Instruktion nicht wirksam war (Naumann, Hochweber & Klieme, 2016). Im BWL-Test trifft vermutlich zweiter Fall zu. Die verwendete Stichprobe studiert in wirtschaftswissenschaftlichen Studiengängen mindestens im vierten Bachelorsemester. Die Testinhalte erfassen grundständiges betriebswirtschaftliches Wissen. Daher ist fraglich, ob in der erreichten Stichprobe (ab dem 4. Semester im Bachelorstudium) überhaupt nennenswerte Lernfortschritte erwartet werden können. Hochschulische Lerngelegenheiten zu im BWL-Test

abgefragten Wissen werden vermutlich früher in grundständigen wirtschaftswissenschaftlichen Studiengängen angeboten. Zur Analyse von Instruktionssensitivität im BWL-Test könnten Studierende zu Beginn des Bachelorstudiums eine geeignetere Stichprobe darstellen.

Für den sNCM-Test besteht jedoch das oben beschriebene Dilemma. Unter der Annahme, dass der Test instruktionssensitiv ist, könnten die Lerngelegenheiten innerhalb eines Semesters nicht ausreichend gewesen sein, um vorherrschende Dogmen in wirtschaftswissenschaftlichen Studiengängen zu ändern, und auch soziale und ökologische Aspekte in betrieblichen Entscheidungsfindungen zu berücksichtigen. Im anderen Fall waren die Instruktionen wirksam, nur der Test war nicht in der Lage die Veränderungen abzubilden. Dagegen spricht jedoch, dass selbstberichtete hochschulische Lerngelegenheiten Lernfortschritt negativ vorhersagen. Der Test scheint also durchaus sensitiv für Veränderungen zu sein. Insgesamt wird dies jedoch als schwache Evidenz für die Instruktionssensitivität des sNCM-Tests gewertet.

Grundannahme 3

Grundannahme 3, dass Kompetenz in Nachhaltigkeitsmanagement vorwiegend hochschulisch vermittelt wird, wird für die Tests NagP und dNCM als plausibel eingestuft. Lernfortschritte sind in diesen Tests in der Schwerpunktgruppe größer als in der Kontrollgruppe (Modell 1). Die Evidenzen aus Modell 2 liefern gemischte Evidenz. Diese sind jedoch nur als Effekte innerhalb der Lehrveranstaltungen zu interpretieren, unter Kontrolle der Lehrveranstaltungen. Selbstberichtete außerhochschulische Lerngelegenheiten haben keine Vorhersagekraft für den Lernfortschritt (Modell 3), was wiederum für Nachhaltigkeitsmanagement als vorwiegend hochschulisch vermittelte Kompetenz spricht.

Für den sNCM-Test widersprechen die präsentierten Evidenzen der Grundannahme 3. Im Test zeigt sich kein Unterschied im Lernfortschritt in Abhängigkeit der hochschulischen Lerngelegenheiten (Modell 1) und die selbstberichteten hochschulischen Lerngelegenheiten sagen den Lernfortschritt negativ vorher (Modell 2). Einzig der nicht signifikante Zusammenhang von außerhochschulischen Lerngelegenheiten zu Lernfortschritt kann als stützende Evidenz der Grundannahme betrachtet werden.

Fazit zur Plausibilität der Testwertinterpretation

Die präsentierten Evidenzen werden für die Tests NagP und dNCM als überwiegend stützend interpretiert. Damit wird die Interpretation, dass Testwerte hochschulisch vermitteltes Wissen spiegeln, vorläufig als plausibel eingestuft. Für den sNCM-Test stützen die Evidenzen

Grundannahme 1 der Testwertinterpretation. Für Grundannahme 3 gibt es nur schwache Evidenzen. Zwar haben die außerhochschulischen Lerngelegenheiten keinen Erklärungswert für die Testwerte im sNCM-Test. Es fehlt aber jeder Nachweis, dass hochschulische Lerngelegenheiten positiven Lernfortschritt vorhersagen. Damit fehlen auch stützende Evidenzen zur Grundannahme 2. Insgesamt wird die Testwertinterpretation damit als wenig plausibel betrachtet. Für den BWL-Test wird die Evidenz zu Grundannahme 3 als stützend interpretiert. Damit ist jedoch keine stützende Evidenz für Grundannahme 2 zu finden. Insgesamt wird daher auch für diesen Test die Testwertinterpretation als wenig plausibel betrachtet.

Für die Verwendung der Tests in zukünftigen Anwendungen sollte neben dem Fazit zur Plausibilität der Testwertinterpretation die mangelnde Generalisierbarkeit der Ergebnisse berücksichtigt werden. Das verwendete Feste-Effekte Modell lässt keine Generalisierung der Ergebnisse zu. Die für die Testwerte gezogenen Interpretationen gelten damit nur für die an der Datenerhebung beteiligten Standorte. Da es bis zum Sommersemester 2018 nur wenige Lehrveranstaltungen im deutschen Hochschulsystem mit Bezug zu Nachhaltigkeitsthemen gab (Seeber et al., 2019), wurde bei der Wahl des Analysemodells der Generalisierbarkeit weniger Bedeutung beigemessen. Falls die Tests außerhalb dieser Stichprobe zur Erfassung von hochschulisch vermittelter Kompetenz in Nachhaltigkeitsmanagement eingesetzt werden sollen, sind zusätzliche Evidenzen zu liefern, welche die Generalisierbarkeit der hier vorgestellten Ergebnisse stützt.

In Abbildung 20 werden die Validitätsevidenzen in die von AERA et al. (2014) vorgeschlagenen Evidenzquellen eingeordnet.

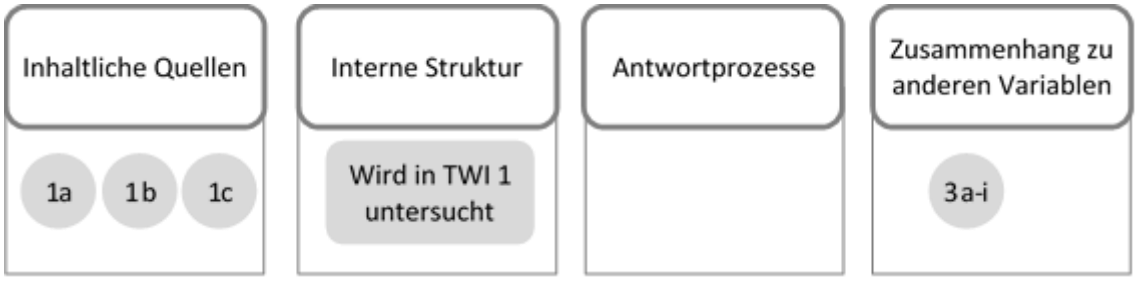


Abbildung 20 Quellen der Validitätsevidenzen für Testwertinterpretation 3

Gesamtdiskussion

9 Gesamtdiskussion

Die Ziele dieser Dissertation waren 1) die Entwicklung eines Schemas zur Einordnung von Testnutzen bei Tests zur Erfassung studentischer Kompetenzen und 2) das Schema auf empirischen Daten anzuwenden. In Kapitel 4.5 wurde bereits die Entwicklung des Schemas diskutiert. In den Kapiteln 6 bis 8 wurden jeweils die Evidenzen für Grundannahmen der jeweiligen Testwertinterpretation diskutiert und ein Fazit zur Plausibilität der Testwertinterpretation gezogen. In diesem Kapitel soll nun die Anwendbarkeit des Schemas als kapitelübergreifender Aspekt der Dissertation diskutiert werden. Abschließend wird auch die Nutzung des argumentationsbasierten Validierungsansatzes im Bereich der studentischen Kompetenzerfassung diskutiert.

9.1 Anwendbarkeit des Schemas

Ausgangspunkt dieser Arbeit waren die häufig nicht auf den Einsatzzweck abgestimmten Validierungskonzepte von Kompetenzerfassungen bei Studierenden (Kuhn et al., 2016). Im ersten Teil dieser Arbeit wurden typische Einsatzzwecke bei der Erfassung studentischer Kompetenzen klassifiziert und ein Validierungsschema entwickelt, das sich auf den Zusammenhang von Test zu Konstrukt, Lehre und berufliche Anforderungen stützt. In dieses Schema können Testwertinterpretationen eingeordnet werden und Validitätsevidenzen abgeleitet werden. Im zweiten Teil der Arbeit wurde das Validierungsschema in einem Forschungsprojekt zur Erfassung studentischer Kompetenzen in Nachhaltigkeitsmanagement angewandt. Die drei Testwertinterpretationen beziehen sich auf den Zusammenhang von 1) Test zu Konstrukt, 2) Test zu beruflichen Anforderungen und 3) Test zu Lehre. Die Testwertinterpretationen können jedoch nicht ausschließlich einem Zusammenhang zugeordnet werden und bauen aufeinander auf. In dieser Arbeit wurden zunächst die Testwerte als plausible Indikatoren der Kompetenz Nachhaltigkeitsmanagement validiert (siehe Kapitel 6). Diese Interpretation stellt eine Bedingung für die Interpretation dar, dass die Testwerte Indikatoren für *hochschulisch vermittelte* Nachhaltigkeitsmanagementkompetenz sind (siehe Kapitel 8).Wie in Kapitel 4 diskutiert, kann die erste Testwertinterpretation als *Konstruktvalidität* verstanden werden, die in vielen der in Kapitel 4 analysierten Texte untersucht wird.

Mögliche Hindernisse bei der Nutzung des argumentationsbasierten Validierungsansatzes

In Kapitel 4 fiel auf, dass sich nur wenige der identifizierten Literaturen auf einen argumentationsbasierten Validierungsansatz bezogen. In den Literaturen, die explizit solch einen Ansatz verfolgten, wurde nicht immer eine explizite Testwertinterpretation formuliert. Woran die geringe Nutzung des Ansatzes liegen könnte, wurde bereits in Kapitel 4 vermutet. Zunächst werden in dieser Dissertation nur Literaturen bis Juli 2017 berücksichtigt. Bei Publikationen aus bereits laufenden oder schon abgeschlossenen Projekten könnten Autor*innen nicht mehr auf aktualisierte Empfehlungen zum Vorgehen bei Validierungen reagiert haben. Allerdings wird die Entwicklung weg von einer *Validität des Tests* hin zu Validierung von Testnutzen schon Jahrzehnte zuvor diskutiert, z.B. bei Messick (1989). Auch in früheren Versionen der *Standards* werden Empfehlungen für die Validierung von Testnutzen gegeben (AERA, APA & NCME, 1999).

Das Verständnis von Validität als Eigenschaften eines Tests findet sich auch in neueren Literaturen. In diesen wird ein Test validiert, um anschließend Forschungsfragen zu untersuchen. Wie in Kapitel 4.5 diskutiert, kann auch die Validierung einer Testwertinterpretation als Forschungsprozess betrachtet werden. Ob widersprüchliche Evidenzen gegen die Plausibilität einer Testwertinterpretation sprechen oder die den Annahmen zugrundeliegenden Theorien überprüft werden sollten, hängt dann von einer Bewertung der Vertrauenswürdigkeit der Theorie ab (bzw. von der Stärke des nomologischen Netzwerks, Cronbach & Meehl, 1955).

In dieser Arbeit liegt ein Beispiel für den zweiten Fall vor. Wie in Kapitel 5 erläutert, lagen bislang keine Instrumente zur Erfassung dieser Kompetenz bei Studierenden vor. Die Validitätsargumente können sich daher in großen Teilen nicht auf empirisch überprüfte Theorien stützen. Vielmehr können die hier (und in den zitierten Artikeln) präsentierten Ergebnisse genutzt werden, um die theoretischen Vorstellungen zur Kompetenzstruktur und –entwicklung anzupassen. In Kapitel 8.2.3.3 etwa wurde diskutiert, dass deklaratives Wissen in BWL, Nachhaltigkeit aus gesamtgesellschaftlicher Perspektive und Nachhaltigkeitsmanagement eine Grundlage für Kompetenz in Nachhaltigkeitsmanagement bilden. Ob ökologische und soziale Aspekte aber dauerhaft in betrieblichen Entscheidungen berücksichtigt werden, wie sie in der simulativ umgesetzten Testkomponente erfasst werden, könnte auch durch Einstellungen zu Nachhaltigkeitsthemen beeinflusst sein. Werden diese Annahmen durch zukünftige

Forschungsergebnisse bestätigt, müssten diese Erkenntnisse bei zukünftigen Testungen berücksichtigt werden.

Eine weitere Herausforderung des Validierungsansatzes besteht darin, für jede Testwertinterpretation ein eigenes Validitätsargument zu formulieren. Für einen Test können so mehrfache Validierungsstudien notwendig sein, was einen zusätzlichen Aufwand bedeutet. Die Validitätsargumente bestehen darüber hinaus nicht dauerhaft, sondern müssen überprüft und ggf. angepasst werden, wenn neue Forschungsergebnisse oder gesellschaftliche Rahmenbedingungen dies erfordern.

Hervorzuheben ist, dass für die Anwendung des in dieser Arbeit verwendeten Verständnisses von Validität keine neuen psychometrischen Methoden benötigt werden. In den Kapiteln 6 bis 8 werden für die Validierung der einzelnen Testwertinterpretationen *t*-Tests, konfirmatorische Faktorenanalysen und Strukturgleichungsmodelle verwendet; in Vorarbeiten wurden ebenfalls IRT-Analysen durchgeführt. Die mangelnde Anwendung des argumentativen Validierungsansatzes sollte also nicht daran liegen, dass neue, noch wenig in Software implementierte Methoden angewandt werden müssen.

9.2 Diskussion des argumentativen Validierungsansatzes

Der in dieser Arbeit verwendete argumentationsbasierte Validierungsansatz legt den Fokus auf die Validierung einer Testwertinterpretation, die für einen bestimmten Testnutzen gilt. In Kapitel 4 wurde festgestellt, dass in bisherigen Testentwicklungen zur Erfassung studentischer Kompetenzen der Schwerpunkt auf der Erforschung eines bestimmten Konstruktes liegt. Das Ziel ist ein besseres Verständnis, um anschließend z.B. Empfehlungen für die Lehre (in Lehramtsstudiengängen) geben zu können. Während in den analysierten Literaturen der Testnutzen eher in der Forschung anzusiedeln ist, wird in den *Standards* der *Test Use* als ein Anwendungsfall eines Tests verstanden, der mit mehr oder minder einflussreichen Konsequenzen für Betroffene verbunden ist. Der Konsequenz von Testeinsätzen, welchen in den *Standards* ein eigenes Kapitel gewidmet wird, spielt in den untersuchten Textstellen bezeichnenderweise keine Rolle.

Diese sollten jedoch untersucht werden, wenn die entwickelten Tests zu anderen Zwecken als zum Erkenntnisgewinn eingesetzt werden (wie in Kapitel 4 beschrieben, werden in mehreren

Texten Ausblicke auf mögliche Einsatzzwecke von Tests gegeben). Die Konsequenzen von Testwerten sollten z.B. untersucht werden, wenn Studierende anhand ihrer Testwerte auf Treatments zugewiesen werden oder sie Rückmeldung zu ihrem aktuellen Könnensstand erhalten. In erstem Fall ist denkbar, dass einige Studierende Nachteile durch die Zuweisung erfahren (vgl. Diskussion zu Einteilung auf Sprachniveaus in Kapitel 3.2.1). In letzterem Fall sollte nicht nur überprüft werden, ob Studierende durchschnittlich von der Rückmeldung profitieren und Lücken schließen können. Vielmehr wäre wichtig zu untersuchen, ob Gruppen von Studierenden, etwa in einem bestimmten Kompetenzbereich, nach einer solchen Rückmeldung eine negative Entwicklung aufweisen.

Literaturverzeichnis

AERA; APA; NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

AERA; APA; NCME. (2014). *Standards for Educational and Psychological Testing*. Washington D.C.: American Educational Research Association.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, *51*(2, Pt.2), 1–38. https://doi.org/10.1037/h0053479

Ames, A. J. & Penfield, R. D. (2015). An NCME Instructional Module on Item-Fit Statistics for Item Response Theory Models. *Educational Measurement: Issues and Practice*, *34*(3), 39–48.

Anderson, L. W. (2002). Curricular Alignment: A Re-Examination. *Theory into Practice*, *41*(4), 255–260.

Aryadoust, V., Ng, L. Y. & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, *38*(1), 6–40.

Bandalos, D. J. & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides & R. E. Schumacker (Hrsg.), *New Developments and Techniques in Structural Equation Modeling* (S. 269–296). Taylor & Francis.

Blömeke, S., Gustafsson, J.-E. & Shavelson, R. J. (2015). Beyond Dichotomies. *Zeitschrift für Psychologie*, *223*(1), 3–13. Zeitschrift für Psychologie, 223(1), 3-13. https://doi.org/10.1027/2151-2604/A000194

Blömeke, S., Jenßen, L., Grassmann, M., Dunekacke, S. & Wedekind, H. (2017). Process mediates structure: The relation between preschool teacher education and preschool teachers' knowledge. *Journal of Educational Psychology*, *109*(3), 338–354. https://doi.org/10.1037/edu0000147

Braun, E., Gusy, B., Leidner, B. & Hannover, B. (2008). Das Berliner Evaluationsinstrument für selbsteingeschätzte, studentische Kompetenzen (BEvaKomp). *Diagnostica*, *54*(1), 30–42. https://doi.org/10.1026/0012-1924.54.1.30

Brückner, S., Förster, M., Zlatkin-Troitschanskaia, O., Happ, R., Walstad, W. B., Yamaoka, M. et al. (2015). Gender effects in assessment of economic knowledge and understanding: Differences among undergraduate business and economics students in Germany, Japan,

and the United States. *Peabody Journal of Education*, *90*(4), 503–518. https://doi.org/10.1080/0161956X.2015.1068079

Brückner, S. & Pellegrino, J. W. (2016). Integrating the analysis of mental operations into multilevel models to validate an assessment of higher education students' competency in business and economics. *Journal of Educational Measurement*, *53*(3), 293–312.

Buchholtz, N. (2014). *Multiperspektivische Ansätze zur Messung des Lehrerprofessionswissens in der Mathematiklehramtsausbildung*. Hamburg, Universität Hamburg, Diss., 2014. Verfügbar unter: http://nbn-resolving.de/urn:nbn:de:gbv:18-65839; http://d-nb.info/1047440245/34; http://ediss.sub.uni-hamburg.de/volltexte/2014/6583/

Bundesministerium für Gesundheit. Approbationsordnung für Ärzte vom 27. Juni 2002 (BGBl. I S. 2405), die zuletzt durch Artikel 3 des Gesetzes vom 16. März 2020 (BGBl. I S. 497) geändert worden ist. ÄApprO 2002. Zugriff am 15.10.2020. Verfügbar unter https://www.bundesaerztekammer.de/recht/gesetze-und-verordnungen/approbationsordnung/

BVerfG, Urteil des Ersten Senats (19.12.2017) 1 BvL 3/14, Rn. 1-253. *BverfGE 147, 253 - 364*.

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Chen, W.-H. & Thissen, D. (1997). Local Dependence Indexes for Item Pairs Using Item Response Theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. https://doi.org/10.3102/10769986022003265

Christensen, K. B., Makransky, G. & Horton, M. (2017). Critical Values for Yen's Q3: Identification of Local Dependence in the Rasch Model Using Residual Correlations. *Applied Psychological Measurement*, *41*(3), 178–194. https://doi.org/10.1177/0146621616677520

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Cronbach, L. J. (1971). Test Validation. In R. Thorndike (Hrsg.), *Educational Measurement* (2. Aufl.). Washington D.C.: American Council on Education.

Cronbach, L. J. (1986). Five Perspectives on the Validity Argument. In Wainer Howard & Braun, Henry, I. (Hrsg.), *Test Validity* (S. 4–18).

Cronbach, L. J. & Meehl, P. M. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin*, *52*(4), 281–302.

DeMars, C. E. (2010). *Item response theory* (Series in understanding statistics. Measurement). Oxford, New York: Oxford University Press.

DIN (2016). *DIN 33430: Anforderungen an berufsbezogene Eignungsdiagnostik*. Berlin: Beuth.

Dorsch, C. (2018). Reflecting on the Smart City: How Student Teachers Learn to Teach Smart Pupils. *GI_Forum*, *6*(2), 186-180.

Elkington, J. (1998). Accounting for the Triple Bottom Line. *Measuring Business Excellence*, *2*(3), 18–22.

Frey, A., Hartig, J. & Rupp, A. A. (2009). An NCME Instructional Module on Booklet Designs in Large-Scale Assessments of Student Achievement: Theory and Practice. *Educational Measurement: Issues and Practice*, *28*(3), 39–53. https://doi.org/10.1111/j.1745-3992.2009.00154.x

Frey, A., Spoden, C. & Born, S. (2020). Construction of Psychometrically Sound Written University Exams. *Psychological Test and Assessment Modeling*, *65*(4), 415–525.

Frey, A., Spoden, C., Fink, A. & Born, S. (2020). Kompetenzorientierte individualisierte Hochschulklausuren und deren prüfungsrechtliche Einordnung. *eleed*, *13*(1). Verfügbar unter: https://eleed.campussource.de/archive/13/5119

Gäde, J. C., Schermelleh-Engel, K. & Werner, C. S. (2020). Klassische Methoden der Reliabilitätsschätzung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3rd ed. 2020, 305-334). Berlin, Heidelberg: Springer.

Gramzow, Y. (2015). *Fachdidaktisches Wissen von Lehramtsstudierenden im Fach Physik. Modellierung und Testkonstruktion* (Studien zum Physik- und Chemielernen, Bd. 181). Berlin: Logos Verl.

Hair, J. F. (2010). *Multivariate data analysis. A global perspective* (7. ed., global ed.). Upper Saddle River, NJ: Pearson.

Hammer, S., Carlson, S. A., Ehmke, T., Koch-Priewe, B., Köker, A., Ohm, U. et al. (2015). Kompetenz von Lehramtsstudierenden in Deutsch als Zweitsprache. Validierung des GSL-Testinstruments. In S. Blömeke & O. Zlatkin-Troitschanskaia (Hrsg.), *Kompetenzen von Studierenden* (Zeitschrift für Pädagogik : Beiheft, Bd. 61, S. 32–54). Weinheim: Beltz Juventa.

Hartig, J., Frey, A. & Jude, N. (2020). Validität von Testwertinterpretationen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3rd ed. 2020, S. 529–545). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-61532-4_21

Hartig, J. & Klieme, E. (2006). Kompetenz und Kompetenzdiagnostik. In K. Schweizer (Hrsg.), *Leistung und Leistungsdiagnostik* (S. 127–143). Heidelberg: Springer Medizin.

Hochschulrahmengesetz (HRG). Zugriff am 02.03.2021. Verfügbar unter: https://www.gesetze-im-internet.de/hrg/HRG.pdf

Hochschulrektorenkonferenz. (2020, 1. September). *Statistische Daten zu Studienangeboten an Hochschulen in Deutschland. Studiengänge, Studierende, Absolventinnen und Absolventen*. Wintersemester 2020/21 (Statistiken zur Hochschulpolitik 01/2020). Zugriff am 04.05.2021. Verfügbar unter: https://www.hrk.de/fileadmin/redaktion/hrk/02-Dokumente/02-03-Studium/02-03-01-Studium-Studienreform/HRK_Statistik_BA_MA_UEbrige_WiSe_2020_21_finale.pdf

Hochschulrektorenkonferenz & ZEIT ONLINE (Hrsg.).. *Studien-Interessentest (SIT)*. Zugriff am 08.02.2021. Verfügbar unter: https://www.hochschulkompass.de/studium/hilfe-bei-der-studienwahl/tests-zur-studienorientierung.html

Hopkins, P., Hughes, P. & Layer, G. (2008). Sustainable graduates: linking formal, informal and campus curricula to embed education for sustainable development in the student learning experience. *Environmental Education Research*, *14*(4), 435–454. https://doi.org/10.1080/13504620802283100

Hu, L.-t. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis. Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

Jahn, G. K. (2014). *Studien zur Überprüfung der Validität eines Instruments zur Erfassung professioneller Unterrichtswahrnehmung von Lehramtsstudierenden*. München, Technische Universität München, Diss., 2014.

Jähnig, C. C. (2014). *Die Messung betriebswirtschaftlichen Wissens von Studierenden. Eine quantitativ-empirische Untersuchung situativer Testaufgaben* (Berufsbildung, Arbeit und Innovation -Dissertationen/Habilitationen, Band 28). Bielefeld: wbv. https://doi.org/10.3278/6004416w

Kane, M. T. (1992). An Argument-Based Approach to Validity. *Psychological Bulletin*, *112*(3), 527-525.

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

Kang, T. & Cohen, A. S. (2007). IRT Model Selection Methods for Dichotomous Items. *Applied Psychological Measurement*, *31*(4), 331–358. https://doi.org/10.1177/0146621606292213

Kang, T., Cohen, A. S. & Sung, H.-J. (2009). Model Selection Indices for Polytomous Items. *Applied Psychological Measurement*, *33*(7), 499–518.

Kemmerer, A. (2019). Konzeption videobasierter und forschungsorientierter Lerneinheiten zur Förderung der Diagnosekompetenz von Lehramtsstudierenden im Fach Englisch. In A. Kreft & M. Hasenzahl (Hrsg.), *Aktuelle Tendenzen in der Fremdsprachendidaktik. Zwischen Professionalisierung, Lernerorientierung und Kompetenzerwerb* (Kolloquium Fremdsprachenunterricht, Bd. 64). Berlin: Peter Lang.

Kirschner, S. (2013). *Modellierung und Analyse des Professionswissens von Physiklehrkräften*. Duisburg, Essen, Univ. Duisburg-Essen, Diss., 2013. Verfügbar unter: http://nbn-resolving.de/urn:nbn:de:hbz:464-20131210-150745-4; http://d-nb.info/1045839434/34; http://duepublico.uni-duisburg-essen.de/servlets/DocumentServlet?id=32764

Klieme, E. & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. *Zeitschrift für Pädagogik*, *52*(6), 876–903.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (Methodology in the social sciences, 3rd ed.). New York: Guilford Press.

Köhler, C., Hartig, J. & Schmid, C. (2020). Deciding between the Covariance Analytical Approach and the Change-Score Approach in Two Wave Panel Data. *Multivariate Behavioral Research*, 1–12.

Kolenikov, S. & Bollen, K. A. (2012). Testing Negative Error Variances. Is a Heywood Case a Symptom of Misspecification? *Sociological Methods & Research*, *41*(1), 124–167. https://doi.org/10.1177/0049124112442138

König, J. & Herzmann, P. (2011). Lernvoraussetzungen angehender Lehrkräfte am Anfang ihrer Ausbildung. Erste Ergebnisse aus der wissenschaftlichen Begleitung des Kölner Modellkollegs Bildungswissenschaften. Paralleltitel: Learning preconditions of future teachers at the beginning of pre-service teacher education. *Lehrerbildung auf dem Prüfstand*, *4*(2), 186–210.

Kuckartz, U. (2018). *Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung* (Grundlagentexte Methoden, 4. Auflage). Weinheim, Basel: BeltzJuventa.

Kuhn, C. (2014). *Fachdidaktisches Wissen von Lehrkräften im kaufmännisch-verwaltenden Bereich. Modellbasierte Testentwicklung und Validierung* (Empirische Berufsbildungs- und Hochschulforschung, Bd. 2). Landau: Verlag Empirische Pädagogik.

Kuhn, C., Happ, R., Zlatkin-Troitschanskaia, O., Beck, K., Förster, M. & Preuße, D. (2014). Kompetenzentwicklung angehender Lehrkräfte im kaufmännisch-verwaltenden Bereich –

Erfassung und Zusammenhänge von Fachwissen und fachdidaktischem Wissen. *Zeitschrift für Erziehungswissenschaft*, *17*(S1), 149–167. https://doi.org/10.1007/s11618-013-0456-3

Kuhn, C., Zlatkin-Troitschanskaia, O., Pant, H. A. [Hans Anand] & Hannover, B. (2016). Valide Erfassung der Kompetenzen von Studierenden in der Hochschulbildung. *Zeitschrift für Erziehungswissenschaft*, *19*(2), 275–298.

Kunina-Habenicht, O., Wilhelm, O., Matthes, F. & Rupp, A. A. (2010). Kognitive Diagnosemodelle: Theoretisches Potential und methodische Probleme. Projekt Kognitive Diagnosemodelle. *Zeitschrift für Pädagogik*, *56*(Beiheft 56), 75–85.

La Porta, F., Franceschini, M., Caselli, S., Cavallini, P., Susassi, S. & Tennant, A. (2011). Unified Balance Scale: an activity-based, bed to community, and aetiology-independent measure of balance calibrated with Rasch analysis. *Journal of Rehabilitation Medicine*, *43*(5), 435–444.

Lauterbach, O. (2015). Erfassung wirtschaftswissenschaftlicher Fachkompetenzen von Studierenden in Startkohorte 5 des nationalen Bildungspanels - Technischer Bericht. *NEPS Working Papers*, (51).

Lehmann, G. (Hrsg.). (2014). *Green Controlling. Leitfaden für die erfolgreiche Integration ökologischer Zielsetzungen in Unternehmensplanung und -steuerung*. Freiburg: Haufe Gruppe.

Leurent, B., Gomes, M., Faria, R., Morris, S., Grieve, R. & Carpenter, J. R. (2018). Sensitivity Analysis for Not-at-Random Missing Data in Trial-Based Cost-Effectiveness Analysis: A Tutorial. *PharmacoEconomics*, *36*(8), 889–901. https://doi.org/10.1007/s40273-018-0650-5

Linninger, C., Kunina-Habenicht, O., Emmenlauer, S., Dicke, T., Schulze-Stocker, F., Leutner, D. et al. (2015). Assessing Teachers' Educational Knowledge. Construct Specification and Validation Using Mixed Methods. *Zeitschrift für Entwicklungspsychologie und pädagogische Psychologie*, *47*(2), 72–83. https://doi.org/10.1026/0049-8637/a000126

Little, T. D., Cunningham, W. A., Shahar, G. & Widaman, K. F. (2002). To Parcel or Not to Parcel. Exploring the Question, Weighing the Merits. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 151–173. https://doi.org/10.1207/S15328007SEM0902_1

Little, T. D., Rhemtulla, M., Gibson, K. & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods*, *18*(3), 285–300. https://doi.org/10.1037/a0033266

Lohse-Bossenz, H., Kunina-Habenicht, O., Dicke, T., Leutner, D. & Kunter, M. (2015). Teachers' knowledge about psychology. Development and validation of a test measuring theoretical

foundations for teaching and its relation to instructional behavior. *Studies in Educational Evaluation*, *44*, 36–49. https://doi.org/10.1016/j.stueduc.2015.01.001

Lord, F. M. (1967). A Paradox in the Interpretation of Group Comparisons. *Psychological Bulletin*, *68*(5), 304–305.

Lüdtke, O., Robitzsch, A., Trautwein, U. & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung. *Psychologische Rundschau*, *58*(2), 103–117.

Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J. S. & Davier, M. von. (2013). Why item parcels are (almost) never appropriate. Two wrongs do not make a right—Camouflaging misspecification with item parcels in CFA models. *Psychological Methods*, *18*(3), 257–284. https://doi.org/10.1037/a0032773

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

Mayring, P. (2015). *Qualitative Inhaltsanalyse. Grundlagen und Techniken* (Beltz Pädagogik, 12., Neuausgabe, 12., vollständig überarbeitete und aktualisierte Aufl.). Weinheim, Bergstr: Beltz, J.

McNeish, D. & Stapleton, L. M. (2016). Modeling Clustered Data with Very Few Clusters. *Multivariate Behavioral Research*, *51*(4), 495–518.

Messick, S. (1989). Meaning and Values in Test Validation. The Science and Ethics of Assessment. *Educational Researcher*, *18*(2), 5–11.

Michaelis, C. (2017). *Kompetenzentwicklung zum nachhaltigen Wirtschaften: Eine Längsschnittstudie in der kaufmännischen Ausbildung*. Frankfurt am Main: Peter-Lang.

Michaelis, C., Aichele, C., Hartig, J., Seeber, S., Dierkes, S., Schumann, M. [Matthias] et al. (2020). Impact of Affective-Motivational Dispositions on Competence in Sustainability Management. In O. Zlatkin-Troitschanskaia, H. A. Pant, M. Toepper & C. Lautenbach (Hrsg.), *Student Learning in German Higher Education. Innovative Measurement Approaches and Research Results* (1st ed. 2020). Wiesbaden: Springer Fachmedien Wiesbaden; Imprint Springer VS.

Moosbrugger, H. & Kelava, A. (2020). Qualitätsanforderungen an Tests und Fragebogen („Gütekriterien“). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3rd ed. 2020, S. 13–38). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-61532-4_2

Naumann, A., Hochweber, J. & Klieme, E. (2016). A Psychometric Framework for the Evaluation of Instructional Sensitivity. *Educational Assessment*, *21*(2), 89–101. https://doi.org/10.1080/10627197.2016.1167591

Naumann, A., Musow, S., Aichele, C., Hochweber, J. & Hartig, J. (2019). Instruktionssensitivität von Tests und Items. *Zeitschrift für Erziehungswissenschaft*, *22*(1), 181–202. https://doi.org/10.1007/s11618-018-0832-0

OECD. (2013). *Assessment of Higher Education Learning Outcomes. Feasibility Study Report* (Volume 2 - Data Analysis and National Experiences). Zugriff am 04.03.2021. Verfügbar unter: https://www.eurashe.eu/library/modernising-phe/AHELO_Feasibility%20Study%20Report%202.pdf

Pellegrino, J. W., Chudowsky, N. & Glaser, R. (2001). *Knowing what Students Know. The science and design of educational assessment* (1st ed.). Washington, DC: National Acad. Press. https://doi.org/10.17226/10019

R Core Team. (2020). R: A language and environment for statistical computing. [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Verfügbar unter: https://www.R-project.org/

Revelle, W. (2020). psych: Procedures for Psychological, Psychometric, and Personality Research [Computer software]. Verfügbar unter: https://CRAN.R-project.org/package=psych

Riese, J., Kulgemeyer, C., Zander, S., Borowski, A., Fischer, H. E., Gramzow, Y. et al. (2015). Modellierung und Messung des Professionswissens in der Lehramtsausbildung Physik. In S. Blömeke & O. Zlatkin-Troitschanskaia (Hrsg.), Kompetenzen von Studierenden, 55–79 [Themenheft].

Robitzsch, A., Kiefer, T. & Wu, M. (2020). TAM: Test Analysis Modules. R package version 3.5-19. [Computer software]. Verfügbar unter: https://CRAN.R-project.org/package=TAM

Røe, C., Damsgård, E., Fors, T. & Anke, A. (2014). Psychometric properties of the pain stages of change questionnaire as evaluated by Rasch analysis in patients with chronic musculoskeletal pain. *BMC Musculoskeletal Disorders*, *15*(1), 95. https://doi.org/10.1186/1471-2474-15-95

Rose, N. (2020). Parameterschätzung und Messgenauigkeit in der Item-Response-Theorie (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3rd ed. 2020, S. 447–500). Berlin, Heidelberg: Springer.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. Verfügbar unter: http://www.jstatsoft.org/v48/i02/

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L. & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, *39*(5), 369–393. Journal of Research in Science Teaching, 39(5), 369-393. https://doi.org/10.1002/TEA.10027

Satorra, A. & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507–514. https://doi.org/10.1007/BF02296192

Schaper, N. & Hilkenmeier, F. (2013, September). *Umsetzungshilfen für kompetenzorientiertes Prüfen. HRK-Zusatzgutachten ausgearbeitet für die HRK.* Bonn: HRK.

Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the Fit of Structural Equation Models. Tests of Significance and Descriptive Goodness-of-Fit Measures. *Methods of Psychological Research Online*, *Vol.8*(2), 23–74.

Schindler, C. J. (2015). *Herausforderung Prüfen. Eine fallbasierte Untersuchung der Prüfungspraxis von Hochschullehrenden im Rahmen eines Qualitätsentwicklungsprogramms*. Dissertation. München. Zugriff am 10.03.2021.

Schmidt, S., Zlatkin-Troitschanskaia, O. & Fox, J.-P. (2016). Pretest-Posttest-Posttest Multilevel IRT Modeling of Competence Growth of Students in Higher Education in Germany. *Journal of Educational Measurement*, *53*(3), 332–351.

Schulz, F., Zehner, F., Schindler, C. & Prenzel, M. (2014). Prüfen und Lernen im Studium: Erste Schritte zur Untersuchung von Prüfungsanforderungen und Lerntypen. *Beiträge zur Hochschulforschung*, *36*(2), 34–58.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2). https://doi.org/10.1214/aos/1176344136

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

Seeber, S., Fischer, A., Michaelis, C. & Müller, J. (2014). Zur Messung von Kompetenzen zum nachhaltigen Wirtschaften mit einem Situational Judgement Test. *berufsbildung*, (146), 6–9.

Seeber, S., Hartig, J., Dierkes, S. & Schumann, M. [Michael]. (2016). Ko-NaMa – Simulation-based Measurement and Validation of a Competence Model for Sustainability Management. In H. A. Pant, O. Zlatkin-Troitschanskaia, C. Lautenbach, M. Toepper & D. Molerov (Hrsg.), *Modeling and Measuring Competencies in Higher Education – Validation*

and Methodological Innovations (KoKoHs). Overview of the Research Projects (S. 53–56). Berlin & Mainz. Humboldt University & Johannes Gutenberg University.

Seeber, S., Hartig, J., Dierkes, S., Schumann, M. [Matthias], Michaelis, C., Repp, A. et al. (2020). *Simulationsbasierte Messung und Validierung eines Kompetenzmodells für das Nachhaltigkeitsmanagement (Ko-NaMa)*. Berlin: IQB - Institut zu Qualitätsentwicklung im Bildungswesen ((Version 1) [Datensatz]. Verfügbar unter: http://doi.org/10.5159/IQB_KoNaMa_v1

Seeber, S., Michaelis, C., Repp, A., Hartig, J., Aichele, C., Schumann, M. [Matthias] et al. (2019). Assessment of Competences in Sustainability Management: Analyses to the Construct Dimensionality. *Zeitschrift für Pädagogische Psychologie*, *33*(2), 148–158. https://doi.org/10.1024/1010-0652/a000240

Seidel, T. (2014). Angebots-Nutzungs-Modelle in der Unterichspsychologie. Integration von Struktur- und Prozessparadigma. *Zeitschrift für Pädagogik*, *60*(6), 850–866.

Shavelson, R. J., Ruiz-Primo, M. A. & Wiley, E. W. (2005). Windows into the mind. *Higher Education*, *49*(4), 413–430.

Ștefănică, F. (2013). Modulbeschreibungen – Deskriptionen realer Ansprüche oder realitätsferne Lyrik? Eine qualitative Analyse am Beispiel (Höhere / Angewandte) Mathematik I/II im Rahmen des Maschinenbaustudiums an ausgewählten Hochschulstandorten Baden-Württembergs. *Zeitschrift für Berufs- und Wirtschaftspädagogik*, *109*, 286–303.

Stiftung für Hochschulzulassung (Hrsg.). (2020, 22. Juni). *Ergänzende Informationen für Ihre Studienplatzbewerbung im Zentralen Vergabeverfahren für bundesweit zulassungsbeschränkte Studiengänge*. Zugriff am 03.09.2020. Verfügbar unter https://www.hochschulstart.de/fileadmin/media/epaper/hilfe20-21/hilfe-zur-bewerbung-20-21/index.html#p=6

Stiller, J., Hartmann, S., Mathesius, S., Straube, P., Tiemann, R., Nordmeier, V. et al. (2016). Assessing scientific reasoning: a comprehensive evaluation of item features that affect item difficulty. *Assessment & Evaluation in Higher Education*, *41*(5), 721–732.

Tan, F. E. S., Jolani, S. & Verbeek, H. (2018). Guidelines for multiple imputations in repeated measurements with time-dependent covariates: a case study. *Journal of Clinical Epidemiology*, *102*, 107–114. https://doi.org/10.1016/j.jclinepi.2018.06.006

Technical recommendations for psychological tests and diagnostic techniques. (1954). *Psychological Bulletin*, *51*(2, Pt.2), 1–38. https://doi.org/10.1037/h0053479

Tremp, P. & Eugster, B. (2006). Universitäre Bildung und Prüfungssystem. Thesen zu Leistungsnachweisen in modularisierten Studiengängen. *Hochschulwesen*, *54*(5).

Tsarouha, E. (2019). *Prüfungspraktiken an deutschen Hochschulen. Eine empirische Studie zu systematischen Einflussgrößen auf die Notengebung in Abschlussprüfungen* (Wissenschaft – Hochschule – Bildung, 1. Auflage 2019). Wiesbaden: Springer Fachmedien Wiesbaden GmbH. Verfügbar unter: http://www.springer.com/

UniReport Satzungen und Ordnungen. (2020, 8. Septembera). *Ordnung des Fachbereichs Psychologie und Sportwissenschaften der Johann Wolfgang Goethe-Universität Frankfurt am Main für den Bachelorstudiengang Psychologie mit dem Abschluss „Bachelor of Science (B.Sc.)" vom 14. Juli 2020* (die Präsidentin der Johann Wolfgang Goethe-Universität Frankfurt am Main, Hrsg.). Zugriff am 04.05.2021. Verfügbar unter: https://www.psychologie.uni-frankfurt.de/91860275/BA_Psychologie_2020_09_08.pdf

UniReport Satzungen und Ordnungen. (2020, 22. Dezemberb). *Rahmenordnung für gestufte und modularisierte Studiengänge der Johann Wolfgang Goethe-Universität Frankfurt am Main vom 30. April 2014 in der Fassung vom 15. Juli 2020* (die Präsidentin der Johann Wolfgang Goethe-Universität Frankfurt am Main, Hrsg.). Johann Wolfgang Goethe-Universität Frankfurt am Main. Zugriff am 04.05.2021. Verfügbar unter: https://www.uni-frankfurt.de/95837297/2020_12_22_Novellierung_RO_mit_Anlagen__Seitenzahlen.pdf

Vogelsang, C. (2014). *Validierung eines Instruments zur Erfassung der professionellen Handlungskompetenz von (angehenden) Physiklehrkräften. Zusammenhangsanalysen zwischen Lehrerkompetenz und Lehrerperformanz* (d-nb.de/cgi-bin/dokserv?id=5101371&prov=M&dok_var=1&dok_ext=htm). Zugl.: Paderborn, Univ., Diss., 2014.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450.

Weinert, F. E. (2001). Vergleichende Leistungsmessung in Schulen - eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 17–31). Weinheim: Beltz.

Wiesbeck, A. B. (2015). *An Evaluation of Simulated Conversations as an Assessment of Pre-Service Teachers' Communication Competence in Parent-Teacher Conversations*. München, Technische Universität München, Diss., 2015. Retrieved from http://nbn-resolving.de/urn:nbn:de:bvb:91-diss-20150821-1271404-1-2; http://d-nb.info/1077063628/34

Wolter, F. & Schiener, J. (2014). En Route to “University Pisa”? On the Measurement of Sociological Competencies. *SOZIALE WELT-ZEITSCHRIFT FUR SOZIALWISSENSCHAFTLICHE FORSCHUNG UND PRAXIS*, *65*(1), 47+.

Yen, W. M. (1993). Scaling Performance Assessments: Strategies for Managing Local Item Dependence. *Journal of Educational Measurement*, *30*(3), 187–213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Zlatkin-Troitschanskaia, O., Pant, H. A. [Hans Anand], Kuhn, C., Toepper, M. & Lautenbach, C. (2016). *Messung akademisch vermittelter Kompetenzen von Studierenden und Hochschulabsolventen. Ein Überblick zum nationalen und internationalen Forschungsstand* (Edition ZfE, Band 1). Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-658-10830-4

Anhang

A. Suchterme und Treffer der Literaturrecherche

Datum	Treffer	Suchterm	zusätzliche Kriterien
		EBSCOhost (PsycArticles, PsycINFO, ERC)	
12.04.2017	696	(german* AND "higher education") AND ("test development" OR competenc* N5 measur* OR competenc* N5 assess*)	Volltext, Englisch, Deutsch, 2010-2017
12.04.2017	1 509	(german* AND "higher education") AND (knowledge N5 measur* OR knowledge N5 assess* OR skill* N5 measur* OR skill* N5 assess*)	Volltext, Englisch, Deutsch, 2010-2017
12.04.2017	74	(deutsch* AND hochschul*) AND (testentwicklung OR testkonstruktion OR kompetenz* mess* OR kompetenz* erfass*)	
12.04.2017	123	(deutsch* AND hochschul*) AND (wissen* mess* OR wissen* erfass*)	
12.04.2017	25	(deutsch* AND hochschul*) AND (fähigkeit* mess* OR fähigkeit* erfass*)	
12.04.2017	16	(deutsch* AND hochschul*) AND (testentwicklung OR testkonstruktion OR kompetenz* N5 mess* OR kompetenz* N5 erfass*)	
12.04.2017	4	(deutsch* AND hochschul*) AND (wissen* N5 mess* OR wissen* N5 erfass*)	
12.04.2017	2	(deutsch* AND hochschul*) AND (fähigkeit* N5 mess* OR fähigkeit* N5 erfass*)	
18.04.2017	16	(deutsch* AND universitär*) AND (testentwicklung OR testkonstruktion OR kompetenz* mess* OR kompetenz* erfass*)	
18.04.2017	31	(deutsch* AND universitär*) AND (wissen* mess* OR wissen* erfass*)	
18.04.2017	15	(deutsch* AND universitär*) AND (fähigkeit* mess* OR fähigkeit* erfass*)	
18.04.2017	4	(deutsch* AND universitär*) AND (testentwicklung OR testkonstruktion OR kompetenz* N5 mess* OR kompetenz* N5 erfass*)	
18.04.2017	4	(deutsch* AND universitär*) AND (wissen* N5 mess* OR wissen* N5 erfass*)	
18.04.2017	0	(deutsch* AND universitär*) AND (fähigkeit* N5 mess* OR fähigkeit* N5 erfass*)	
		Web of Science/ Knowledge	
12.04.2017	7	TOPIC: ((german* AND "higher education") AND ("test development" OR competenc* NEAR measur* OR competenc* NEAR assess*))	
12.04.2017	12	**TOPIC: ((german* AND "higher education") AND (knowledge NEAR measur* OR knowledge NEAR assess* OR skill* NEAR measure OR skill* NEAR assess*))**	
		FIS Bildung (u.a. ERIC)	
12.04.2017	0	**(** (Freitext: GERMAN* *und* "HIGHER EDUCATION") *und* (Freitext: "TEST DEVELOP*") **)** *und* (Jahr >=2010) **)** *und* (Sprache: deutsch *oder* englisch)	
12.04.2017	13	**(** (Freitext: GERMAN* *und* "HIGHER EDUCATION") *und* (Freitext: COMPETENC* *und* MEASUR*) **)** *und* (Jahr >=2010) **)** *und* (Sprache: deutsch *oder* englisch)	
12.04.2017	14	**(** (Freitext: GERMAN* *und* "HIGHER EDUCATION") *und* (Freitext: COMPETENC* *und* ASSESS*) **)** *und* (Jahr >=2010) **)** *und* (Sprache: deutsch *oder* englisch)	

12.04.2017	6	**((** (Freitext: GERMAN* *und* "HIGHER EDUCATION") *und* (Freitext: KNOWLEDGE *und* MEASUR*) **)** *und* (Jahr >=2010) **)** *und* (Sprache: deutsch *oder* englisch)
12.04.2017	12	**((** (Freitext: GERMAN* *und* "HIGHER EDUCATION") *und* (Freitext: KNOWLEDGE *und* ASSESS*) **)** *und* (Jahr >=2010) **)** *und* (Sprache: deutsch *oder* englisch)
12.04.2017	4	**((** (Freitext: GERMAN* *und* "HIGHER EDUCATION") *und* (Freitext: SKILL* *und* ASSESS*) **)** *und* (Jahr >=2010) **)** *und* (Sprache: deutsch *oder* englisch)
12.04.2017	6	**((** (Freitext: GERMAN* *und* "HIGHER EDUCATION") *und* (Freitext: SKILL* *und* MEASUR*) **)** *und* (Jahr >=2010) **)** *und* (Sprache: deutsch *oder* englisch)
12.04.2017	46	**((** (Freitext: DEUTSCH* *und* HOCHSCHUL*) *und* (Freitext: TEST*ENTWICKLUNG *oder* TEST*KONSTRUKTION) **)** *und* (Jahr >=2010) **)** *und* (Sprache: deutsch *oder* englisch)
12.04.2017	199	**((** (Freitext: DEUTSCH* *und* HOCHSCHUL*) *und* (Freitext: KOMPETENZ* *und* MESS*) **)** *und* (Jahr >=2010) **)** *und* (Sprache: deutsch *oder* englisch)
12.04.2017	165	**((** (Freitext: DEUTSCH* *und* HOCHSCHUL*) *und* (Freitext: KOMPETENZ* *und* ERFASS*) **)** *und* (Jahr >=2010) **)** *und* (Sprache: deutsch *oder* englisch)
12.04.2017	195	**((** (Freitext: DEUTSCH* *und* HOCHSCHUL*) *und* (Freitext: WISSEN* *und* ERFASS*) **)** *und* (Jahr >=2010) **)** *und* (Sprache: deutsch *oder* englisch)
12.04.2017	184	**((** (Freitext: DEUTSCH* *und* HOCHSCHUL*) *und* (Freitext: WISSEN* *und* MESS*) **)** *und* (Jahr >=2010) **)** *und* (Sprache: deutsch *oder* englisch)
12.04.2017	31	**((** (Freitext: DEUTSCH* *und* HOCHSCHUL*) *und* (Freitext: FAEHIGKEIT* *und* MESS*) **)** *und* (Jahr >=2010) **)** *und* (Sprache: deutsch *oder* englisch)
12.04.2017	38	**((** (Freitext: DEUTSCH* *und* HOCHSCHUL*) *und* (Freitext: FAEHIGKEIT* *und* ERFASS*) **)** *und* (Jahr >=2010) **)** *und* (Sprache: deutsch *oder* englisch)
12.04.2017	51	**((** (Freitext: DEUTSCH* *und* HOCHSCHUL*) *und* (Freitext: TEST*ENTWICKLUNG *oder* TEST*KONSTRUKTION *oder* TESTINSTRUMENT) **)** *und* (Jahr >=2010) **)** *und* (Sprache: deutsch *oder* englisch)
18.04.2017	38	**((** (Freitext: DEUTSCH* *und* UNIVERSITAE*) *und* (Freitext: TEST*ENTWICKLUNG *oder* TEST*KONSTRUKTION *oder* TESTINSTRUMENT) **)** *und* (Jahr >=2010) **)** *und* (Sprache: deutsch *oder* englisch)
18.04.2017	108	**((** (Freitext: DEUTSCH* *und* UNIVERSITAE*) *und* (Freitext: KOMPETENZ* *und* MESS*) **)** *und* (Jahr >=2010) **)** *und* (Sprache: deutsch *oder* englisch)
18.04.2017	123	**((** (Freitext: DEUTSCH* *und* UNIVERSITAE*) *und* (Freitext: KOMPETENZ* *und* ERFASS*) **)** *und* (Jahr >=2010) **)** *und* (Sprache: deutsch *oder* englisch)
18.04.2017	158	**((** (Freitext: DEUTSCH* *und* UNIVERSITAE*) *und* (Freitext: WISSEN* *und* ERFASS*) **)** *und* (Jahr >=2010) **)** *und* (Sprache: deutsch *oder* englisch)
18.04.2017	115	**((** (Freitext: DEUTSCH* *und* UNIVERSITAE*) *und* (Freitext: WISSEN* *und* MESS*) **)** *und* (Jahr >=2010) **)** *und* (Sprache: deutsch *oder* englisch)
18.04.2017	28	**((** (Freitext: DEUTSCH* *und* UNIVERSITAE*) *und* (Freitext: FAEHIGKEIT* *und* MESS*) **)** *und* (Jahr >=2010) **)** *und* (Sprache: deutsch *oder* englisch)

		ERIC	
18.04.2017	58	german AND "higher education" AND (competence measure OR competence assessment)	since 2008
18.04.2017	1	german AND "higher education" AND "test development"	
18.04.2017	70	german AND "higher education" AND (knowledge measure OR knowledge assessment)	
18.04.2017	71	german AND "higher education" AND (skill measure OR skill assessment)	

B. Literaturnachweis der qualitativen Arbeit

Aschinger, F., Epstein, H., Mueller, S., Schaeper, H., Voettiner, A. & Weiss, T. (2011). Higher education and the transition to work. *Zeitschrift für Erziehungswissenschaft, 14* (14), 267–282. https://doi.org/10.1007/s11618-011-0190-7

Berger, S., Bouley, F., Fritsch, S., Krille, C., Seifried, J. & Wuttke, E. (2015). Fachwissen und fachdidaktisches Wissen im wirtschaftspädagogischen Studium. Entwicklung eines Testinstruments und erste empirische Befunde. In B. Koch-Priewe, A. Köker, J. Seifried & E. Wuttke (Hrsg.), *Kompetenzerwerb an Hochschulen: Modellierung und Messung. Zur Professionalisierung angehender Lehrerinnen und Lehrer sowie frühpädagogischer Fachkräfte* (S. 105–128). Bad Heilbrunn: Klinkhardt.

Berger, S., Fritsch, S., Seifried, J., Bouley, F., Mindnich, A., Wuttke, E. et al. (2013). Entwicklung eines Testinstruments zur Erfassung des fachlichen und fachdidaktischen Wissens von Studierenden der Wirtschaftspädagogik. Erste Erfahrungen und Befunde. In O. Zlatkin-Troitschanskaia, R. Nickolaus & K. Beck (Hrsg.)Kompetenzmodellierung und Kompetenzmessung bei Studierenden der Wirtschaftswissenschaften und der Ingenieurwissenschaften, 93–107 [Themenheft]. Landau: Empirische Pädagogik.

Biermann, A., Kaub, K., Friedrich, A., Wach, F.-S., Ruffing, S., Reichl, C. et al. (2017). SioS-L – Studie zu individuellen und organisationalen Einflüssen auf den Studienerfolg in der Lehrerbildung. In C. Gräsel & K. Trempler (Hrsg.), *Entwicklung von Professionalität pädagogischen Personals* (S. 75–92). Wiesbaden: Springer Fachmedien Wiesbaden.

Blömeke, S., Jenßen, L., Grassmann, M., Dunekacke, S. & Wedekind, H. (2017). Process mediates structure: The relation between preschool teacher education and preschool teachers' knowledge. *Journal of Educational Psychology, 109* (3), 338–354. https://doi.org/10.1037/edu0000147

Borowski, A., Kirschner, S., Liedtke, S. & Fischer, H. E. (2011). Vergleich des Fachwissens von Studierenden, Referendaren und Lehrenden in der Physik. *Physik und Didaktik in Schule und Hochschule, 10* (1), 1–9.

Bouley, F., Berger, S., Fritsch, S., Wuttke, E., Seifried, J., Schnick-Vollmer, K. et al. (2015). Der Einfluss von universitären und außeruniversitären Lerngelegenheiten auf das Fachwissen und fachdidaktische Wissen von angehenden Lehrkräften an kaufmännisch-berufsbildenden Schulen. In S. Blömeke & O. Zlatkin-Troitschanskaia (Hrsg.)Kompetenzen von Studierenden, 100–115 [Themenheft].

Brandenburger, M., Mikelskis-Seifert, S. & Labudde, P. (2014). Problemlösen in der Mechanik: eine Untersuchung mit Studierenden. *PhyDid B, Didaktik der Physik, Beiträge zur DPG-Frühjahrstagung, 2014,* 8. Verfügbar unter http://www.phydid.de/index.php/phydid-b/article/view/511

Bremerich-Vos, A., Dämmer, J., Willenberg, H. & Schwippert, K. (2011). Professionelles Wissen von Studierenden des Lehramts deutsch. In S. Blömeke, A. Bremerich-Vos, H. Haudeck, G. Kaiser, G. Nold, K. Schwippert et al. (Hrsg.), *Kompetenzen von Lehramtsstudierenden in gering strukturierten Domänen. Erste Ergebnisse aus TEDS-LT* (S. 47–76). Münster: Waxmann.

Bromme, R. & Thomm, E. (2016). Knowing Who Knows: Laypersons' Capabilities to Judge Experts' Pertinence for Science Topics. *Cognitive Science, 40* (1), 241–252.

Brückner, S., Förster, M., Zlatkin-Troitschanskaia, O., Happ, R., Walstad, W. B., Yamaoka, M. et al. (2015). Gender effects in assessment of economic knowledge and understanding: Differences among undergraduate business and economics students in Germany, Japan, and the United States. *Peabody Journal of Education, 90* (4), 503–518. https://doi.org/10.1080/0161956X.2015.1068079

Brückner, S. & Pellegrino, J. W. (2016). Integrating the analysis of mental operations into multilevel models to validate an assessment of higher education students' competency in business and economics. *Journal of Educational Measurement, 53* (3), 293–312.

Buchholtz, N. (2014). *Multiperspektivische Ansätze zur Messung des Lehrerprofessionswissens in der Mathematiklehramtsausbildung*. Hamburg, Universität Hamburg, Diss., 2014.

Buchholtz, N. & Kaiser, G. (2013). Improving Mathematics Teacher Education in Germany. Empirical Results from a longitudinal Evaluation of innovative Programs. *International Journal of Science and Mathematics Education, 11* (4), 949–977. https://doi.org/10.1007/s10763-013-9427-7

Buchholtz, N. & Kaiser, G. (2013). Professionelles Wissen im Studienverlauf: Lehramt Mathematik. In S. Blömeke, A. Bremerich-Vos, G. Kaiser, G. Nold, H. Haudeck, J.-U. Kessler et al. (Hrsg.), *Professionelle Kompetenzen im Studienverlauf. Weitere Ergebnisse zur Deutsch-, Englisch- und Mathematiklehrerausbildung aus TEDS-LT* (info/1031260226/04; Verlagsangaben: http://deposit.d-nb.de/cgi-bin/dokserv?id=4254478&prov=M&dok_var=1&dok_ext=htm, S. 107–143).

Buchholtz, N., Kaiser, G. & Stancel-Piqtak, A. (2011). Professionelles Wissen von Studierenden des Lehramts Mathematik. In S. Blömeke, A. Bremerich-Vos, H. Haudeck, G. Kaiser, G. Nold,

K. Schwippert et al. (Hrsg.), *Kompetenzen von Lehramtsstudierenden in gering strukturierten Domänen. Erste Ergebnisse aus TEDS-LT* (S. 101–133). Münster: Waxmann.

Buchholtz, N., Leung, F. K. S., Ding, L., Kaiser, G., Park, K. & Schwarz, B. (2013). Future mathematics teachers' professional knowledge of elementary mathematics from an advanced standpoint. *ZDM, 45* (1), 107–120. https://doi.org/10.1007/s11858-012-0462-6

Cauet, E. (2016). *Testen wir relevantes Wissen? Zusammenhang zwischen dem Professionswissen von Physiklehrkräften und gutem und erfolgreichem Unterrichten* (d-nb.de/cgi-bin/dokserv?id=367909e9ceb6498e8e2f0ff9ffa24974&prov=M&dok_var=1&dok_ext=htm; Inhaltsverzeichnis: http://d-nb.info/1103521837/04). Dissertation, Universität Duisburg-Essen, 2016.

Döhrmann, M., Kaiser, G. & Blömeke, S. (2010). Messung des mathematischen und mathematikdidaktischen Wissens. Theoretischer Rahmen und Teststruktur. In S. Blömeke, G. Kaiser & R. Lehmann (Hrsg.), *TEDS-M 2008 (3). Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich* (S. 169–196). Münster: Waxmann.

Döhrmann, M., Kaiser, G. & Blömeke, S. (2010). Messung mathematischen und mathematikdidaktischen Wissens. Theoretischer Rahmen und Teststruktur. In S. Blömeke, G. Kaiser & R. Lehmann (Hrsg.), *TEDS-M 2008 (2). Professionelle Kompetenz und Lerngelegenheiten angehender Primarstufenlehrkräfte im internationalen Vergleich* (S. 169–194). Münster: Waxmann.

Dübbelde, G. (2013). *Diagnostische Kompetenzen angehender Biologie-Lehrkräfte im Bereich der naturwissenschaftlichen Erkenntnisgewinnung*. Kassel, Univ., Diss., 2013.

Dunekacke, S., Jenßen, L. & Blömeke, S. (2015). Mathematikdidaktische Kompetenz von Erzieherinnen und Erziehern. Validierung des KomMa-Leistungstests durch die videogestützte Erhebung von Performanz. In S. Blömeke & O. Zlatkin-Troitschanskaia (Hrsg.), *Kompetenzen von Studierenden* (Zeitschrift für Pädagogik : Beiheft, Bd. 61, S. 80–99). Weinheim: Beltz Juventa.

Formazin, M., Schroeders, U., Köller, O., Wilhelm, O. & Westmeyer, H. (2011). Studierendenauswahl im Fach Psychologie. Testentwicklung und Validitätsbefunde. *Psychologische Rundschau, 62* (4), 221–236. Verfügbar unter http://dx.doi.org/10.1026/0033-3042/a000093

Förster, M., Brückner, S., Beck, K., Zlatkin-Troitschanskaia, O. & Happ, R. (2016). Individuelle und kontextuelle Prädiktoren des Fachwissenserwerbs zum Internen Rechnungswesen im Hochschulstudium. Paralleltitel: Individual and contextual predictors of the acquisition of content knowledge in internal accounting in higher education. *Zeitschrift für Erziehungswissenschaft, 19* (2), 375–393.

Förster, M., Happ, R. & Zlatkin-Troitschanskaia, O. (2012). Valide Erfassung des volkswirtschaftlichen Fachwissens von Studierenden der Wirtschaftswissenschaften und der Wirtschaftspädagogik. Eine Untersuchung der diagnostischen Eignung des Wirtschaftskundlichen Bildungstests (WBT). *Berufs- und Wirtschaftspädagogik Online* (22), 23.

Förster, M. & Zlatkin-Troitschanskaia, O. (2010). Wirtschaftliche Fachkompetenz bei Studierenden mit und ohne Lehramtsperspektive in den Diplom- und Bachelorstudiengängen - Messverfahren und erste Befunde. Paralleltitel: Measurement of professional competence in the domain of Economics of university students in Economics and in Business and Economics Education - Comparison of the (old) diploma and the (new) bachelor degree. In K. Beck & O. Zlatkin-Troitschanskaia (Hrsg.), *Lehrerprofessionalität. Was wir wissen und was wir wissen müssen* (Lehrerbildung auf dem Prüfstand, Jg. 3, Sonderheft, S. 106–125). Landau: Verl. Empirische Pädagogik.

Förster, M., Zlatkin-Troitschanskaia, O., Brückner, S., Happ, R., Hambleton, R. K., Walstad, W. B. et al. (2015). Validating test score interpretations by cross-national comparison: Comparing the results of students from Japan and Germany on an American test of economic knowledge in higher education. *Zeitschrift für Psychologie, 223* (1), 14–23. https://doi.org/10.1027/2151-2604/a000195

Fritsch, S., Berger, S., Seifried, J., Bouley, F., Wuttke, E., Schnick-Vollmer, K. et al. (2015). The impact of university teacher training on prospective teachers' CK and PCK – a comparison between Austria and Germany. *Empirical Research in Vocational Education and Training, 7:4*.

Gold, B., Förster, S. & Holodynski, M. (2013). Evaluation eines videobasierten Trainingsseminars zur Förderung der professionellen Wahrnehmung von Klassenführung im Grundschulunterricht*. *Zeitschrift für pädagogische Psychologie, 27* (3), 141–155. https://doi.org/10.1024/1010-0652/a000100

Gramzow, Y. (2015). *Fachdidaktisches Wissen von Lehramtsstudierenden im Fach Physik. Modellierung und Testkonstruktion* (Studien zum Physik- und Chemielernen, Bd. 181). Berlin: Logos Verl.

Groß Ophoff, J., Schladitz, S., Leuders, J., Leuders, T. & Wirtz, M. A. (2015). Assessing the development of educational research literacy: The effect of courses on research methods in studies of educational science. *Peabody Journal of Education, 90* (4), 560–573. https://doi.org/10.1080/0161956X.2015.1068085

Groß Ophoff, J., Schladitz, S., Lohrmann, K. & Wirtz, M. (2014). Evidenzorientierung in bildungswissenschaftlichen Studiengängen. In K. Drossel, R. Strietholt & W. Bos (Hrsg.), *Empirische Bildungsforschung und evidenzbasierte Reformen im Bildungswesen* (S. 251–275). Münster u.a.: Waxmann.

Großschedl, J., Harms, U., Glowinski, I. & Waldmann, M. (2014). Professionswissen angehender Biologielehrkräfte. Das KiL-Projekt. *Der mathematische und naturwissenschaftliche Unterricht, 67* (8), 457–462.

Großschedl, J., Harms, U., Kleickmann, T. & Glowinski, I. (2015). Preservice Biology Teachers' Professional Knowledge: Structure and Learning Opportunities. *Journal of Science Teacher Education, 26* (3), 291–318.

Großschedl, J., Neubrand, C., Kirchner, A., Oppermann, L., Basel, N. & Gantner, S. (2015). Entwicklung und Validierung eines Testinstruments zur Erfassung des evolutionsbezogenen Professionswissens von Lehramtsstudierenden (ProWiE). Paralleltitel: Development and Validation of a Test Instrument for the Assessment of the Evolution-related Professional Knowledge of Pre-service Teachers (ProWiE). *Zeitschrift für Didaktik der Naturwissenschaften, 21* (1), 173–185.

Hammer, S., Carlson, S. A., Ehmke, T., Koch-Priewe, B., Köker, A., Ohm, U. et al. (2015). Kompetenz von Lehramtsstudierenden in Deutsch als Zweitsprache. Validierung des GSL-Testinstruments. In S. Blömeke & O. Zlatkin-Troitschanskaia (Hrsg.), *Kompetenzen von Studierenden* (Zeitschrift für Pädagogik : Beiheft, Bd. 61, S. 32–54). Weinheim: Beltz Juventa.

Händel, M., Tupac-Yupanqui, A. & Lockl, K. (Dezember 2012). *Metakognitives Wissen und der Einsatz von Lernstrategien bei Studierenden* (NEPS Working Paper Nr. 20). Bamberg: German National Educational Panel Study (NEPS) - Universität Bamberg. Verfügbar unter https://www.neps-data.de/Portals/0/Working%20Papers/WP_XX.pdf

Hartmann, S., Mathesius, S., Stiller, J., Straube, P., Krüger, D. & Upmeier zu Belzen, A. (2015). Kompetenzen der naturwissenschaftlichen Erkenntnisgewinnung als Teil des Professionswissens zukünftiger Lehrkräfte: Das Projekt Ko-WADiS. In B. Koch-Priewe, A. Köker, J. Seifried & E. Wuttke (Hrsg.), *Kompetenzerwerb an Hochschulen: Modellierung und Messung. Zur Professionalisierung angehender Lehrerinnen und Lehrer sowie frühpädagogischer Fachkräfte* (S. 39–58). Bad Heilbrunn: Klinkhardt.

Hartmann, S., Upmeier zu Belzen, A., Krüger, D. & Pant, H. A. (2015). Scientific reasoning in higher education: Constructing and evaluating the criterion-related validity of an assessment of preservice science teachers' competencies. *Zeitschrift für Psychologie, 223* (1), 47–53. https://doi.org/10.1027/2151-2604/a000199

Heigl, N. R. (2014). *Cross-curricular problem solving. Entwicklung und Erprobung eines Testverfahrens und Analyse von Bedingungen fächerübergreifender Problemlösekompetenzen bei Studierenden*. Eichstätt-Ingolstadt, Katholische Universität Eichstätt-Ingolstadt, Diss., 2014.

Heigl, N. R., Zoelch, C. & Thomas, J. (2011). Cross-Curricular Problem Solving. An Empirical Study of Demands on Teaching and Learning in Higher Education. In J. Özyurt, A. Anschütz, S. Bernholt & J. Lenk (Hrsg.), *Interdisciplinary perspectives on cognition, education and the brain* (Hanse-Studien, Bd. 7, S. 175–186). Oldenburg: BIS-Verl. der Carl-von-Ossietzky-Univ.

Hendler, J., Mischo, C., Wahl, S. & Strohmer, J. (2011). Das sprachbezogene Wissen angehender frühpädagogischer Fachkräfte im Wissenstest und in der Selbsteinschätzung. Paralleltitel: The early childhood practitioners' knowledge about language in a knowledge test and in the self-assessment. *Empirische Pädagogik, 25* (4), 518–542.

Hohenstein, F. (2015). *Pädagogisch-psychologisches Wissen von Lehramtsstudierenden. Entwicklung und Validierung eines Testverfahrens*. Dissertation, Christian Albrechts Universität, 2015.

Hohenstein, F., Kleickmann, T., Zimmermann, F., Köller, O. & Möller, J. (2017). Erfassung von pädagogischem und psychologischem Wissen in der Lehramtsausbildung: Entwicklung eines Messinstruments. *Zeitschrift für Pädagogik, 63* (1), 91–112.

Jahn, G. K. (2014). *Studien zur Überprüfung der Validität eines Instruments zur Erfassung professioneller Unterrichtswahrnehmung von Lehramtsstudierenden*. München, Technische Universität München, Diss., 2014.

Jähnig, C. C. (2013). Assessing Business Knowledge of Students in German Higher Education. In U. Faßhauer, B. Fürstenau & E. Wuttke (Eds.), *Jahrbuch der berufs- und*

wirtschaftspädagogischen Forschung 2013 (Schriftenreihe der Sektion Berufs- und Wirtschaftspädagogik der Deutschen Gesellschaft für Erziehungswissenschaft (DGfE), pp. 47–59). Opladen u.a.: Budrich. Verfügbar unter http://www.pedocs.de/volltexte/2013/8063; http://nbn-resolving.de/urn:nbn:de:0111-opus-80634

Jähnig, C. C. (2013). *Die Messung betriebswirtschaftlichen Wissens von Studierenden. Eine quantitativ-empirische Untersuchung situativer Testaufgaben* (info/1050899768/04; Verlagsangaben: http://deposit.d-nb.de/cgi-bin/dokserv?id=4663692&prov=M&dok_var=1&dok_ext=htm). Zugl.: Göttingen, Univ., Diss., 2013.

Jansing, B., Haudeck, H., Keßler, C., Nold, G. & Stancel-Piqtak, A. (2013). Professionelles Wissen im Studienverlauf: Lehramt Englisch. In S. Blömeke, A. Bremerich-Vos, G. Kaiser, G. Nold, H. Haudeck, J.-U. Kessler et al. (Hrsg.), *Professionelle Kompetenzen im Studienverlauf. Weitere Ergebnisse zur Deutsch-, Englisch- und Mathematiklehrerausbildung aus TEDS-LT* (info/1031260226/04; Verlagsangaben: http://deposit.d-nb.de/cgi-bin/dokserv?id=4254478&prov=M&dok_var=1&dok_ext=htm, S. 77–106).

Jenßen, L., Dunekacke, S., Eid, M. & Blömeke, S. (2015). The relationship of mathematical competence and mathematics anxiety: An application of latent state-trait theory. *Zeitschrift für Psychologie, 223* (1), 31–38. https://doi.org/10.1027/2151-2604/a000197

Jurkowski, S. & Hänze, M. (2012). Förderung transaktiven Interaktionsverhaltens. Effekte eines Trainings transaktiver Interaktionsbeiträge auf den Lernerfolg beim kooperativen Lernen. *Zeitschrift für Entwicklungspsychologie und pädagogische Psychologie, 44* (4), 209–220. https://doi.org/10.1026/0049-8637/a000074

Jüttner, M. & Neuhaus, B. (2013). Das Professionswissen von Biologielehrkräften - Ein Vergleich zwischen Biologielehrkräften, Biologen und Pädagogen. *Zeitschrift für Didaktik der Naturwissenschaften, 19,* 31–49.

Kaiser, G., Busse, A., Hoth, J., König, J. & Blömeke, S. (2015). About the Complexities of Video-Based Assessments: Theoretical and Methodological Approaches to Overcoming Shortcomings of Research on Teachers' Competence. *International Journal of Science & Mathematics Education, 13* (2), 369–387.

Kehne, M., Seifert, A. & Schaper, N. (2013). Struktur eines Instruments zur Kompetenzerfassung in der Sportlehrerausbildung. *Sportunterricht, 62* (2), 53–57. Verfügbar unter http://www.hofmann-

verlag.de/project/zs_archiv/archiv/sportunterricht/2013/Sportunterricht-Ausgabe-Februar-2013.pdf#page=53

Kemper, C. J., Mitschke, T., Rollett, W., Kemper, V. & Oberfeld, D. (2016). Verbesserung der Präsentationskompetenz in der Lehre an deutschen Hochschulen. Entwicklung des Mainzer Verfahrens zur Peer-Evaluation studentischer Präsentationen (MPEP). *Das Hochschulwesen, 64* (3), 95–103.

Kirschner, S. (2013). *Modellierung und Analyse des Professionswissens von Physiklehrkräften*. Duisburg, Essen, Univ. Duisburg-Essen, Diss., 2013.

Kleickmann, T., Großschedl, J., Harms, U., Heinze, A., Herzog, S., Hohenstein, F. et al. (2014). Professionswissen von Lehramtsstudierenden der mathematisch-naturwissenschaftlichen Fächer - Testentwicklung im Rahmen des Projekts KiL. *Unterrichtswissenschaft, 42* (3), 280–288. Verfügbar unter http://www.beltz.de/fachmedien/erziehungs_und_sozialwissenschaften/zeitschriften/unterrichtswissenschaft/article/Journal.html?tx_beltz_journal[article]=27596&cHash=c5d8eb5f9b634b730d656de413db3ad9; https://www.digizeitschriften.de/dms/img/?PID=PPN513613439_0042|LOG_0037

Klug, J. (2011). *Modeling and Training a New Concept of Teachers' Diagnostic Competence*. Dissertation, Technische Universität Darmstadt, 2011.

Kollar, I., Ufer, S., Reichersdorfer, E., Vogel, F., Fischer, F. & Reiss, K. (2014). Effects of collaboration scripts and heuristic worked examples on the acquisition of mathematical argumentation skills of teacher students with different levels of prior achievement. *Learning and Instruction, 32,* 22–36. https://doi.org/10.1016/j.learninstruc.2014.01.003

König, J. (2012). Die Entwicklung von pädagogischem Unterrichtswissen. Theoretischer Rahmen, Testinstrument, Skalierung und Ergebnisse. In J. König & A. Seifert (Hrsg.), *Lehramtsstudierende erwerben pädagogisches Professionswissen. Ergebnisse der Längsschnittstudie LEK zur Wirksamkeit der erziehungswissenschaftlichen Lehrerausbildung.* Münster: Waxmann.

König, J. & Blömeke, S. (2010). Messung des pädagogischen Wissens. Theoretischer Rahmen und Teststruktur. In S. Blömeke, G. Kaiser & R. Lehmann (Hrsg.), *TEDS-M 2008 (2). Professionelle Kompetenz und Lerngelegenheiten angehender Primarstufenlehrkräfte im internationalen Vergleich* (S. 253–273). Münster: Waxmann.

König, J., Blömeke, S., Paine, L., Schmidt, W. H. & Hsieh, F.-J. (2011). General Pedagogical Knowledge of Future Middle School Teachers. On the Complex Ecology of Teacher

Education in the United States, Germany, and Taiwan. *Journal of Teacher Education, 62* (2), 188–201. https://doi.org/10.1177/0022487110388664

König, J. (2010). Längsschnittliche Erhebung pädagogischer Kompetenzen von Lehramtsstudierenden (LEK). Theoretischer Rahmen, Fragestellungen, Untersuchungsanlage und erste Ergebnisse zu Lernvoraussetzungen von angehenden Lehrkräften. Paralleltitel: Longitudinal survey of pedagogical competencies of teacher students (LEK). *Lehrerbildung auf dem Prüfstand, 3* (1), 56–83.

König, J. (2012). Zum Einfluss der Schulpraxis auf den Erwerb von pädagogischem Wissen: Spielen erste Unterrichtsversuche eine Rolle? In T. Hascher & G. H. Neuweg (Hrsg.), *Forschung zur (Wirksamkeit der) Lehrer/innen/bildung* (Österreichische Beiträge zur Bildungsforschung, S. 143–159). Berlin u.a.: Lit.

König, J. (2015). Measuring classroom management expertise (CME) of teachers. A video-based assessment approach and statistical results. *Cogent Education, 2* (1). https://doi.org/10.1080/2331186X.2014.991178

König, J. & Blömeke, S. (2012). Future teachers' general pedagogical knowledge from a comparative perspective: does school experience matter? *ZDM, 44* (3), 341–354. https://doi.org/10.1007/s11858-012-0394-1

König, J., Blömeke, S. & Doll, J. (2011). Pädagogisches Wissen von Deutsch-, Englisch- und Mathematiklehramtsstudierenden. In S. Blömeke, A. Bremerich-Vos, H. Haudeck, G. Kaiser, G. Nold, K. Schwippert et al. (Hrsg.), *Kompetenzen von Lehramtsstudierenden in gering strukturierten Domänen. Erste Ergebnisse aus TEDS-LT* (S. 135–157). Münster: Waxmann.

König, J., Blömeke, S., Klein, P., Suhl, U., Busse, A. & Kaiser, G. (2014). Is teachers' general pedagogical knowledge a premise for noticing and interpreting classroom situations? A video-based assessment approach. *Teaching & Teacher Education, 38,* 76–88.

König, J., Buchholtz, C. & Dohmen, D. (2015). Analyse von schriftlichen Unterrichtsplanungen. Empirische Befunde zur didaktischen Adaptivität als Aspekt der Planungskompetenz angehender Lehrkräfte. *Zeitschrift für Erziehungswissenschaft, 18* (2), 375–404. https://doi.org/10.1007/s11618-015-0625-7

König, J. & Herzmann, P. (2011). Lernvoraussetzungen angehender Lehrkräfte am Anfang ihrer Ausbildung. Erste Ergebnisse aus der wissenschaftlichen Begleitung des Kölner Modellkollegs Bildungswissenschaften. Paralleltitel: Learning preconditions of future teachers at the beginning of pre-service teacher education. *Lehrerbildung auf dem Prüfstand, 4* (2), 186–210.

König, J. & Kramer, C. (2016). Teacher professional knowledge and classroom management. On the relation of general pedagogical knowledge (GPK) and classroom management expertise (CME). *ZDM, 48* (1-2), 139–151. https://doi.org/10.1007/s11858-015-0705-4

König, J. & Lebens, M. (2012). Classroom Management Expertise (CME) von Lehrkräften messen: Überlegungen zur Testung mithilfe von Videovignetten und erste empirische Befunde. *Lehrerbildung auf dem Prüfstand, 5* (1), 3–28.

König, J. & Rothland, M. (2012). Motivations for choosing teaching as a career. Effects on general pedagogical knowledge during initial teacher education. *Asia-Pacific Journal of Teacher Education, 40* (3), 289–315. https://doi.org/10.1080/1359866X.2012.700045

König, J., Tachtsoglou, S. & Seifert, A. (2012). Individuelle Voraussetzungen, Lerngelegenheiten und der Erwerb von pädagogischem Professionswissen. In J. König & A. Seifert (Hrsg.), *Lehramtsstudierende erwerben pädagogisches Professionswissen. Ergebnisse der Längsschnittstudie LEK zur Wirksamkeit der erziehungswissenschaftlichen Lehrerausbildung* (S. 234–283). Münster: Waxmann.

Kotzebue, L. v. (2014). Diagram competence as a task in biology education for the teacher training. *Diagrammkompetenz als biologiedidaktische Aufgabe für die Lehrerbildung. Konzeption, Entwicklung und empirische Validierung eines Strukturmodells zum diagrammspezifischen Professionswissen im biologischem Kontext*. München, Technische Universität München, Diss., 2014.

Kotzebue, L. v. & Nerdel, C. (2012). Professionswissen von Biologielehrkräften zum Umgang mit Diagrammen. *Zeitschrift für Didaktik der Naturwissenschaften, 18,* 181–200.

Kufner, S. (2013). *Diagnose und Prognose von Handlungskompetenz im Bereich adaptiven Lehrens bei Studierenden - eine Videostudie*. Passau, Universität Passau, Diss., 2013.

Kuhn, C. (2014). *Fachdidaktisches Wissen von Lehrkräften im kaufmännisch-verwaltenden Bereich. Modellbasierte Testentwicklung und Validierung* (Empirische Berufsbildungs- und Hochschulforschung, Bd. 2). Landau: Verlag Empirische Pädagogik.

Kuhn, C., Happ, R., Zlatkin-Troitschanskaia, O., Beck, K., Förster, M. & Preuße, D. (2014). Kompetenzentwicklung angehender Lehrkräfte im kaufmännisch-verwaltenden Bereich – Erfassung und Zusammenhänge von Fachwissen und fachdidaktischem Wissen. *Zeitschrift für Erziehungswissenschaft, 17* (S1), 149–167. https://doi.org/10.1007/s11618-013-0456-3

Kunina-Habenicht, O., Schulze-Stocker, F., Kunter, M., Baumert, J., Leutner, D., Förster, D. et al. (2013). Die Bedeutung der Lerngelegenheiten im Lehramtsstudium und deren individuelle Nutzung für den Aufbau des bildungswissenschaftlichen Wissens. *Zeitschrift für Pädagogik,*

59 (1), 1–23. Verfügbar unter http://www.beltz.de/fachmedien/erziehungs_und_sozialwissenschaften/zeitschriften/zeitschrift_fuer_paedagogik/article/Journal.html?tx_beltz_journal[article]=12169&cHash=40060aa0e47941d82c13f2962192f1df; http://nbn-resolving.de/urn:nbn:de:0111-pedocs-119245

Kunter, M., Kunina-Habenicht, O., Baumert, J., Dicke, T., Holzberger, D., Lohse-Bossenz, H. et al. (2017). Bildungswissenschaftliches Wissen und professionelle Kompetenz in der Lehramtsausbildung. In C. Gräsel & K. Trempler (Hrsg.), *Entwicklung von Professionalität pädagogischen Personals* (S. 37–54). Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-07274-2_3

Lauterbach, O. (2015). Erfassung wirtschaftswissenschaftlicher Fachkompetenzen von Studierenden in Startkohorte 5 des nationalen Bildungspanels - Technischer Bericht. *NEPS Working Papers* (51).

Linninger, C., Kunina-Habenicht, O., Emmenlauer, S., Dicke, T., Schulze-Stocker, F., Leutner, D. et al. (2015). Assessing Teachers' Educational Knowledge. Construct Specification and Validation Using Mixed Methods. *Zeitschrift für Entwicklungspsychologie und pädagogische Psychologie, 47* (2), 72–83. https://doi.org/10.1026/0049-8637/a000126

Lohse-Bossenz, H., Kunina-Habenicht, O., Dicke, T., Leutner, D. & Kunter, M. (2015). Teachers' knowledge about psychology. Development and validation of a test measuring theoretical foundations for teaching and its relation to instructional behavior. *Studies in Educational Evaluation, 44,* 36–49. https://doi.org/10.1016/j.stueduc.2015.01.001

Mashkovskaya, A. (2013). *Der C-Test als Lesetest bei Muttersprachlern*. Duisburg, Essen, Univ. Duisburg-Essen, Diss., 2013.

Meschede, N. (2013). *Professionelle Wahrnehmung der inhaltlichen Strukturierung im naturwissenschaftlichen Grundschulunterricht. Theoretische Beschreibung und empirische Erfassung* (info/1049163826/04; Verlagsangaben: http://deposit.d-nb.de/cgi-bin/dokserv?id=4619031&prov=M&dok_var=1&dok_ext=htm). Zugl.: Kiel, Univ., Diss., 2013.

Nouns, Z. M. & Georg, W. (2010). Progress testing in German speaking countries. *Medical Teacher, 32* (6), 467–470. https://doi.org/10.3109/0142159X.2010.485656

Peter, J., Leichner, N., Mayer, A.-K. & Krampen, G. (2015). A short test for the assessment of basic knowledge in psychology. *Psychology Learning & Teaching, 14* (3), 224–235. https://doi.org/10.1177/1475725715605763

Plöger, W. (2014). Wie gut können Lehrpersonen Unterricht analysieren? Ergebnisse eines empirischen Forschungsprojektes. *Pädagogische Rundschau, 68* (3), 273–287.

Richter, T., Naumann, J. & Horz, H. (2010). Eine revidierte Fassung des Inventars zur Computerbildung (INCOBI-R). *Zeitschrift für pädagogische Psychologie, 24* (1), 23–37.

Riese, J., Kulgemeyer, C., Zander, S., Borowski, A., Fischer, H. E., Gramzow, Y. et al. (2015). Modellierung und Messung des Professionswissens in der Lehramtsausbildung Physik. In S. Blömeke & O. Zlatkin-Troitschanskaia (Hrsg.)Kompetenzen von Studierenden, 55–79 [Themenheft].

Riese, J. & Reinhold, P. (2012). Die professionelle Kompetenz angehender Physiklehrkräfte in verschiedenen Ausbildungsformen. Empirische Hinweise für eine Verbesserung des Lehramtsstudiums. Paralleltitel: The professional competencies of trainee teachers in physics in different educational programs. *Zeitschrift für Erziehungswissenschaft, 15* (1), 111–143. Verfügbar unter http://dx.doi.org/10.1007/s11618-012-0259-y

Rosman, T., Mayer, A.-K. & Krampen, G. (2016). Measuring psychology students' information-seeking skills in a situational judgment test format: Construction and validation of the PIKE-P test. *European Journal of Psychological Assessment, 32* (3), 220–229.

Roters, B., Nold, G., Haudeck, H., Kessler, J.-U. & Stancel-Piqtak, A. (2011). Professionelles Wissen von Studierenden des Lehramts Englisch. In S. Blömeke, A. Bremerich-Vos, H. Haudeck, G. Kaiser, G. Nold, K. Schwippert et al. (Hrsg.), *Kompetenzen von Lehramtsstudierenden in gering strukturierten Domänen. Erste Ergebnisse aus TEDS-LT* (S. 77–100). Münster: Waxmann.

Rott, B., Leuders, T. & Stahl, E. (2015). Assessment of mathematical competencies and epistemic cognition of preservice teachers. *Zeitschrift für Psychologie, 223* (1), 39–46. https://doi.org/10.1027/2151-2604/a000198

Schaeper, H. (2013). The German National Educational Panel Study (NEPS). Assessing competencies over the life course and in higher education. In S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), *Modeling and Measuring Competencies in Higher Education. Tasks and Challenges* (pp. 147–158). Rotterdam: SensePublishers. Verfügbar unter http://link.springer.com/chapter/10.1007/978-94-6091-867-4_11

Schladitz, S., Rott, B., Winter, A., Wischgoll, A., Groß Ophoff, J., Hosenfeld, I. et al. (2013). LeScEd - learning the science of education research competence in educational sciences. In S. Blömeke & O. Zlatkin-Troitschanskaia (Eds.), *The German funding initiative 'Modeling and Measuring Competencies in Higher Education'. 23 research projects on engineering,*

economics and social sciences, education and generic skills of higher education students (KoKoHs Working Papers, Bd. 3, pp. 82–84). Verfügbar unter http://www.dzhw.eu/pdf/pub_art/22/2013_kokohs_wp3_bloemeke-zlatkin-troitschanskaia.pdf#page=85

Schladitz, S., Groß Ophoff, J. & Wirtz, M. (2015). Konstruktvalidierung eines Tests zur Messung bildungswissenschaftlicher Forschungskompetenz. In S. Blömeke & O. Zlatkin-Troitschanskaia (Hrsg.)Kompetenzen von Studierenden, 167–184 [Themenheft].

Schmelzing, S. (2010). *Das fachdidaktische Wissen von Biologielehrkräften: Konzeptionalisierung, Diagnostik, Struktur und Entwicklung im Rahmen der Biologielehrerbildung*. Berlin: Logos-Verl.***66930.

Schmidt, S., Brückner, S., Zlatkin-Troitschanskaia, O. & Förster, M. (2015). Das wirtschaftswissenschaftliche Wissen in der Hochschulbildung. Eine Analyse der messinvarianten Erfassung finanzwirtschaftlichen Fachwissens bei Studierenden. Paralleltitel: Knowledge of business and economics in higher education. *Empirische Pädagogik, 29* (1), 106–124.

Schnick-Vollmer, K., Berger, S., Bouley, F., Fritsch, S., Schmitz, B., Seifried, J. et al. (2015). Modeling the competencies of prospective business and economics teachers: Professional knowledge in accounting. *Zeitschrift für Psychologie, 223* (1), 24–30. https://doi.org/10.1027/2151-2604/a000196

Seidel, T. & Stürmer, K. (2014). Modeling and Measuring the Structure of Professional Vision in Preservice Teachers. *American Educational Research Journal, 51* (4), 739–771. https://doi.org/10.3102/0002831214531321

Seidel, T., Blomberg, G. & Stürmer, K. (2010). "Observer" - Validierung eines videobasierten Instruments zur Erfassung der professionellen Wahrnehmung von Unterricht. Projekt OBSERVE. In E. Klieme, D. Leutner & M. Kenk (Hrsg.), *Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes* (Zeitschrift für Pädagogik, Beiheft. 56, S. 296–306). Weinheim: Beltz. Verfügbar unter http://www.pedocs.de/volltexte/2010/3438; http://nbn-resolving.de/urn:nbn:de:0111-opus-34384

Seifert, A. & König, J. (2012). Pädagogisches Unterrichtswissen – bildungswissenschaftliches Wissen: Validierung zweier Konstrukte. In J. König & A. Seifert (Hrsg.), *Lehramtsstudierende erwerben pädagogisches Professionswissen. Ergebnisse der Längsschnittstudie LEK zur*

Wirksamkeit der erziehungswissenschaftlichen Lehrerausbildung (S. 215–233). Münster: Waxmann.

Seifert, A. & Schaper, N. (2010). Überprüfung eines Kompetenzmodells und Messinstruments zur Strukturierung allgemeiner pädagogischer Kompetenz in der universitären Lehrerbildung. Paralleltitel: Development and psychometrical testing of an instrument for the use of measuring pedagogical competence in universitary teacher education. *Lehrerbildung auf dem Prüfstand, 3* (2), 179–198.

Seifert, A. & Schaper, N. (2010). Welche Dimensionen bzw. Modellstruktur liegt der Messung pädagogischer Kompetenz in der universitären Lehrerausbildung zugrunde? In B. Schwarz, P. Nenniger & R. S. Jäger (Hrsg.), *Erziehungswissenschaftliche Forschung - nachhaltige Bildung. Beiträge zur 5. DGfE-Sektionstagung "Empirische Bildungsforschung" - AEPF-KBBB im Frühjahr 2009* (Erziehungswissenschaft, Bd. 28, S. 224–234). Landau: Verl. Empirische Pädag.

Seifert, A. & Schaper, N. (2012). Die Entwicklung von bildungswissenschaftlichem Wissen: Theoretischer Rahmen, Testinstrument, Skalierung und Ergebnisse. In J. König & A. Seifert (Hrsg.), *Lehramtsstudierende erwerben pädagogisches Professionswissen. Ergebnisse der Längsschnittstudie LEK zur Wirksamkeit der erziehungswissenschaftlichen Lehrerausbildung* (S. 183–214). Münster: Waxmann.

Stark, R., Herzmann, P. & Krause, U.-M. (2010). Effekte integrierter Lernumgebungen. Vergleich problembasierter und instruktionsorientierter Seminarkonzeptionen in der Lehrerbildung. Paralleltitel: Effects of integrated learning environments - a comparison between problembased and instruction-oriented seminar conceptions in teacher training. *Zeitschrift für Pädagogik, 56* (4), 548–563. Verfügbar unter http://www.pedocs.de/volltexte/2013/7159; http://nbn-resolving.de/urn:nbn:de:0111-opus-71591

Stender, A. (2014). *Unterrichtsplanung. Vom Wissen zum Handeln - Theoretische Entwicklung und empirische Überprüfung des Transformationsmodells der Unterrichtsplanung*. Kiel, Christian-Albrechts-Universität, Diss., 2014.

Stiller, J., Hartmann, S., Mathesius, S., Straube, P., Tiemann, R., Nordmeier, V. et al. (2016). Assessing scientific reasoning: a comprehensive evaluation of item features that affect item difficulty. *Assessment & Evaluation in Higher Education, 41* (5), 721–732.

Stürmer, K. (2011). *Voraussetzungen für die Entwicklung professioneller Unterrichtswahrnehmung im Rahmen universitärer Lehrerausbildung*. München, Technische Universität München, Diss., 2011.

Stürmer, K., Könings, K. D. & Seidel, T. (2013). Declarative knowledge and professional vision in teacher education: Effect of courses in teaching and learning. *British Journal of Educational Psychology, 83* (3), 467–483.

Stürmer, K. & Seidel, T. (2015). Assessing professional vision in teacher candidates: Approaches to validating the observer extended research tool. *Zeitschrift für Psychologie, 223* (1), 54–63. https://doi.org/10.1027/2151-2604/a000200

Stürmer, K., Seidel, T. & Kunina-Habenicht, O. (2015). Unterricht wissensbasiert beobachten. Unterschiede und erklärende Faktoren bei Referendaren zum Berufseinstieg. Paralleltitel: Knowledge-based classroom observation. *Zeitschrift für Pädagogik, 61* (3), 345–360. Verfügbar unter http://www.beltz.de/fachmedien/erziehungs_und_sozialwissenschaften/zeitschriften/zeitschrift_fuer_paedagogik/article/Journal.html?tx_beltz_journal[article]=30358&cHash=ea137ce8cb37461f3e1fdc269bf341f9

Taskinen, P. H., Steimel, J., Gräfe, L., Engell, S. & Frey, A. (2015). A competency model for process dynamics and control and its use for test construction at university level. *Peabody Journal of Education, 90* (4), 477–490. https://doi.org/10.1080/0161956X.2015.1068074

Thoma, G.-B., Dalehefte, I. M. & Köller, O. (2014). Entwicklung und Validierung eines Multiple-Choice-Tests zur Erfassung von Wissen über das menschliche Gehirn und Nervensystem. *Psychologie in Erziehung und Unterricht, 61* (3), 231–236. Verfügbar unter http://dx.doi.org/10.2378/peu2014.art18d

Tiede, J. & Grafe, S. (2016). Media Pedagogy in German and U.S. Teacher Education. *Comunicar, 24* (49), 19–28.

Trempler, K., Hetmanek, A., Kiesewetter, J., Wermelt, M., Fischer, F., Fischer, M. et al. (2015). Nutzung von Evidenz im Bildungsbereich. Validierung eines Instruments zur Erfassung von Kompetenzen der Informationsauswahl und Bewertung von Studien. In S. Blömeke & O. Zlatkin-Troitschanskaia (Hrsg.), *Kompetenzen von Studierenden* (Zeitschrift für Pädagogik : Beiheft, Bd. 61, S. 144–166). Weinheim: Beltz Juventa.

Türling, J. M. (2014). (Trainee) teachers' professional competence by learning from errors. An empirical study in accounting lessons. *Die professionelle Fehlerkompetenz von (angehenden) Lehrkräften. Eine empirische Untersuchung im Rechnungswesenunterricht*.

Vogelsang, C. (2014). *Validierung eines Instruments zur Erfassung der professionellen Handlungskompetenz von (angehenden) Physiklehrkräften. Zusammenhangsanalysen zwischen Lehrerkompetenz und Lehrerperformanz* (d-nb.de/cgi-bin/dokserv?id=5101371&prov=M&dok_var=1&dok_ext=htm). Zugl.: Paderborn, Univ., Diss., 2014.

Voss, T. & Kunter, M. (2011). Pädagogisch-psychologisches Wissen von Lehrkräften. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV* (S. 193–214). Münster: Waxmann.

Voss, T., Kunter, M. & Baumert, J. (2011). Assessing teacher candidates' general pedagogical/psychological knowledge. Test construction and validation. *Journal of Educational Psychology, 103* (4), 952–969. https://doi.org/10.1037/a0025125

Voss, T., Kunter, M., Seiz, J., Hoehne, V. & Baumert, J. (2014). Die Bedeutung des pädagogisch-psychologischen Wissens von angehenden Lehrkräften für die Unterrichtsqualität. *Zeitschrift für Pädagogik, 60* (2), 184–201.

Weinhardt, M. & Kelava, A. (2016). Die performanzorientierte Erfassung psychosozialer Beratungskompetenz in Forschung und Lehre im Rahmen einer Simulationsumgebung. *Neue Praxis, 46* (4), 363–377.

Weresch-Deperrois, I. (2014). *Entwicklung eines standardorientierten Situational Judgement Tests zur Erfassung professioneller pädagogischer Kompetenz*. Frankfurt am Main, Johann Wolfgang Goethe-Univ., Diss., 2014.

Wiesbeck, A. B. (2015). Simulierte Gespräche als ein Assessment zur Messung der Gesprächsführungskompetenz Lehramtsstudierender in Lehrer-Elterngesprächen - Eine Validierungsstudie. *An Evaluation of Simulated Conversations as an Assessment of Pre-Service Teachers' Communication Competence in Parent-Teacher Conversations*. München, Technische Universität München, Diss., 2015.

Winter-Hölzl, A., Wäschle, K., Wittwer, J., Watermann, R. & Nückles, M. (2015). Entwicklung und Validierung eines Tests zur Erfassung des Genrewissens Studierender und Promovierender der Bildungswissenschaften. In S. Blömeke & O. Zlatkin-Troitschanskaia (Hrsg.)Kompetenzen von Studierenden, 185–202 [Themenheft].

Wischgoll, A. (2016). Combined Training of One Cognitive and One Metacognitive Strategy Improves Academic Writing Skills. *FRONTIERS IN PSYCHOLOGY, 7*. https://doi.org/10.3389/fpsyg.2016.00187

Woitkowski, D. (2015). *Fachliches Wissen Physik in der Hochschulausbildung. Konzeptualisierung, Messung, Niveaubildung* (info/1072090864/04; Verlagsangaben: http://deposit.d-nb.de/cgi-bin/dokserv?id=5288832&prov=M&dok_var=1&dok_ext=htm). Zugl.: Paderborn, Univ., Diss., 2015.

Woitkowski, D., Riese, J. & Reinhold, P. (2011). Modellierung fachwissenschaftlicher Kompetenz angehender Physiklehrkräfte. Paralleltitel: Modelling Content Knowledge of Prospective Physics Teachers. *Zeitschrift für Didaktik der Naturwissenschaften, 17,* 289-313; 1,2 MB. Verfügbar unter http://archiv.ipn.uni-kiel.de/zfdn/pdf/17_Woitkowski.pdf

Wolter, F. & Schiener, J. (2014). En Route to “University Pisa”? On the Measurement of Sociological Competencies. *SOZIALE WELT-ZEITSCHRIFT FUR SOZIALWISSENSCHAFTLICHE FORSCHUNG UND PRAXIS, 65* (1), 47+.

Wüstenberg, S., Greiff, S. & Funke, J. (2012). Complex problem solving — More than reasoning? *Intelligence, 40* (1), 1–14. https://doi.org/10.1016/j.intell.2011.11.003

Wuttke, E. & Seifried, J. (2013). Diagnostic competence of (prospective) teachers in vocational education. An analysis of error identification in accounting lessons. In K. Beck & O. Zlatkin-Troitschanskaia (Eds.), *From Diagnostics to Learning Success. Proceedings in Vocational Education and Training* (Professional and VET learning, vol. 1, pp. 225–240). Rotterdam: SensePublishers.

Zlatkin-Troitschanskaia, O., Förster, M., Brückner, S., Hansen, M. & Happ, R. (2013). Modellierung und Erfassung der wirtschaftswissenschaftlichen Fachkompetenz bei Studierenden im deutschen Hochschulbereich. In O. Zlatkin-Troitschanskaia, R. Nickolaus & K. Beck (Hrsg.)Kompetenzmodellierung und Kompetenzmessung bei Studierenden der Wirtschaftswissenschaften und der Ingenieurwissenschaften, 108–133 [Themenheft]. Landau: Empirische Pädagogik.

Zlatkin-Troitschanskaia, O., Happ, R., Förster, M., Preuße, D., Schmidt, S. & Kuhn, C. (2013). Analyse der Ausprägung und Entwicklung der Fachkompetenz von Studierenden der Wirtschaftswissenschaften und der Wirtschaftspädagogik. In O. Zlatkin-Troitschanskaia, R. Nickolaus & K. Beck (Hrsg.)Kompetenzmodellierung und Kompetenzmessung bei Studierenden der Wirtschaftswissenschaften und der Ingenieurwissenschaften, 69–92 [Themenheft]. Landau: Empirische Pädagogik.

Zlatkin-Troitschanskaia, O., Förster, M., Brückner, S. & Happ, R. (2014). Insights from a German assessment of business and economics competence. In H. Coates (Hrsg.), *Higher Education*

Learning Outcomes Assessment. International Perspectives (Higher Education Research and Policy, Bd. 6, S. 175–197). Frankfurt am Main: Peter Lang GmbH.

Zylka, J. (2012). *Medienkompetenzen und Instrumente zu ihrer Messung. Entwicklung eines Wissenstests zu informationstechnischem Wissen von Lehrkräften, Lehramtsanwärtern und Lehramtsstudierenden* (d-nb.de/cgi-bin/dokserv?id=4362768&prov=M&dok_var=1&dok_ext=htm; Inhaltsverzeichnis: http://d-nb.info/1035786354/04). Zugl.: Weingarten, Pädag. Hochsch., Diss., 2012.

C. Übersicht der nicht eindeutig zuordenbaren Textstellen aus der Kodierung

a. Textstellen, die Kategorie 1 und 2 zugeordnet wurden

Quelle	Textstelle
Buchholtz (2014), S. 92	Welche strukturellen Zusammenhänge bestehen zwischen dem Fachwissen und dem fachdidaktischem Wissen von Lehramtsstudierenden im Bachelor und im Masterstudium und lassen sich diese Zusammenhänge bzw. deren Veränderungen erklären?
Buchholtz & Kaiser (2013), S. 135	Zeigen sich im Master-/ Hauptstudium ähnliche Zusammenhänge zwischen den einzelnen Wissensdimensionen wie im Bachelor-/Grundstudium?
Förster & Zlatkin-Troitschanskaia (2010), S. 108	Um die Ausprägung und Entwicklung der (kognitiven) Lehrprofessionalitätsdimensionen (fach- und fachdidaktische Kompetenz) bei den Studierenden des neuen konsekutiven Bachelor-Master-Modells für Wirtschaftspädagogik („Interventionsgruppe") im Zeitverlauf zu untersuchen, werden Längsschnitterhebungen zu drei Zeitpunkten durchgeführt
Großschedl et al. (2014), S. 458	Es soll untersucht werden, wie sich die Beziehungen zwischen den Bereichen des Professionswissens im Studienverlauf entwickeln.
Heigl (2014), S. 68	Neben der empirischen Klärung des Konstrukts steht deshalb auch die Suche nach Fördermaßnahmen im Zentrum der Diagnostik fächerübergreifenden Problemlösens.
Hohenstein (2015), S. 28	Die vorliegende Arbeit hat das Ziel, einen Wissenstest zur Erfassung von pädagogisch-psychologischem Wissen von Lehramtsstudierenden zu entwickeln und zu validieren. Vor der Konstruktion eines solchen Tests stellt sich die Frage, welche konkreten Wissensbestände die Lehramtsstudierenden in ihrer universitären Ausbildung erwerben können und welche Lerngelegenheiten an den Hochschulen angeboten werden.
Kehne, Seifert & Schaper (2013), S. 54	Welche professionellen Kompetenzen und Überzeugungen entwickeln Studierende in der ersten Phase der Lehrerausbildung?
Kirschner (2013), S. 50	FF4: Wie hängen CK und PCK bei Lehramtsstudierenden der Physik, Lehrkräften der Physik im Vorbereitungsdienst und Lehrkräften der Physik zusammen? H4: Die Korrelationen unterscheiden sich nicht zwischen den Gruppen.
Kleickmann et al. (2014), S. 281	Wie ist das professionelle Wissen, das im Studium erworben wird, bei Lehramtskandidatinnen und -kandidaten strukturiert? Bewährt sich hier auch die Unterteilung in Fachwissen, fachdidaktisches Wissen und pädagogisch/psychologisches Wissen oder zeigen sich andere Strukturen? Wie stark korrelieren die Bereiche?
König (2010), S. 64	Zeigen sich Unterschiede in der Struktur des pädagogischen Wissens zwischen den beiden Befragungen?
König (2012), S. 180)	Gleichzeitig stellt der empirische Beleg für den Wissenszuwachs der Studierenden, welcher an allen Standorten gleichermaßen mit großer praktischer Bedeutsamkeit belegbar ist, den zentralen Hinweis dafür dar, dass das Testinstrument standortübergreifend curricular valide ist. Offensichtlich misst der Test Inhalte, die zu einem Kerncurriculum zählen, welches gelehrt und studiert wird.
König & Blömeke (2010), S. 269	Entsprechende Erkenntnisse tragen sowohl zur Klärung theoretischer Grundfragen der Erziehungswissenschaft und Bildungsforschung als auch zur Klärung der Wirksamkeit von Lehrerausbildung bei.

Fortsetzung auf der nächsten Seite

Quelle	Textstelle
König, Blömeke & Doll (2011), S. 152	Ferner ist die Tendenz erkennbar, dass mit zunehmender fachlicher Spezialisierung der angehenden Lehrkräfte der Zusammenhang zwischen ihrem Wissen in Pädagogik und ihrem Fachwissen niedriger ausfällt. Bei geringer fachlicher Spezialisierung während der Ausbildung dürfte sich der Wissenserwerb der Lehrkräfte in den Bereichen Pädagogik, Fach und Fachdidaktik also vermutlich stärker ausbalanciert vollziehen.
Kotzebue (2014), S. 5	Übergeordnetes Ziel dieser Dissertation ist es daher, einen Beitrag zur Aufklärung der Kompetenzstruktur zum fachlichen und fachdidaktischen Umgang mit Diagrammen durch Lehrkräfte zu leisten, um damit die Voraussetzungen für die Optimierung der Lehrerausbildung im Bereich der Diagrammkompetenz zu klären und Implikationen für die Einbindung dieser Thematik in das Curriculum der Lehrerausbildung zu erhalten.
Kunina-Habenicht et al. (2013), S. 9	Der standardisierte Test zur Erfassung des bildungswissenschaftlichen Wissens, der in dieser Studie zum Einsatz kam, wurde im Rahmen des BilWiss-Projekts entwickelt. Hierbei steht nicht das Handeln der Lehrkräfte in der Praxis im Mittelpunkt, sondern konzeptuellanalytisches Wissen, welches im universitären Lehramtsstudium aufgebaut werden soll.
Kunter et al. (2017), S. 42	Welches bildungswissenschaftliche Wissen wird am Ende des Lehramtsstudiums erreicht?
Riese et al. (2015) [Abstract]	Die Ziele des vorgestellten Projekts liegen in der Modellierung und Messung domänenspezifischer und generischer Kompetenzen, die Lehramtsstudierende der Physik im Hochschulstudium erwerben sollen.
Riese et al. (2015), S. 56	Für eine Analyse des Professionswissens von Lehramtsstudierenden ist aber eine differenziertere Modellierung wünschenswert, die z. B. die innere Struktur von Fachwissen und fachdidaktischem Wissen und die Beziehung zwischen diesen Wissensbereichen abbilden kann, um damit z. B. die Wirkung unterschiedlicher Lerngelegenheiten und längsschnittliche Entwicklungsverläufe untersuchen zu können.
Rott, Leuders & Stahl (2015), S. 41	We expect that we can distinguish between students' epistemic orientations (certain vs. uncertain) on the one hand and the sophistication of their argumentations on the other hand reflecting the reported discussion on context specificity of epistemic judgments. This structure should be seen more clearly in 4th semester students than in 1st semester students, due to the influence of reflective elements of courses in mathematics and mathematics education.
Schladitz et al. (2015), S. 83	Subsequently, in two main experiments we will analyze how the psychometric properties are affected by training in the previously mentioned content areas.
Schmelzing (2010), S. 124	Neben der Diagnostik des fachdidaktischen Wissens stellt sich aus Sicht der Biologielehrerbildung die grundlegende Frage wie sich das fachdidaktische Wissen von Biologielehrkräften entwickelt und wie sich diese Entwicklungsprozesse im Rahmen der Biologielehrerbildung unterstützen lassen.
Seidel & Stürmer (2014) [Abstract]	Professional vision has been identified as an important element of teacher expertise that can be developed in teacher education.
Seidel & Stürmer (2014), S. 3	Quantitative instruments as the one presented and proposed in this article allow more efficient data analysis. Such a measure can provide a first indicator regarding, for example, the current state of teacher competencies. These indicators can be used promptly for feedback on teaching and formative assessment.
Seifert & Schaper (2010), S. 181	Vor diesem Hintergrund kann die Messung professioneller Handlungskompetenz von schulischen Lehrkräften oftmals zugleich auch zur Evaluation eines Lehrer ausbildenden Curriculums bzw. bestimmter Teile eines entsprechenden Curriculums angelegt werden.
Seifert & Schaper (2010), S.193f	Zentrale Fragestellung dieser Studie war, inwieweit die Struktur des entwickelten Instruments zur Erfassung allgemeiner pädagogischer Kompetenz die Struktur des Kompetenzerwerbs im erziehungswissenschaftlichen Studium widerspiegelt.

Fortsetzung auf der nächsten Seite

Quelle	Textstelle
Stender (2014), S. 111	Inwieweit unterscheiden sich die Ausprägungen der formalen und funktionalen Merkmale von Handlungsskripten mit zunehmender Erfahrung von Lehrpersonen?
Stender (2014), S. 117	Inwieweit verändert sich der Einfluss des Professionswissens von Lehrpersonen auf die Qualität der Handlungsskripte mit zunehmender Erfahrung der (angehenden) Lehrpersonen?
Stürmer (2011), S. 7	Zum anderen zeigt sich die Notwendigkeit, empirisch valide Messverfahren zu entwickeln (Cochran-Smith, 2003; Darling-Hammond, 2006; Desimone, 2009; Jahn, Prenzel, Stürmer, & Seidel, in Druck), um professionelle Kompetenz abbilden und ihre Entwicklung im Rahmen universitärer Lehrausbildung untersuchen zu können.
Tiede & Grafe (2016), [Abstract]	To understand, assess and eventually improve the status of media pedagogical teacher education, comprehensive research is required.
Türling 82014), S. 25	Neben der Modellierung und Messung der PFK zielt das Projekt auf die Identifizierung und Klassifizierung domänenspezifischer Fehlerarten (Buchführung) sowie eine Analyse der Kompetenzentwicklung von (angehenden) Lehrkräften im Verlauf der Lehrerausbildung (Universität, Referendariat, Berufseintritt) ab.
Weinhardt & Kelava (2016), S. 364	Um einige dieser Probleme in der Erfassung von Handlungskompetenzen zu vermeiden, stellt der vorliegende Artikel einen Ansatz zur performanzorientierten Erfassung psychosozialer Beratungskompetenz vor, der sowohl für den Einsatz in der Forschung (z.B. hinsichtlich der Modellierung von Beratungskompetenzerwerbsprozessen) als auch in der Lehre (z.B. für die Erstellung teilnehmerspezifischer Lernmaterialien) verwendet werden kann.
Weresch-Deperrois (2014), S. 74	Unter dieser Prämisse kann der intendierte Situational Judgement Test – sofern sich der Test als valide zeigt – als diagnostisches Instrument gelten, weil hierdurch die Kompetenz angehender Lehrpersonen aufgrund eines Sammelns, Aufbereitens und Verarbeitens von Informationen diagnostiziert und gegebenenfalls adäquate Handlungen (wie etwa ein verpflichtendes zusätzliches Selbststudium) eingeleitet werden könnten.
Woitkowski (2015), S. 137	Das fachliche Wissen von Physikstudenten (Fach - und Lehramtsstudiengänge) lässt sich durch die drei distalen Merkmale Abiturnote, Geschlecht und Studienfortschritt, gemessen in Semesterwochenstunden in belegten Fachveranstaltungen, gut voraussagen.
Woitkowski (2015), S. 7	Perspektivisch sollten die damit erhaltenen Daten als Grundlage für die Entwicklung strukturierter und adaptiver Unterstützungsmaßnahmen oder die Erfassung komplexerer, auf dem fachlichen Wissen aufbauender Kompetenzen nutzbar sein.
Zlatkin-Troitschanskaia et al. (2013), S. 77	Um neben der Ausprägung insbesondere auch die Entwicklung von Studierenden in den ausgewählten Kompetenzaspekten analysieren zu können, wurde die ILLEV-Studie als Längsschnittstudie angelegt.

b. Textstellen, die Kategorie 1 und 3 zugeordnet wurden

Quelle	Textstelle
Kirschner (2013), S. 31	Für eine Validierung und zur Klärung der Relevanz des Konstrukts für Unterricht muss erforscht werden, welche Einflüsse CK, PCK und PK auf das Handeln der Lehrkräfte sowie auf Schülerleistung und –motivation haben. Ist der Zusammenhang gering, stellt dies die Validität des Konstrukts in Frage.
König et al. (2014), S. 77	Against this background, in this article we empirically investigate the question of how and to what extent the declarative conceptual GPK of early career (i.e., four years teaching experience or less) mathematics middle school teachers can be understood as a premise for their skills to notice and interpret pedagogical situations in a mathematics classroom presented to them via video-vignettes.
Kotzebue (2014), S. 14	Innerhalb dieser Schwerpunktsetzung wurden anhand des theoretischen Hintergrunds zur Diagrammkompetenz und zum Professionswissen Komponenten erarbeitet, über die (angehende) Biologielehrkräfte verfügen müssen, um sach- und adressatengerecht mit Diagrammen umgehen und im Unterricht biologische Inhalte vermitteln zu können.
Linninger et al. (2015), S. 80f	However, this increased insight on the characteristics of EK does not diminish the validity of the test's application within our research program, since the aim of this longitudinal study – and the reason for the development of the test in the first place – is, in fact, to analyze longitudinal effects of EK on other aspects of teachers' competence as well as on their practice at school. S. 80f
Lohse-Bossenz et al. (2015), S. 39f	The main aim of this study, therefore, is to describe the development of a test which measures teacher candidates' psychological knowledge at the end of university, and to obtain first evidence on the relationship between theoretical knowledge and professional behavior.
Vogelsang (2014), S. 280	Bilden die Kennwerte des Paderborner Instruments zur Erfassung professioneller Handlungskompetenz (angehender) Physiklehrkräfte Prädiktoren für die Performanz von (angehenden) Physiklehrkräften im Physikunterricht?

c. Textstellen, die Kategorie 2 und 3 zugeordnet wurden

Quelle	Textstelle
Jahn (2014), S. 50	Um die Erfassung professioneller Unterrichtswahrnehmung im Verlauf der universitären Lehrerbildung über das Abbilden von integrierten Wissensstrukturen hinaus zu rechtfertigen, ist ein Zusammenhang zum späteren professionellen Handeln im Unterricht entscheidend. Dieser Aspekt gewinnt noch weiter an Bedeutung, wird der nächste Schritt von der Erfassung hin zur systematischen Förderung professioneller Unterrichtswahrnehmung unternommen.
Lauterbach (2015), S. 7	[Das als Online-Befragung durchgeführte Expertenrating diente der Prüfung der Aufgaben hinsichtlich ihrer Eignung,]… die Relevanz entsprechender Kompetenzen für Studium und Beruf zu erfassen.
Stürmer & Seidel (2015), S. 57	However, the suitability of the Observer as a tool for assessing professional vision within the theory-based (university) and practice-based (induction at schools) phases of teacher education has yet to be proven.
Wiesbeck (2015), S. 1	Hence, there is a need for instruments that can assess whether training measures are effective and whether pre-service teachers are adequately prepared for communicating successfully with parents.

d. Textstellen, die den Kategorien 1, 2 und 3 zugeordnet wurden

Quelle	Textstelle
Linninger et al. (2015), S. 73	Thus, EK is supposed to provide the theoretical subject-unspecific foundation for teachers' professional behavior both inside and outside the classroom and should therefore be fostered by academic teacher education.
Lohse-Bossenz et al. (2015), S. 39	To sum up, we follow Zeichner (2005) who recommended the development of tests to assess teacher knowledge, and add that such tests should maximize individual variability in test scores in order to sufficiently explore the relationships between teacher education programs, teachers' professional knowledge, and teachers' professional behavior (Borko, 2004).
Wuttke & Seifried (2013), S. 227	An essential part of and a first step in supporting learning from errors is a teacher's ability to identify or diagnose the errors made by students. Next, he/she needs adequate strategies to handle the errors and to give feedback. Furthermore, the teacher's belief system with regard to the possibilities and constraints of error learning is important (Figure 1; Türling, Seifried & Wuttke, 2012). We assume that teachers can develop these competences in the course of their training and professional life.

D. Übersicht der Textstellen, die nicht in das Validierungsschema eingeordnet werden konnten

Quelle	Textstelle	Anmerkungen
Hendler et al. (2011), S. 525	Eine Forschungsfrage, die in der in diesem Artikel beschriebenen Studie nachgegangen wird, ist deshalb, ob sich Anfängerinnen in der Erzieherinnenausbildung (Fachschule vs. Hochschule) hinsichtlich ihres sprachbezogenen Wissens bereits zu Beginn ihrer Fachschulausbildung bzw. ihres Studiums unterscheiden.	Hinweis auf Angebots-Nutzungs-Modell
König & Herzmann (2011), S. 187	In welchem Umfang beispielsweise unterscheiden sich angehende Lehrkräfte bereits am Anfang ihres Studiums in bereichsspezifischen Lernvoraussetzungen? Haben Unterschiede in solchen Lernvoraussetzungen differenzielle Entwicklungen in der Lehrerausbildung zur Folge?	Hinweis auf Angebots-Nutzungs-Modell
Kuhn (2014), S. 116	Der zuletzt genannte Aspekt wird im Folgenden als vernachlässigbar erachtet, da mit dieser Studie derzeit rein grundlagenorientierte Zwecke verfolgt werden, die keine direkte Verwertung des Tests bzw. der Testergebnisse (bspw. zur Feststellung der Eignung für den Lehrerberuf) vorsehen.	Aussage: Testwerte haben keine Konsequenzen
Kunina-Habenicht et al. (2013), S. 9	Die interindividuellen Unterschiede in der Nutzung der universitären Lerngelegenheiten.	Hinweis auf Angebots-Nutzungs-Modell
Peter et al. (2015), S. 7	It should be noted that, despite good reliability, this test should not be used for individual high stakes decisions as the number of items is too low and reliability scores do not justify an application in individual diagnostics.	Aussage zu Konsequenzen: nicht für Individualdiagnostik verwenden Im Sinne der Standards: Ausschluss von Testeinsätzen
Vogelsang (2014), S. 134f	Daher kann eine Stichprobe zur Überprüfung der Validität beliebige Lehrpersonen in allen Ausbildungsphasen umfassen, da nach dieser Annahme nur das Vorhandensein dieses Wissens entscheidend für das Handeln mit ‚guter' Qualität sein sollte. Allerdings werden die meisten Testverfahren zur Kompetenzmessung mit dem Ziel der Evaluation von Lehrerausbildungsprogrammen verwendet, weshalb tendenziell eher angehende Lehrkräfte betrachtet werden sollten, wobei auch hier die genaue Wahl von der konkreten Kompetenzmodellierung abhängt.	betrifft allgemeine Aussage zu Testverfahren, nicht das im Text behandelte Instrument
Wolter & Schiener (2014), S. 67	Sollte sich dieser Befund in weiteren Studien bestätigen, wäre dies ein wichtiges Argument für die Debatte, ob Jobben neben dem Studium den Studienerfolg – mutmaßlich aus Zeitgründen – beeinträchtigt.	Es werden außerhochschulische Bedingungsfaktoren für Studienerfolg angesprochen. Wieder Hinweis auf Angebots-Nutzungs-Modell

E. Übersicht der Validitätsevidenzen für Texte, die eine Testwertinterpretation nennen

Quellen von Validitätsevidenzen nach AERA, APA & NCME (2014)

Quelle	Kodierung TWI	inhaltsbasiert	Antwortprozesse	interne Struktur	Zusammenhang zu anderen Variablen	Zusammenhang Test - Kriterium
Blömeke et al. (2017), S. 340	Konstrukt	Expertenurteile zur Repräsentativität des Konstruktes und Power, Unterschiede im Antwortverhalten vorherzusagen und zu erklären		Modellvergleiche + MG-CFA metrische Invarianz der Struktur über Subgruppen	Zusammenhang Testwert - hochschulische Lerngelegenheiten in verschiedenen Ausbildungsarten -zu schulischer Leistung in Mathematik	Zusammenhang Testwert -Fähigkeit Lehrer-Kind Interaktionen in KiTa wahrzunehmen (erfasst über Videos von typischen Situationen)
Brückner et al. (2015) S. 442	Konstrukt			Passung zu IRT-Modell	Einfluss von schulischem Vorwissen auf Testwert	
Brückner & Pellegrino (2016); S. 294	Konstrukt		Think-Aloud um konstruktrelevante von irrelevanten Bearbeitungsprozessen zu trennen			
Gramzow (2015), S. 156	Konstrukt, Lehre	Experteneinschätzung zur Angemessenheit der Items bzgl. Konstrukt Prüfung der curicularen Passung	Methode des Lauten Denkens zur Identifikation von Wissensquellen für Bearbeitung der Items	vier fachdidaktische Facetten: CFA, Modellvergleiche	Abgrenzung zu anderen Wissensbereichen Wissen primär an der Universität erwerbbar?: Testwerte - Fortschritt hochschulische fachdidaktische Ausbildung	
Hammer et al. (2015), S. 39	Berufliche Anforderungen				Abgrenzung von DaZ zu - linguistischem, pädagogischem und mathematikdidaktischem Wissen; Zusammenhang von Testwert zu Studierendenmerkmalen und universitären Lerngelegenheiten	

Fortsetzung auf der nächsten Seite

Quellen von Validitätsevidenzen nach AERA, APA & NCME (2014)

Quelle	Kodierung TWI	inhaltsbasiert	Antwortprozesse	interne Struktur	Zusammenhang zu anderen Variablen	Zusammenhang Test - Kriterium
Jahn (2014), S. 49	Berufliche Anforderungen	realitätsnähere Itemformate: Videoclips		CFA, Modellvergleiche	Zusammenhang Testwert - Erhebungsbedingungen (Generalisierbarkeit) -universitäre Lerngelegenheiten	für Prädiktion Studien-Berufserfolg: Zusammenhang Testwert und Variablen, die „typischerweise als Prädiktoren für Studien- und Berufserfolg herangezogen werden"
Kuhn (2014), S. 7 + 114	Konstrukt	Experteneinschätzung zur inhaltlichen Passung und Angemessenheit für verschiedene Stufen von Ausbildungsfortschritt	Kognitive Interviews (Lautes Denken) zur Identifikation von Lösungswegen	Passung zu IRT-Modell CFA, Modellvergleiche DIF-Analysen	Zusammenhang Testwert – (vor-) hochschulische Lerngelegenheiten - andere wirtschaftswissenschaftliche Tests	
Kuhn et al. (2014), S. 161	Konstrukt			CFA, Korrelationen von 3 Tests als Nachweis, dass Dimensionen trennbar sind	Zusammenhang Testwert - Abiturnote -Leistungskurs (Indikator für Vorwissen) -Anzahl hochschulischer Lerngelegenheiten	
Riese et al. (2015), S. 68	Lehre	Experteneinschätzung, ob Testinhalte für Physik-studierende während des gesamten Studiums sind; Lehrbuchanalysen		Modellvergleiche in IRT	Zusammenhang Testwert-Studienfortschritt, -weitere Fähigkeitstests, -Selbsteinschätzungen zu Interesse, Motivation	
Stiller et al. (2016), S. 8	Konstrukt			Analyse von Faktoren, die Itemschwierigkeit beeinflussen		

Fortsetzung auf der nächsten Seite

Quellen von Validitätsevidenzen nach AERA, APA & NCME (2014)

Quelle	Kodierung TWI	inhaltsbasiert	Quelle	Kodierung TWI	inhaltsbasiert	Quelle
Vogelsang (2014), S. 494	Konstrukt, berufliche Anforderungen	Experteneinschätzung zu inhaltlicher Passung der Items		CFA, Modellvergleiche	Zusammenhang Testwert und über Interview erhobene Werte zu Belief System	Zusammenhang von Testwert zu beruflichem Handeln
Wiesbeck (2015), S. 84	Berufliche Anforderungen		Analysen von Antwortprozesse um Authentizität einzuschätzen	CFA, Modellvergleich Korrelation von "Testhälften" (verschiedene Videovignetten und verschiedene Schauspieler): Hinweis auf Generalisierbarkeit	Zusammenhang Testwert - Maße aus anderen Quellen für gleiches Konstrukt; -externe Kriterien: Abiturnote, Fachsemester, Vorwissen, selbst-berichtete Kommunikationskompetenz	

F. Beispielitems der Leistungstests

a. BWL-Test

Was ist der Unterschied zwischen Brutto- und Nettoinvestitionen?

- Bei der Ermittlung der Bruttoinvestitionen wird die Mehrwertsteuer berücksichtigt.
- Während die Ersatzinvestitionen bei der Bestimmung der Bruttoinvestitionen berücksichtigt werden, werden diese in die Ermittlung der Nettoinvestitionen nicht einbezogen.
- Die Bruttoinvestitionen des Inlands sind stets kleiner als die Nettoinvestitionen im Ausland.
- Die Nettoinvestitionen bzw. Bruttoinvestitionen geben den im Inland bzw. Ausland investierten Betrag eines Unternehmens an.

b. NagP-Test

Zu den ökologischen Zielen der Nachhaltigkeit gehört…

- Atomenergie durch Kohlekraftwerke ersetzen.
- die Entwicklung neuer schädlingsresistenter Pflanzenarten.
- die Sicherung des Gesundheitszustandes der Weltbevölkerung.
- der Erhalt der Artenvielfalt.

c. dNCM-Test

In welchem Maße werden die ökonomische, ökologische und soziale Zieldimension im klassischen Triple-Bottom-Line-Ansatz berücksichtigt?

- Die ökologische und soziale Zieldimension wird nur gemäß den gesetzlichen Anforderungen berücksichtigt.
- Wegen der besonderen globalen Bedeutung der ökologischen Dimension wird diese im Vergleich zur ökonomischen und sozialen Dimension in stärkerem Maße berücksichtigt.
- Bei diesem Ansatz werden die Zieldimensionen grundsätzlich in gleichem Maße berücksichtigt.
- Weil die Eigenkapitalgeber in besonderem Maße das unternehmerische Risiko tragen, wird die ökonomische Zieldimension bei diesem Ansatz vorrangig berücksichtigt.

d. sNCM-Test

Situation 10 (Transport) Item 1

Liebe Michaela,

um die Belieferung unseres Kunden Best Bik'O in Basel wie geplant ausführen zu können, bitte ich Dich, die ins Auge gefassten Transportmittel miteinander zu vergleichen. Für die wählbaren Transportrouten habe ich bereits eine Vorauswahl getroffen, die ich als Anhang beifüge. Die Menge beläuft sich auf eine Containerladung.

Vielen Dank und herzliche Grüße
Thomas

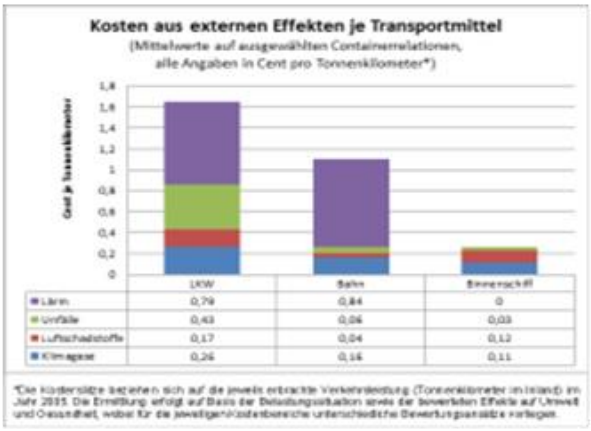
Hier finden Sie die von Thomas bereitgestellten Dokumente:
Übersicht mit den wählbaren **Transportrouten (Anlage I)**
graphische Gegenüberstellung der **durchschnittlichen Transportkosten (Anlage II)**
Übersicht der **durchschnittlichen Kosten aus externen Effekten (Anlage III)**
Thomas bittet Michaela, die Dokumente auszuwerten und ihm auf Basis einer Nutzwertanalyse per E-Mail mitzuteilen, welches Transportmittel sie empfiehlt. Weitere zentrale nicht-monetäre Faktoren sollen mit einem Gewicht von 20 % berücksichtigt werden und sind mit diesem Gewicht bereits in der Analyse enthalten. Das Gesamtgewicht muss dem Wert 1 entsprechen.

Bitte verfassen Sie die Antwort-E-Mail.

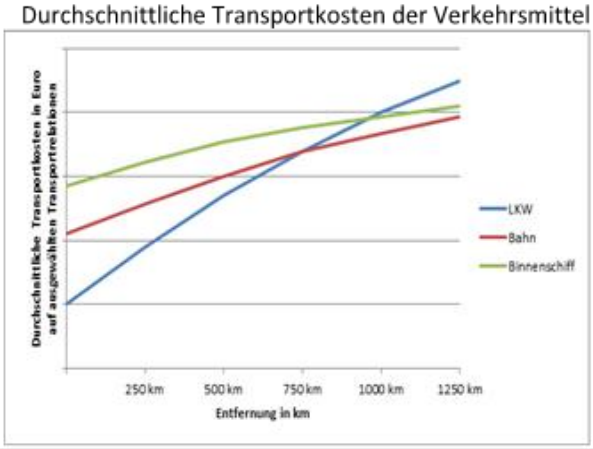
Anlagen

Übersicht der Transportrouten

Durchschnittliche Kosten aus externen Effekten



Durchschnittliche Transportkosten der Verkehrsmittel



Lieber Thomas,

auf Basis meiner Analyse weist der Transport per
[Drop-Down-Menü]
a) Bahn
b) LKW
c) Binnenschiff
den höchsten Nutzen für das Unternehmen auf. Die Kriterien habe ich hierbei wie folgt gewichtet:

[Eintragen der Gewichte in die Tabelle, alle anderen Werte sind vorgegeben]

Kriterien	Gewicht	LKW		Bahn**		Binnenschiff**	
		Punkte	Teilnutzwert	Punkte	Teilnutzwert	Punkte	Teilnutzwert
Frachtkosten*	[ ]	4	1,60	3	1,20	1,5	0,60
Externe Effekte	[ ]	1,5	0,60	3	1,20	4	1,60
Nicht-monetäre Kriterien:	0,20		0,60		0,65		0,45
Schnelligkeit	0,05	3	0,15	4	0,20	1	0,05
Pünktlichkeit/Zuverlässigkeit	0,10	2	0,20	3	0,30	3	0,30
Flexibilität	0,05	5	0,25	3	0,15	2	0,10
Gesamtwert	1,00		[ ]		[ ]		2,65

*Die Punktevergabe für die Transportkosten bezieht sich auf die Strecke Göttingen - Basel.
**Die Transportkosten für die Anlieferung und Abholung zum/am Binnenhafen bzw. Güterbahnhof sind bereits einkalkuliert und spiegeln sich in der Punktevergabe entsprechend wider.

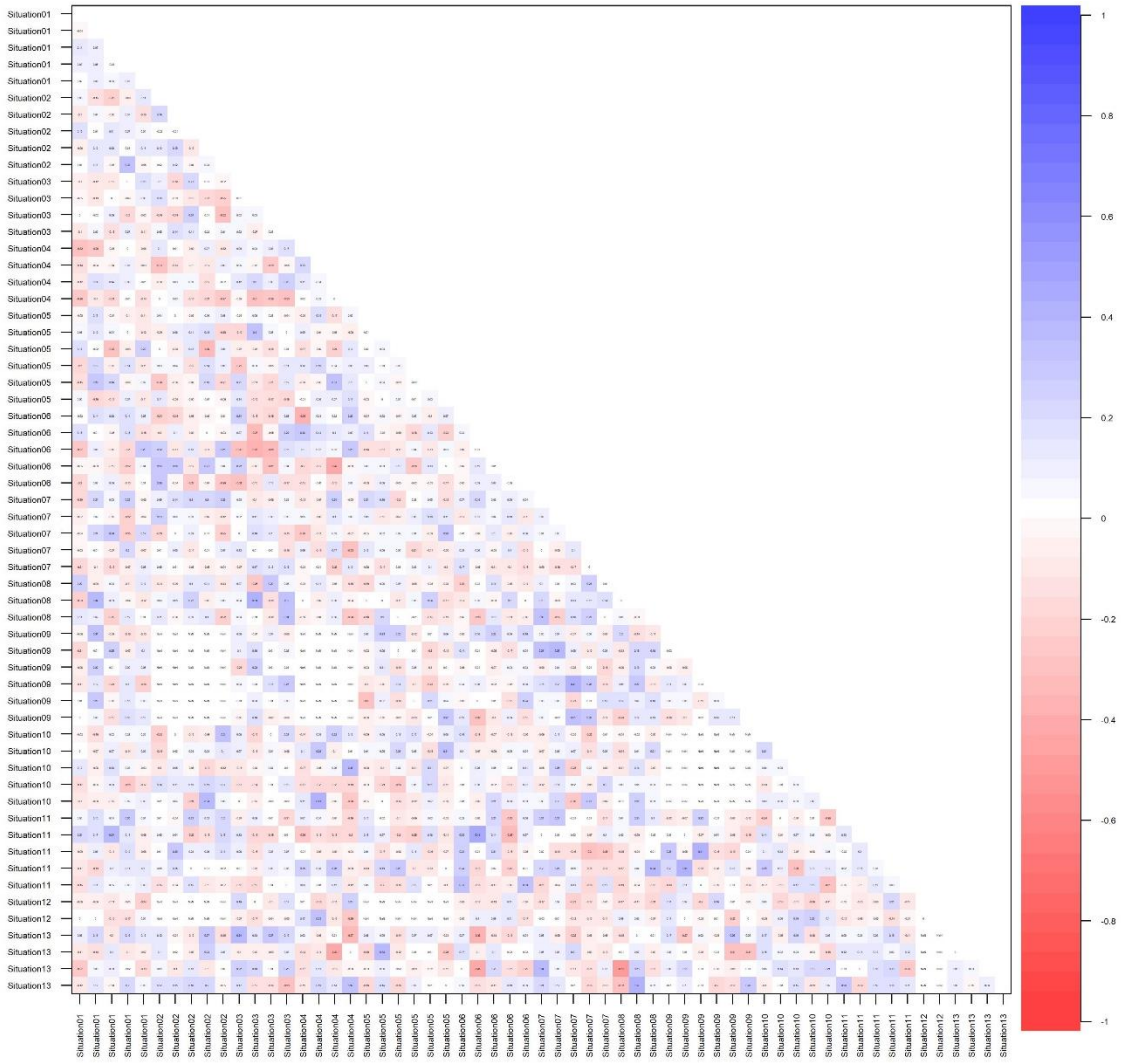
Die Gewichtung habe ich so vorgenommen, weil...
[Multiple Choice mit folgenden Antwortmöglichkeiten:]
a) wir als Unternehmen die Verantwortung für die durch uns verursachten Externen Effekte übernehmen sollten.
b) weil davon auszugehen ist, dass auch unsere Kunden zunehmend mehr Wert auf einen möglichst umweltfreundlichen Transport legen.
c) Frachtkosten und Kosten für die Beeinträchtigung der Umwelt in etwa gleich berücksichtigt werden sollten.
d) der marktseitige Preisdruck uns dazu zwingt, die Transportkosten zugunsten unserer Wettbewerbsfähigkeit gering zu halten.
e) die Frachtkosten ohnehin vom Kunden getragen werden und deswegen im Sinne des Umweltschutzes nachrangig sind.
f) für den Kunden der Preis das wichtigste Entscheidungskriterium ist.
g) aus betriebswirtschaftlicher Sicht keine Veranlassung dazu besteht, die externen Effekte bei der Transportmittelwahl übermäßig zu berücksichtigen.

Viele Grüße,
Michaela

Michaela Krüger, Dipl.-Kffr.

Leiterin Transport und Logistik
E-Mail: michaela.krueger@pyramid.de
Pyramid - Innovation & Design GmbH
Geschäftsführer: Heiko Hartmann, Thomas Werner
Amtsgericht Göttingen, HRB 6166
Nikolausberg 20, 37073

G. aQ3-Statistiken im sNCM-Test zum ersten Messzeitpunkt



Anmerkungen. Die Items sind in Reihenfolge der Situationen dargestellt. Nahe der Diagonalen finden sich deshalb die aQ3-Statistiken für Items innerhalb einer Situation.

H. Korrelation der Skalenmittelwerte der selbstberichteten hochschulischen und außerhochschulischen Lerngelegenheiten zu Nachhaltigkeitsthemen.

	NoNa-MW1	GeNa-MW1	BetNa-MW1	NoNa-MW2	GeNa-MW2	BetNa-MW2	NaIn-MW1	NaUm-MW1	InGe-MW1	NaIn-MW2	NaUm-MW2	InGe-MW2
NoNa-MW1	1.84 (1.38)											
GeNa-MW1	.75	1.72 (1.37)										
BetNa-MW1	.67	.61	1.97 (1.69)									
NoNa-MW2	.45			2.23 (1.16)								
GeNa-MW2		.40		.76	2.16 (1.16)							
BetNa-MW2			.39	.69	.64	2.36 (1.26)						
NaIn-MW1	.20	.25	.17				2.47 (1.47)					
NaUm-MW1	.30	.38	.26				.58	2.48 (1.69)				
InGe-MW1	.31	.36	.23				.52	.62	2.36 (1.64)			
NaIn-MW2				.23	.28	.18	.63			2.52 (1.05)		
NaUm-MW2				.27	.32	.27		.64		.53	2.52 (1.24)	
InGe-MW2				.19	.30	.22			.52	.55	.65	2.31 (1.16)

Anmerkungen. MW1 = Skalenmittelwert des ersten Messzeitpunkts. MW2 = Skalenmittelwert des zweiten Messzeitpunkts. Die Items wurden auf einer vier-stufigen Likert-Skala von 1 (*Trifft gar nicht zu*) bis 4 (*Trifft voll zu*) bearbeitet. In der Diagonalen sind die Skalenmittelwerte und Skalenstandardabweichungen (in Klammern) angegeben. Werte abseits der Diagonalen sind Pearsons Korrelationskoeffizienten. Im linken oberen Quadranten sind die Korrelationen der hochschulischen Lerngelegenheiten abgebildet im Quer- und Längsschnitt. Im rechten unteren Quadranten sind die Korrelationen der außer-hochschulischen Lerngelegenheiten im Quer- und Längsschnitt abgebildet. Im linken unteren Quadranten sind die Korrelationen zwischen hochschulischen und außer-hochschulischen Lerngelegenheiten zum jeweiligen Messzeitpunkt abgebildet.

I. Mittelwerte und beobachtete Kovarianzmatrix der in Kapitel 8.2.3 in den Analysenmodellen verwendeten Variablen

	M	na1_1	na2_1	na3_1	na1_2	na2_2	na3_2	bwl1_1	bwl2_1	bwl3_1	bwl1_2	bwl2_2	bwl3_2	ncm1_1	ncm2_1	ncm3_1
na1_1	0.53	.059														
na2_1	0.552	.013	.048													
na3_1	0.53	.008	.014	.04												
na1_2	0.604	.015	.021	.014	.05											
na2_2	0.56	.015	.011	.011	.017	.061										
na3_2	0.522	.015	.015	.019	.023	.023	.056									
bwl1_1	0.454	.021	.004	.004	.001	.013	.012	.065								
bwl2_1	0.46	-.001	-.003	.006	-.002	-.001	.001	.018	.077							
bwl3_1	0.492	.016	.006	-.007	-.001	.003	-.001	.023	.008	.068						
bwl1_2	0.476	.01	.01	.006	.021	.003	.01	.014	.008	.02	.047					
bwl2_2	0.427	-.003	.007	.005	.004	.004	.004	.004	.009	-.002	.006	.045				
bwl3_2	0.458	.01	.012	.002	.019	.014	.014	.012	.003	.026	.022	.009	.076			
ncm1_1	0.431	.007	.009	.002	.011	.012	.008	.01	.003	.003	.004	.008	.01	.05		
ncm2_1	0.405	.015	.009	.007	.008	.01	.003	.005	-.006	.005	.005	.001	.012	.015	.054	
ncm3_1	0.438	.006	.006	.006	.009	.014	.004	.008	.005	.004	.008	.001	.012	.013	.003	.05
ncm1_2	0.419	.011	.01	.009	.015	.016	.013	.005	.001	.006	.008	.003	.01	.01	.016	.005
ncm2_2	0.45	.007	.005	.003	.018	.008	.015	.005	.005	.000	.012	.005	.012	.009	.01	.006
ncm3_2	0.482	.009	.008	.006	.017	.018	.013	.009	.012	.013	.013	-.002	.011	.008	.005	.007
sit1_1	0.82	.004	.006	.001	.002	.018	.009	.011	-.021	.009	.005	.017	-.004	.01	.002	.004
sit2_1	0.854	.004	.02	.001	.008	.03	.008	.007	-.014	.009	-.004	.002	.008	.01	.017	.016
sit3_1	0.859	.004	.001	.001	.003	.016	.009	.014	-.02	.007	.000	.006	.013	.003	.035	-.001
sit1_2	0.846	.01	.003	.011	.005	.024	.026	.013	.007	.022	.018	.023	.031	.005	.016	.013
sit2_2	0.873	.009	.016	.006	.006	.018	.013	.005	-.003	-.004	.006	.007	.007	.014	.012	.007
sit3_2	1.062	.003	.012	.019	.001	.013	.016	.019	-.015	.005	.004	-.002	-.008	.013	-.002	.021
GotD	0.231	.000	-.015	-.008	-.016	-.016	-.02	.006	.019	.025	.001	.008	.01	.01	.017	-.011
GotE	0.106	.003	-.004	.007	.005	.001	.008	.013	-.008	-.002	.009	.000	.011	.003	.004	.009
Lei	0.058	-.003	-.008	-.008	-.007	-.003	.000	.000	.002	.007	.001	.000	.001	-.003	-.006	.002
Mar	0.192	-.005	-.003	-.001	-.018	-.002	-.007	.003	.008	-.001	-.002	.004	.001	-.017	-.007	.016
GotB	0.048	-.01	.015	.001	.006	.000	.013	-.005	.001	-.004	.005	.000	.014	.000	-.001	-.003
Ulm	0.135	.009	.008	.01	.02	.024	.022	-.002	-.003	-.001	-.005	-.005	-.014	.023	.009	.002
Hoh	0.135	.008	.004	.007	.006	.008	-.005	-.009	-.016	-.017	-.011	-.002	-.018	-.009	-.012	-.007
aOTL_1	1.819	-.015	.006	.001	-.004	-.013	-.009	.000	.006	.002	.013	-.016	.008	-.002	.025	-.002
naOTL_1	2.385	-.001	-.006	-.011	.013	.013	.019	-.003	-.025	-.004	-.007	-.006	.014	.014	-.005	.000
aOTL_2	2.201	.012	.028	.026	.05	.024	.042	-.017	-.027	-.027	.008	-.01	-.028	.014	.019	-.01
naOTL_2	2.415	-.003	.008	-.009	.026	-.008	.009	.006	.007	.000	-.001	.002	-.004	.018	-.007	.004

	ncm1_2	ncm2_2	ncm3_2	sit1_1	sit2_1	sit3_1	sit1_2	sit2_2	sit3_2	GotD	GotE	Lei	Mar	GotB	Ulm	Hoh	aOTL_1
na1_1																	
na2_1																	
na3_1																	
na1_2																	
na2_2																	
na3_2																	
bwl1_1																	
bwl2_1																	
bwl3_1																	
bwl1_2																	
bwl2_2																	
bwl3_2																	
ncm1_1																	
ncm2_1																	
ncm3_1																	
ncm1_2	.037																
ncm2_2	.016	.046															
ncm3_2	.008	.015	.045														
sit1_1	.011	.016	.002	.119													
sit2_1	.009	-.001	.017	.017	.136												
sit3_1	.013	.013	.01	-.006	.046	.138											
sit1_2	.011	.004	-.001	.026	.014	.07	.175										
sit2_2	.001	-.003	.005	-.004	.047	.000	-.01	.11									
sit3_2	.011	.004	.004	-.006	.008	-.011	.041	.029	.172								
GotD	.000	-.006	.003	-.009	.000	-.004	.027	-.002	-.03	.178							
GotE	.006	.007	-.008	.013	-.001	.016	-.025	.005	.013	-.024	.095						
Lei	-.004	-.005	-.002	-.002	-.003	-.006	.006	.003	.003	-.013	-.006	.054					
Mar	-.012	-.009	-.012	.01	-.003	.008	.002	.024	.002	-.044	-.02	-.011	.155				
GotB	.005	.003	.006	.005	.004	.003	.009	.004	-.007	-.011	-.005	-.003	-.009	.046			
Ulm	.012	.014	.017	-.007	.017	.012	-.006	.003	.025	-.031	-.014	-.008	-.026	-.006	.116		
Hoh	-.006	-.008	.003	-.01	-.01	-.021	.001	-.021	.012	-.031	-.014	-.008	-.026	-.006	-.018	.116	
aOTL_1	.001	.02	.000	-.039	-.005	-.003	-.006	-.001	-.025	.009	.027	-.01	-.015	.021	-.01	-.008	.343
naOTL_1	.002	.001	.001	-.016	.015	.037	.036	-.018	-.014	-.01	-.015	.000	-.009	.02	.036	.007	.063
aOTL_2	.03	.039	.036	.001	.006	.005	-.002	-.004	.001	-.07	-.012	-.043	-.073	.034	.106	.099	.185
naOTL_2	.001	.007	.015	-.001	.018	-.006	.011	-.024	-.036	.000	-.034	-.004	-.027	.018	.043	.017	.024

	naOTL_1	aOTL_2	naOTL_2
na1_1			
na2_1			
na3_1			
na1_2			
na2_2			
na3_2			
bwl1_1			
bwl2_1			
bwl3_1			
bwl1_2			
bwl2_2			
bwl3_2			
ncm1_1			
ncm2_1			
ncm3_1			
ncm1_2			
ncm2_2			
ncm3_2			
sit1_1			
sit2_1			
sit3_1			
sit1_2			
sit2_2			
sit3_2			
GotD			
GotE			
Lei			
Mar			
GotB			
Ulm			
Hoh			
aOTL_1			
naOTL_1	.261		
aOTL_2	.114	.571	
naOTL_2	.179	.127	.289

Anmerkungen. *M* = Mittelwert.

.

J. Bootstrap-Konfidenzintervalle für negativ geschätzte Varianzen in den Analysen zu Testwertinterpretation 3

Test	Modell	95% Konfidenzintervall
NagP	1	[-.013; .005]
NagP	2	[-.012; .005]
NagP	3	[-.015; .004]
sNCM	2	[-.019; .012]

Anmerkungen. Konfidenzintervall wurde mit jeweils 1000 Replikationen berechnet. Modell 1 = Feste Effekte von Standorten auf beide Messzeitpunkte und lineare Kontraste zwischen Schwerpunkt- und Kontrollgruppe für beide Messzeitpunkte. Modell 2 = Feste Effekte von Standorten auf beide Messzeitpunkte und selbstberichtete akademische Lerngelegenheiten als Prädiktoren. Modell 3 = Feste Effekte von Standorten auf beide Messzeitpunkte und selbstberichtete nicht-akademische Lerngelegenheiten als Prädiktoren.

a. Bootstrap-Konfidenzintervalle für negativ geschätzte Varianzen in den Alternativmodellen zur Analyse von Testwertinterpretation 3

Test	Modell	95% Konfidenzintervall
NagP	3	[<-.001.; < .001]
sNCM	2	[-.23 ; .07]

Anmerkungen. Konfidenzintervall wurde mit jeweils 1000 Replikationen berechnet. Modell 2 = Post-hoc Korrektur der Standardfehler für Daten geclustert in Lehrveranstaltungen. Selbstberichtete akademische Lerngelegenheiten als Prädiktoren auf beide Messzeitpunkte und Vorhersage der Selbstberichte zum zweiten Messzeitpunkt durch Standorte. Modell 3 = Post-hoc Korrektur der Standardfehler für Daten geclustert in Lehrveranstaltungen. Selbstberichtete nicht-akademische Lerngelegenheiten als Prädiktoren auf beide Messzeitpunkte.

K. Vollständige Tabelle für standardisierte Regressionsgewichte der Standorte und lineare Kontraste der Schwerpunkt- und Kontrollgruppen mit den BWL-Testwerten als abhängigen Variablen (Modell 1)

	Modell 1		
	β	*SE*	*p*
Feste Effekte der Standorte			
BWL.T1 ~			
Göttingen D	.195	.068	.004
Göttingen E	.325	.073	<.001
Leipzig	.164	.068	.016
Marburg	.104	.065	.109
Göttingen B	-.047	.073	.515
Ulm	.179	.068	.008
Hohenheim	.013	.074	.860
BWL.T2 ~			
Göttingen D	-.051	.150	.735
Göttingen E	.014	.141	.921
Leipzig	-.225	.154	.143
Marburg	-.058	.101	.567
Göttingen B	.450	.154	.004
Ulm	-.178	.140	.204
Hohenheim	-.011	.156	.946
Latente Interzepte			
BWL.T1	.00		
BWL.T2	.326	.388	.401
Prädiktion der Testwerte durch Lerngelegenheiten			
BWL.T1 ~			
Ø Kontrollgruppen	.098	.021	<.001
Ø Schwerpunktgruppen	.026	.053	.629
Kontrast.T1	-.072	.046	.119
BWL.T2 ~			
Ø Kontrollgruppen	-.037	.056	.510
Ø Schwerpunktgruppen	.119	.132	.368
Kontrast.T2	.156	.094	.097
BWL.T1	.760	.184	<.001
R^2 (erklärte Varianz)		.72	

Anmerkungen. Regressionsgewichte der festen Effekte sind in Relation zur Referenzgruppe „Göttingen C“ (Kontrollgruppe) angegeben. „Ø Schwerpunktgruppen“: die mittleren Regressionsgewichte der Standorte der Schwerpunktgruppe, gewichtet nach der jeweiligen Gruppengröße. „Ø Kontrollgruppen“: die mittleren Regressionsgewichte der Standorte der Kontrollgruppe, gewichtet nach der jeweiligen Gruppengröße. „Kontrast“ = Ø Schwerpunktgruppen *minus* Ø Kontrollgruppen.

Modellfit: $\chi^2(38) = 32.40$, $p = .73$, CFI = 1.00, TLI = 1.06, RMSEA = .00, 90% KI [.00 - .02], SRMR = .03.

L. Vollständige Tabelle für die standardisierten Regressionsgewichte der Standorte und lineare Kontraste der Schwerpunkt- und Kontrollgruppen mit den NagP-, dNCM- und sNCM-Testwerten als abhängigen Variablen (Modell 1)

	NagP [a]			dNCM			sNCM		
	β	*SE*	*p*	β	*SE*	*p*	β	*SE*	*p*
Feste Effekte der Standorte									
Testwert.T1 ~									
Göttingen D	.046	.058	.436	.147	.059	.012	.201	.071	.004
Göttingen E	.110	.064	.088	.191	.065	.003	.200	.079	.011
Leipzig	-.091	.060	.128	-.058	.060	.332	-.157	.076	.038
Marburg	.046	.055	.405	.016	.055	.771	.151	.066	.021
Göttingen B	-.475	.088	<.001	-.527	.109	<.001	.181	.078	.020
Ulm	.233	.058	<.001	.206	.058	<.001	.265	.067	<.001
Hohenheim	.283	.063	<.001	.163	.063	.010	.054	.075	.467
Testwert.T2 ~									
Göttingen D	-.077	.115	.503	-.182	.133	.173	.115	.143	.419
Göttingen E	.019	.106	.858	-.038	.121	.752	.337	.121	.005
Leipzig	.037	.122	.760	-.119	.133	.373	.407	.146	.005
Marburg	-.057	.081	.485	-.140	.088	.112	.264	.094	.005
Göttingen B	.804	.168	<.001	.623	.215	.004	.204	.165	.217
Ulm	.140	.108	.195	.079	.130	.543	.203	.132	.126
Hohenheim	-.068	.130	.604	-.123	.143	.387	.219	.144	.130
Latente Interzepte									
Testwert.T1	.000			.000			.000		
Testwert.T2	.039	.039	.323	.079	.040	.052	-.020	.054	.706
Autoregressiver Effekt									
Testwert T2 ~									
Testwert.T1	1.161	.103	<.001	.791	.235	.001	.781	.186	<.001
Lineare Kontraste									
Testwert.T1:									
Ø Kontrollgruppen	.014	.018	.462	.043	.019	.021	.052	.023	.026
Ø Schwerpunktgruppen	.002	.056	.969	-.074	.062	.234	.143	.058	.013
Kontrast.T1	-.011	.051	.822	-.117	.055	.033	.091	.048	.055
Testwert.T2 ~									
Ø Kontrollgruppen	-.012	.042	.781	-.061	.047	.202	.123	.047	.009
Ø Schwerpunktgruppen	.286	.106	.007	.185	.125	.140	.210	.126	.095
Kontrast.T2	.297	.080	<.001	.245	.095	.010	.087	.094	.357
R^2 (erklärte Varianz)		1.00			.50			.91	

Anmerkungen. Regressionsgewichte der festen Effekte sind in Relation zur Referenzgruppe „Göttingen C" (Kontrollgruppe) angegeben. „Ø Schwerpunktgruppen": die mittleren Regressionsgewichte der Standorte der Schwerpunktgruppe, gewichtet nach der jeweiligen Gruppengröße. „Ø Kontrollgruppen": die mittleren Regressionsgewichte der Standorte der Kontrollgruppe, gewichtet nach der jeweiligen Gruppengröße. „Kontrast" = Ø Schwerpunktgruppen *minus* Ø Kontrollgruppen. Die Tabelle gibt zweiseitige *p*-Werte an.

Modellfit NagP: $\chi^2(38) = 35.22$, $p = .60$, CFI = 1.00, TLI = 1.02, RMSEA = .00, 90% KI [.00 - .03], SRMR = .04. dNCM: $\chi^2(39) = 36.26$, $p = .60$, CFI = 1.00, TLI = 1.03, RMSEA = .00, 90% KI [.00 - .03], SRMR = .04. sNCM: $\chi^2(37) = 5.14$, $p = .07$, CFI = .89, TLI = .83, RMSEA = .03, 90% KI [.00 - .04], SRMR = .05.

[a] Für das Modell wurde eine Varianz auf einen Wertebereich größer gleich Null fixiert.

M. Vollständige Tabelle für die standardisierten Regressionsgewichte der Standorte und der selbstberichteten hochschulischen Lerngelegenheiten mit den NagP-, dNCM- und sNCM-Testwerten als abhängigen Variablen (Modell 2)

	NagP [a]			dNCM			sNCM [a]		
	β	*SE*	*p*	β	*SE*	*p*	β	*SE*	*p*
Feste Effekte der Standorte									
Testwert.T1 ~									
Göttingen D	.047	.059	.419	.148	.059	.012	.208	.071	.003
Göttingen E	.112	.064	.084	.192	.066	.003	.208	.079	.009
Leipzig	-.092	.060	.123	-.058	.060	.332	-.163	.077	.033
Marburg	.045	.055	.410	.017	.055	.761	.153	.065	.020
Göttingen B	-.474	.088	<.001	-.526	.110	<.001	.190	.078	.016
Ulm	.233	.058	<.001	.206	.059	<.001	.269	.066	<.001
Hohenheim	.285	.063	<.001	.165	.064	.010	.062	.076	.415
Testwert.T2 ~									
Göttingen D	-.078	.115	.495	-.196	.133	.140	.157	.137	.253
Göttingen E	.017	.106	.872	-.056	.121	.640	.370	.118	.002
Leipzig	.040	.122	.746	-.113	.132	.391	.393	.129	.002
Marburg	-.056	.081	.487	-.142	.087	.103	.270	.089	.003
Göttingen B	.797	.174	<.001	.533	.223	.017	.371	.160	.020
Ulm	.129	.121	.286	-.013	.141	.925	.375	.141	.008
Hohenheim	-.082	.146	.577	-.238	.158	.132	.435	.146	.003
Latente Interzepte									
Testwert.T1	.000			.000			.000		
Testwert.T2	.032	.054	.554	.015	.057	.790	.104	.069	.131
Prädiktion der Testwerte durch Lerngelegenheiten									
Testwert.T1 ~									
Hochschulische LG.T1	-.021	.066	.755	-.009	.072	.904	-.045	.075	.550
Testwert.T2 ~									
Testwert.T1	1.162	.104	<.001	.777	.237	.001	.765	.111	<.001
Hochschulische LG.T2	.026	.121	.831	.218	.132	.100	-.402	.154	.009
Prädiktion der Lerngelegenheiten									
Hochschulische LG.T2 ~									
Hochschulische LG.T1	.452	.053	<.001	.449	.053	<.001	.453	.052	<.001
Göttingen D	.024	.067	.717	.025	.067	.709	.030	.066	.647
Göttingen E	.032	.057	.579	.034	.057	.551	.034	.057	.552
Leipzig	-.042	.070	.547	-.042	.070	.546	-.040	.069	.561
Marburg	.005	.044	.918	.005	.044	.917	.006	.043	.895
Göttingen B	.302	.074	<.001	.299	.074	<.001	.309	.073	<.001
Ulm	.426	.052	<.001	.428	.052	<.001	.425	.052	<.001
Hohenheim	.468	.067	<.001	.467	.067	<.001	.468	.066	<.001
R^2 (erklärte Varianz)	1.00			.52			1.00		

Anmerkungen. Regressionsgewichte der festen Effekte sind in Relation zur Referenzgruppe „Göttingen C" (Kontrollgruppe) angegeben. Die Tabelle gibt zweiseitige *p*-Werte an.

Modellfit NagP: $\chi^2 = 41.28(48)$ $p = .743$, CFI = 1.00, TLI = 1.03, RMSEA = .00, 90% KI [.00 - .02], SRMR = .04. dNCM: $\chi^2(49) = 49.30$ $p = .46$, CFI = 1.00, TLI = 1.00, RMSEA = .00, 90 % KI = [.00 - .03], SRMR = .04. sNCM: $\chi^2(47) = 57.89$, $p = .13$, CFI = .95, TLI = .92, RMSEA = .02, 90% KI [.00 - .04], SRMR = .04.

[a] Für das Modell wurde eine Varianz auf einen Wertebereich größer gleich Null fixiert.

N. Vollständige Tabelle für die standardisierten Regressionsgewichte der Standorte und der selbstberichteten außerhochschulischen Lerngelegenheiten mit den NagP-, dNCM- und sNCM-Testwerten als abhängigen Variablen (Modell 3)

	NagP [a]			**dNCM**			**sNCM**		
	β	*SE*	*p*	β	*SE*	*p*	β	*SE*	*p*
Feste Effekte der Standorte									
Testwert.T1 ~									
Göttingen D	.045	.058	.443	.148	.059	.012	.200	.070	.004
Göttingen E	.105	.064	.102	.190	.065	.004	.204	.079	.009
Leipzig	-.099	.060	.099	-.064	.060	.290	-.157	.076	.038
Marburg	.045	.055	.410	.015	.055	.790	.151	.065	.021
Göttingen B	-.500	.088	<.001	-.537	.110	<.001	.182	.080	.023
Ulm	.210	.059	<.001	.190	.060	.002	.264	.068	<.001
Hohenheim	.262	.064	<.001	.148	.065	.023	.054	.076	.482
Testwert.T2 ~									
Göttingen D	-.091	.115	.431	-.184	.134	.170	.130	.138	.346
Göttingen E	.011	.106	.915	-.038	.121	.754	.330	.118	.005
Leipzig	.024	.123	.845	-.125	.133	.350	.428	.141	.002
Marburg	-.066	.081	.419	-.143	.088	.105	.261	.091	.004
Göttingen B	.782	.172	<.001	.600	.215	.005	.244	.160	.128
Ulm	.121	.112	.281	.080	.133	.550	.249	.130	.055
Hohenheim	-.094	.134	.483	-.126	.145	.387	.267	.140	.056
Latente Interzepte									
Testwert.T1	.000			.000			.000		
Testwert.T2	.015	.064	.816	.100	.072	.167	.072	.081	.376
Prädiktion der Testwerte durch Lerngelegenheiten									
Testwert.T1 ~									
Außerhochschulische LG.T1	.112	.067	.096	.084	.073	.248	.002	.076	.976
Testwert.T2 ~									
Testwert.T1	1.165	.103	<.001	.770	.232	.001	.786	.182	<.001
Außerhochschulische LG.T2	.026	.096	.784	-.005	.109	.960	-.177	.111	.129
Prädiktion der Lerngelegenheiten									
Außerhochschulische LG.T2 ~									
Außerhochschulische LG.T1	.658	.048	<.001	.655	.048	<.001	.656	.048	< .001
R^2 (erklärte Varianz)		1.00			.49			.95	

Anmerkungen. Regressionsgewichte der festen Effekte sind in Relation zur Referenzgruppe „Göttingen C“ (Kontrollgruppe) angegeben. Die Tabelle gibt zweiseitige *p*-Werte an.

Modellfit NagP: $\chi^2(55) = 6.10$, $p = .30$, CFI =.98, TLI = .98, RMSEA = .01, 90% KI [.00 - .03], SRMR = .04. dNCM: $\chi^2(56) = 51.98$, $p = .63$, CFI = 1.00, TLI = 1.03, RMSEA = .00, 90% KI [.00 - .02], SRMR = .04. sNCM: $\chi^2(54) = 66.27$, $p = .12$, CFI = .93, TLI = .90, RMSEA = .02, 90% KI [.00 - .04], SRMR = .05.

[a] Für das Modell wurde eine Varianz auf einen Wertebereich größer gleich Null fixiert.

O. Variance Inflation Factor (VIF) für Prädiktoren in Modellen mit negativ geschätzter Varianz für Testwerte zum zweiten Messzeitpunkt.

Test	Model	Prädiktor	Toleranz	VIF
NagP	Model 1	NaGP.T1	.56	1.79
	Model 2	NaGP.T1	.56	1.79
		Hochschulische LG.T2	.37	2.69
	Model 3	NaGP.T1	.55	1.82
		Außerhochschulische LG.T2	.56	1.79
sNCM	Model 2	sNCM.T1	.82	1.22
		Hochschulische LG.T2	.37	2.67

Anmerkungen. Toleranz = 1 – R^2_k, wobei R^2_k die erklärte Varianz in Prädiktor *k* durch alle anderen Prädiktoren im Modell. VIF ist der Kehrwert der Toleranz.