TUM School of Education

Susanne Klatten-Stiftungslehrstuhl für Empirische Bildungsforschung

Studien zur Überprüfung der Validität eines Instruments zur Erfassung professioneller Unterrichtswahrnehmung von Lehramtsstudierenden

Gloria Katharina Jahn

Vollständiger Abdruck der von der Fakultät *TUM School of Education* der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Philosophie (Dr. phil.)

genehmigten Dissertation.

Vorsitzende:	Univ.-Prof. Dr. Doris Lewalter
Prüfer/innen der Dissertation:	1. Univ.-Prof. Dr. Manfred Prenzel
	2. Univ.-Prof. Dr. Christina Seidel

Die Dissertation wurde am 16.07.2014 bei der Technischen Universität eingereicht und durch die TUM School of Education am 12.09.2014angenommen.

GLIEDERUNG

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

DANKSAGUNG

Sehr viele Menschen haben dazu beigetragen, dass ich während meiner Promotion viel gelernt habe – sowohl fachlich als auch über mich selbst. An dieser Stelle ist es mir nicht möglich, meinen Dank ausreichend auszudrücken. Deshalb werde ich mich bei diesen Menschen auch noch persönlich bedanken. Doch so viel soll gesagt sein:

Allen voran möchte ich meinem Erstbetreuer und meiner Zweitbetreuerin für ihre engagierte und kompetente Unterstützung bei der Anfertigung meiner Dissertation danken. Prof. Dr. Manfred Prenzel gilt mein besonderer Dank für sein Vertrauen in meine Fähigkeiten und seine Offenheit für meine Forschungsideen. Prof. Dr. Tina Seidel danke ich speziell dafür, dass sie mir stets geholfen hat, meine Ideen zu strukturieren. Darüber hinaus gilt mein herzlichster Dank meiner Mentorin Dr. Katharina Müller, die mich vom ersten Arbeitstag an auf meinem wissenschaftlichen Weg mit vielen wertvollen Gesprächen begleitet hat.

Ebenso bedanke ich mich bei meinen Kolleginnen im Observe-Projekt für die bereichernden Diskussionen und die gemeinsame Projektarbeit. In diesem Zusammenhang möchte ich ausdrücklich auch meinen Hilfskräften für ihre tatkräftige Unterstützung während der Datenerhebung und Datenaufbereitung danken. Während meiner Promotion hatte ich das große Glück, an einem außergewöhnlichen Lehrstuhl in einer unheimlich anregenden Atmosphäre zu arbeiten. Vielen Dank euch allen! Ein besonderer Dank geht an die vielen Korrektur-Leser am Lehrstuhl und weit darüber hinaus. Nicht genug kann ich meiner Kollegin und langjährigen Freundin Dr. Jessica Mattern für ihre unentwegte Unterstützung danken – fachlich wie emotional. Danke, dass du immer die richtigen Worte gefunden hast, mich zu motivieren!

Ganz besonders möchte ich meiner Familie, allen voran meinen Eltern, dafür danken, dass sie mich stets darin unterstützt, das zu tun, was ich liebe. Abschließend möchte ich von ganzem Herzen meinem Freund dafür danken, dass er immer an mich glaubt und mich bestärkt.

ZUSAMMENFASSUNG

Trotz der zunehmenden Kompetenzorientierung in der universitären Lehrerbildung mangelt es noch an Instrumenten, die Kompetenzen kontextualisiert erfassen. Ein positives Beispiel stellt das Tool Observer zur Erfassung professioneller Unterrichtswahrnehmung dar. Um nicht nur Rückmeldung auf individueller Ebene, sondern auch auf systemischer Ebene geben zu können, ist ein großflächiger Einsatz des Instruments notwendig. Deshalb zielt die vorliegende Arbeit darauf ab, das Instrument für einen Einsatz im Large-Scale-Kontext zu validieren. Aus diesem Einsatzkontext ergeben sich spezifische Anforderungen an das Instrument, die die drei Validitätsaspekte *Struktur*, *Generalisierbarkeit* und *Externalität* betreffen. Deren Prüfung und Synthese in eine übergreifende Validitätsaussage ist Gegenstand der vorliegenden Arbeit. In der ersten Forschungsfrage wird überprüft, inwieweit sich die Struktur professioneller Unterrichtswahrnehmung mit dem Tool Observer an einer großen und heterogenen Stichprobe empirisch abbilden lässt. Die zweite Forschungsfrage untersucht, ob ein Zusammenhang zwischen der Kompetenzerfassung und unterschiedlichen Erhebungsbedingungen existiert. In der dritten Forschungsfrage werden die Kompetenzstruktur und die Itemschwierigkeiten Studierender verschiedener Lehramtsstudiengänge verglichen. Für die vierte Forschungsfrage werden Prädiktoren für den Studien- und Berufserfolg als externale Kriterien herangezogen. Es wird geprüft, inwieweit ein Zusammenhang mit professioneller Unterrichtswahrnehmung besteht. Die Befunde zeigen, dass das Instrument in einer großen und heterogenen Stichprobe die Struktur professioneller Unterrichtswahrnehmung empirisch abbildet, und dass es über verschiedene Erhebungsbedingungen hinweg stabil eingesetzt werden kann. Zudem wird die Kompetenzstruktur für Studierende unterschiedlicher Lehramtsstudiengänge repliziert, wobei die Itemschwierigkeiten mit Blick auf die verschiedenen Lehramtsstudiengänge interpretiert werden müssen. Weiter ergibt sich ein heterogenes Bild hinsichtlich eines Zusammenhangs mit Prädiktoren für den Studien- und Berufserfolg: Ein lernbegleitungsorientierter Lehrbegriff und das Persönlichkeitsmerkmal Verträglichkeit erklären zusammen 15 % Varianz professioneller Unterrichtswahrnehmung, während pädagogisches Interesse und Interesse an Bildungswissenschaften keine zusätzliche Varianz aufklären. Die vorliegende Arbeit zeigt Möglichkeiten und Grenzen eines Einsatzes des Tools Observer im Large-Scale-Kontext auf und leistet damit einen substantiellen Beitrag zu dessen Validierung.

1 EINLEITUNG UND ZIELBESTIMMUNG

Ein Universitätsstudium soll gemäß der Empfehlung des Wissenschaftsrates (2006) „angemessen auf den Arbeitsmarkt vorbereiten und den Absolventen dabei die Möglichkeit bieten, durch lebenslanges Lernen den Anforderungen des Beschäftigungssystems auch längerfristig gewachsen zu bleiben“ (S. 9). Eine Aufgabe der empirischen Bildungsforschung ist es, zu überprüfen, inwieweit Bildungsinstitutionen die von ihnen gesetzten Ziele erreichen (Hartig, 2008) und somit effektiv sind. Im Hinblick auf die universitäre Lehrerbildung wurde im Zuge des Bologna-Prozesses und bedingt durch schwache bis mittelmäßige Leistungen deutscher Schülerinnen und Schüler in internationalen Vergleichsstudien verstärkt eine empirische Überprüfung der Effektivität der universitären Lehrerbildung gefordert (vgl. Blömeke, 2004). Eine grundlegende Voraussetzung für eine derartige Überprüfung ist die Definition von messbaren Kriterien dafür, was Studierenden können sollen (Hartig, 2008). Für den Bereich der universitären Lehrerbildung wurden derart konkrete Kriterien mit der Verabschiedung der Standards der Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK, 2004) etabliert. In diesen werden keine curricularen Inhalte, sondern handlungsrelevante Kompetenzen für die Schul- und Unterrichtspraxis festgelegt. Es werden vier Kompetenzbereiche (Unterrichten, Erziehen, Beurteilen und Innovieren) beschrieben, die sich in insgesamt elf Kompetenzen differenzieren lassen (KMK, 2004). Damit wurde auch auf die verstärkt auftretende Kritik reagiert, in der eine mangelnde Berufsfeldorientierung und die Beliebigkeit der Inhalte des Lehramtsstudiums beklagt wurden (Schaefers, 2002; Terhart, 2000). Im Hinblick auf die Forderung nach der empirischen Überprüfung der Lehrerbildung bedarf es neben der Festlegung von Kriterien auch der Entwicklung von Messinstrumenten, die es ermöglichen, Lernerfolge Studierender zu diagnostizieren und zurückzumelden. Voraussetzung für die Erfassung professioneller Kompetenzen ist zum einen eine klare theoretische Modellierung und zum anderen eine empirische Überprüfung des theoretisch postulierten Kompetenzmodells (Koeppen, Hartig, Klieme & Leutner, 2008). Darüber hinaus ist eine kontextualisierte und damit verhaltensnahe Erfassung zentral für die Validität von Kompetenzmessungen (Darling-Hammond, 2006). Denn nur so kann die Komplexität und Kontextgebundenheit unterrichtlichen Handelns berücksichtigt werden (Oser, Heinzer & Salzmann, 2010). Die Modellierung und Erfassung professioneller Kompetenzen von Lehrpersonen ist mittlerweile zentrales Thema empirischer Forschung. Projekte wie COACTIV (Kunter et al., 2011) haben dazu beigetragen, Kompetenzen von Lehrpersonen zu konzeptualisieren und stan-

dardisierte Testverfahren zu entwickeln. Auch im Bereich der standardisierten Erfassung pädagogisch-psychologischen Wissens wurden Fortschritte erzielt (z. B. Voss, Kunter & Baumert, 2011). Es werden vermehrt Versuche unternommen, bisher gängige Selbsteinschätzungen im Fragebogenformat (Frey, 2006) und distale Indikatoren wie Noten durch verhaltensnahe Kompetenzerfassungen zu ergänzen (siehe Abschnitt 2.1.2.2.2). Allerdings steht die Entwicklung von Instrumenten, die über Erhebungen im Fragebogenformat hinausgehen und Kompetenzen kontextualisiert erfassen, noch am Anfang (Kunter & Klusmann, 2010; Seidel, Blomberg & Stürmer, 2010a; Seidel & Stürmer, in Druck).

1.1 Anliegen der Arbeit

Wie erläutert stellen die KMK-Standards der Lehrerbildung (2004) mit den beschriebenen Kompetenzen zentrale Kriterien für die Beurteilung des Lernerfolgs Studierender im Rahmen ihrer universitären Lehrerbildung bereit. Die universitäre Lehrerbildung in Deutschland besteht grundsätzlich aus dem Studium der Fächer, Fachdidaktiken und Bildungswissenschaften sowie aus Praktikumsphasen (Terhart, 2009). Für den Aufbau der beschriebenen Kompetenzen werden wichtige Lerngelegenheiten vor allem in den bildungswissenschaftlichen Studienanteilen angeboten, die Themen aus unterschiedlichen Disziplinen, insbesondere aus Erziehungswissenschaft, Psychologie und Soziologie, beinhalten (Kunina-Habenicht et al., 2012). Allerdings sind die beschriebenen Kompetenzen recht weit gefasst, so dass es für die Entwicklung eines Messinstruments erforderlich ist, einen Fokus zu setzen. Dies wurde beispielsweise im DFG-Projekt Observe (Seidel, Stürmer, Prenzel, Jahn & Schäfer, eingereicht) umgesetzt, in dem die professionelle Unterrichtswahrnehmung Lehramtsstudierender fokussiert wird. Diese ist im Kompetenzbereich Unterrichten der KMK-Standards der Lehrerbildung (2004) zu verorten. Sie stellt einen Indikator dafür dar, inwieweit Lehramtsstudierende in der Lage sind, ihr Wissen über effektives Lehren und Lernen, das nach Shulman (1987) inhaltlich dem pädagogisch-psychologischen Wissen zuzuordnen ist, auf authentische Unterrichtssituationen anzuwenden (Seidel & Stürmer, in Druck). Somit entspricht professionelle Unterrichtswahrnehmung einer Vorstufe professioneller Handlungskompetenz (Seidel et al., 2010a).

Die vorliegende Arbeit ist in das DFG-Projekt Observe eingebettet. Im Projektverlauf wurde ein videobasiertes Online-Tool zur kontextualisierten und gleichzeitig standardisierten Erfassung professioneller Unterrichtswahrnehmung entwickelt (Seidel, Blomberg & Stür-

mer, 2010b) mit dem Ziel, interindividuelle Unterschiede und Entwicklungen Lehramtsstudierender im Rahmen der universitären Lehrerbildung abzubilden. Gerade in Anbetracht der großen Rolle, die Kompetenzmessungen für die Optimierung von Bildungsprozessen und Weiterentwicklung von Bildungssystemen spielen (Koeppen et al., 2008), gewinnt die Beurteilung der Güte der eingesetzten Messinstrumente an Bedeutung. Neben der Gewährleistung der Objektivität und Reliabilität eines Messinstruments ist insbesondere die Überprüfung der Validität entscheidend, da sich diese auf die Interpretation und Nutzung der Messergebnisse bezieht (American Educational Research Association [AERA], American Psychological Association [APA] & National Council on Measurement in Education [NCME], 1999; Messick, 1995). Bei mangelnder Validität sind alle Schlüsse, die aus entsprechenden Messungen gezogen werden und als Grundlage für Bildungsentscheidungen auf individueller oder politischer Ebene dienen, problematisch und nur begrenzt nützlich (Zumbo, 2007). An dieser Stelle setzt die vorliegende Arbeit an und zielt darauf ab, einen substantiellen Beitrag zur Validierung des videobasierten Online-Tools Observer zu leisten sowie Möglichkeiten und Grenzen eines großflächigen Einsatzes aufzuzeigen. Dazu werden zunächst Anforderungen an das Instrument hinsichtlich der Validität der Interpretation und Nutzung der Messergebnisse spezifiziert und bisherige Projektstudien im Hinblick auf die Überprüfung dieser Anforderungen eingeordnet. Anschließend steht die Überprüfung vier spezifischer Anforderungen im Mittelpunkt, die mit einem übergreifenden Einsatz im Large-Scale-Kontext einhergehen. Diese Überprüfungen zielen auf drei Validitätsaspekte ab, nämlich (1) *Struktur*, (2) *Generalisierbarkeit* über verschiedene Erhebungsbedingungen und unterschiedliche Lehramtsstudiengänge hinweg sowie (3) *Externalität*.

1.2 Vorgehen

Die vorliegende Arbeit gliedert sich in einen theoretischen und einen empirischen Teil. Im ersten Kapitel des Theorieteils wird eine Arbeitsdefinition von Kompetenz gegeben und hinsichtlich ihrer Nützlichkeit für die vorliegende Arbeit analysiert. Außerdem werden Herausforderungen der Kompetenzerfassung und verschiedene methodische Ansätze, auch speziell aus Large-Scale-Perspektive, diskutiert. Das zweite Kapitel fokussiert professionelle Unterrichtswahrnehmung als Beispiel für eine professionelle Kompetenz von Lehrpersonen. Professionelle Unterrichtswahrnehmung wird zunächst definiert und ihre Struktur theoretisch modelliert, bevor auf ihre Erfassung im Rahmen des DFG-Projekts Observe

eingegangen wird. Im dritten Kapitel werden gemäß dem argumentbasierten Ansatz von Validierung (Kane, 1992, 2001, 2013a) Anforderungen an ein Instrument zur Erfassung professioneller Unterrichtswahrnehmung beschrieben. Dabei liegt der Schwerpunkt auf Anforderungen hinsichtlich der Validität der Interpretation und Nutzung der Messergebnisse. Es wird ein integriertes Modell von Validität erarbeitet, vor dessen Hintergrund bisherige Projektstudien eingeordnet und noch ausstehende Überprüfungen mit Blick auf einen großflächigen Einsatz des Messinstruments identifiziert werden. Ausgehend vom dargelegten theoretischen Hintergrund werden vier zentrale Forschungsfragen und Hypothesen abgeleitet. Der empirische Teil ist entlang dieser vier Forschungsfragen aufgebaut. Es werden jeweils das methodische Vorgehen beschrieben, die Ergebnisse präsentiert sowie inhaltlich und methodisch diskutiert und weitere Implikationen dargestellt. Abschließend werden die verschiedenen Evidenzgrundlagen zu einer übergreifenden Validitätsaussage bezüglich des Einsatzes des Tools Observer im Large-Scale-Kontext synthetisiert. Im Anschluss daran wird der in der vorliegenden Arbeit gewählte Ansatz zur Validierung des Instruments beurteilt. Abschließend wird ein Ausblick auf anschließende Forschungsfragen skizziert.

2 THEORETISCHER HINTERGRUND

2.1 Erfassung professioneller Kompetenzen von Lehramtsstudierenden

Das erste Kapitel des theoretischen Hintergrunds fokussiert die Erfassung professioneller Kompetenzen von Lehramtsstudierenden. Dabei wird zunächst der Begriff Kompetenz diskutiert und für die vorliegende Arbeit definiert. Anschließend wird auf methodische Ansätze zur Kompetenzerfassung näher eingegangen.

2.1.1 Der Kompetenzbegriff

Der Begriff Kompetenz wird im alltäglichen Sprachgebrauch häufig verwendet und kann dabei unterschiedliche Bedeutungen haben. Auch im bildungswissenschaftlichen Kontext ist Kompetenz seit dem Programme for International Student Assessment (PISA) ein sehr prominenter Begriff (Prenzel, Gogoling & Krüger, 2007), der nicht einheitlich gebraucht wird (Hartig, 2008). Versteht man Kompetenzen als Ergebnisse von Bildungsprozessen (vgl. Hartig, 2008), ist im Hinblick auf eine empirische Erfassung eine einheitliche und präzise Definition von Kompetenz von großer Bedeutung. Denn nur dadurch kann im Rahmen von empirischen Studien sichergestellt werden, dass eine adäquate Operationalisierung der zu untersuchenden Kompetenz erfolgt, Befunde einzelner Studien angemessen interpretiert und bewertet werden sowie Befunde verschiedener Studien vergleichbar sind. Eine allgemeingültige, wissenschaftlichen Kriterien genügende Definition von Kompetenz ist jedoch aufgrund der alltagssprachlichen sowie wissenschaftlichen Bedeutungsvielfalt unrealistisch (Hartig, 2008). Deshalb gilt es, für spezifische Fragestellungen explizite Arbeitsdefinitionen zu formulieren (Hartig, 2008). Vor dem Hintergrund der Entwicklung von Messinstrumenten zur validen Erfassung von Ergebnissen von Bildungsprozessen wird im Folgenden eine Arbeitsdefinition von Kompetenz für die vorliegende Arbeit präzisiert, kurz in der Kompetenzdebatte verortet und hinsichtlich ihrer Nützlichkeit analysiert. Auf eine ausführliche Darstellung und Diskussion verschiedener Wurzeln und Definitionen des Kompetenzbegriffs wird an dieser Stelle verzichtet und auf Klieme und Hartig (2007) verwiesen.

2.1.1.1 Arbeitsdefinition von Kompetenz

Einen bis heute einflussreichen Systematisierungsversuch unterschiedlicher Kompetenzbegriffe legte Weinert (2001) vor, in dem die große Bandbreite an Bedeutungen wie auch die Widersprüchlichkeit einzelner Verwendungen des Begriffs Kompetenz aufgezeigt werden (Hartig, 2008). Darauf aufbauend definierte Weinert (2002) Kompetenzen als „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 Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können“ (S. 27f). Diese Definition findet bis heute vielfache Anwendung. Dementsprechend werden im Modell professioneller Handlungskompetenz von Baumert und Kunter (2006) neben kognitiven Aspekten auch Überzeugungen und Werthaltungen, motivationale Orientierungen und selbstregulative Fähigkeiten als einzelne Facetten von Kompetenz erfasst. Für die vorliegende Arbeit wird eine engere Definition von Kompetenz gewählt, die im Rahmen des DFG Schwerpunktprogramms „Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen“ (Klieme & Leutner, 2006b), in dem das Projekt Observe eingebettet ist, vorgenommen wurde. Kompetenzen werden demnach definiert als „*kontextspezifische kognitive Leistungsdispositionen*, die sich funktional auf Situationen und Anforderungen in bestimmten *Domänen* beziehen“ (Klieme & Leutner, 2006a, S.879; Hervorhebungen im Original). Diese Arbeitsdefinition stimmt mit der Verwendung des Kompetenzbegriffs im Rahmen von internationalen Vergleichsstudien wie PISA überein (Hartig, Klieme & Leutner, 2008), in denen Kompetenzen als „prinzipiell erlernbare, mehr oder minder bereichsspezifische Kenntnisse, Fertigkeiten und Strategien“ (Baumert, Stanat & Demmrich, 2001, S.22) verstanden werden. Eine derart auf kognitive Leistungsdispositionen konzentrierte Arbeitsdefinition scheint zunächst im Widerspruch zu der Definition von Weinert (2002) zu stehen, die motivationale, volitionale und soziale Aspekte miteinschließt. Allerdings ist die Arbeit von Weinert (2001) selbst diesbezüglich nicht kohärent (Koeppen et al., 2008). Zwar wird darin das Konzept der umfassenden Handlungskompetenz aufgegriffen, für eine vergleichende Kompetenzmessung im Large-Scale-Kontext wird aber wiederum eine Fokussierung auf eher fachbezogene kognitive Leistungsdimensionen vorgeschlagen (Jude & Klieme, 2008). Weinert betont, dass nur durch die getrennte Erfassung von kognitiven und motivationalen Aspekten ihre Interaktion systematisch analysiert werden kann (Koeppen et al., 2008; Weinert, 2001).

2.1.1.2 Nützlichkeit der gewählten Arbeitsdefinition

Für die vorliegende Arbeit wurde eine auf kognitive Aspekte eingeschränkte Arbeitsdefinition von Kompetenz gewählt. Im Folgenden wird dargelegt, warum diese Arbeitsdefinition im Hinblick auf die Überprüfung der Validität eines Messinstruments zur Erfassung von Kompetenzen als Ergebnisse von Bildungsprozessen nützlich ist.

2.1.1.2.1 Abgrenzung zu den Konstrukten Motivation und Intelligenz

Mit der Einschränkung auf kognitive Leistungsdispositionen wird durch die gewählte Arbeitsdefinition eine klare Abgrenzung von Kompetenzen zum Konstrukt der Motivation ermöglicht (Hartig, 2008). Es gilt jedoch festzuhalten, dass Kompetenzen auch gemäß der gewählten Arbeitsdefinition nicht unabhängig von motivationalen Aspekten sind. Denn bestimmte personale Voraussetzungen beeinflussen den Kompetenzerwerb (Spinath, 2012). Gemäß dem Modell der Determinanten und Konsequenzen der professionellen Kompetenz von Lehrkräften (vgl. Kunter, Kleickmann, Klusmann & Richter, 2011) wirken sich persönliche Voraussetzungen auf die Nutzung der angebotenen Lerngelegenheiten, die professionelle Kompetenz und das professionelle Verhalten aus. Darüber hinaus können motivationale Aspekte die Testsituation selbst beeinflussen z. B. mangelnde Motivation.

Durch die postulierte Erlernbarkeit und Kontextspezifität lassen sich Kompetenzen mithilfe der gewählten Arbeitsdefinition weiter von Konstrukten wie Intelligenz bzw. kognitiven Grundfähigkeiten konzeptuell abgrenzen (Hartig & Klieme, 2006). Für das Verständnis von Kompetenzen als Ergebnisse von Bildungsprozessen ist es entscheidend, dass sie grundsätzlich erlernbar und somit im Rahmen von Bildungsprozessen veränderbar sind (vgl. Shavelson, 2010). Denn nur wenn eine gezielte Förderung von Kompetenzen beispielsweise im Verlauf der Lehrerbildung möglich ist, können Kompetenzen von Absolventen zur Überprüfung der Wirksamkeit der Lehrerbildung herangezogen werden. Im Gegensatz zu Kompetenzen werden Intelligenz bzw. kognitive Grundfähigkeiten als relativ stabile Faktoren angesehen (Hartig & Klieme, 2006). Während diese durch generalisierbare Leistungsdispositionen gekennzeichnet sind, sind Kompetenzen an spezifische Kontexte gebunden, aus denen sich Anforderungen ergeben, die es zu bewältigen gilt (Hartig & Klieme, 2006).

2.1.1.2.2 Verbindung von Wissenschafts- und Berufsfeldorientierung

Im Hinblick auf die Überprüfung der Wirksamkeit der Lehrerbildung bietet der Kontextbezug von Kompetenzen in der gewählten Arbeitsdefinition einen weiteren Vorteil. Dieser ergibt sich aus der Spannung zwischen Wissenschafts- und Berufsfeldorientierung, in dem sich die Lehrerbildung befindet (Schaefers, 2002). Den einen Pol bildet die Vorbereitung auf den Beruf. Diesbezüglich beklagen Lehramtsstudierende häufig den mangelnden Praxisbezug im Studium (Kunina-Habenicht et al., 2012). Ebenso wird in Reformvorschlägen zur Lehrerbildung ein verstärkter Berufsbezug gefordert (vgl. Abel, 2010). Auf der anderen Seite spielt auch die Wissenschaftlichkeit eine zentrale Rolle in der Lehrerbildung, weil Studierende dadurch auf lebenslanges Lernen vorbereitet werden (Rutheman, 2004). Dies stellt gemäß der Empfehlung des Wissenschaftsrats (2006) ein zentrales Ziel eines jeden Universitätsstudiums dar. Gerade angesichts sich ständig verändernder und nicht exakt vorhersagbarer beruflicher Anforderungen gewinnt die akademische, forschungsbasierte Ausbildung an der Universität enorm an Bedeutung (Prenzel, Reiss & Seidel, 2011). Wissenschaftliche Konzepte können bei der Analyse und Interpretation von neuen Handlungssituationen hilfreich sein, indem sie ein fruchtbares Such- und Assimilationsraster bilden, das es ermöglicht, Fragen zu stellen, Hypothesen aufzustellen und Deutungsmuster zu finden (Messner & Reusser, 2000).

Nachdem sowohl Wissenschafts- als auch Berufsfeldorientierung entscheidend für eine effektive Lehrerbildung sind und demnach systematisch verbunden werden sollten (Prenzel, 2009; Rutheman, 2004), fordert Terhart (2000) eine berufsorientierte Wissenschaftlichkeit. Das Lehramtsstudium sollte wissenschaftlich und forschungsbasiert sein und für das professionelle Handeln im Lehrberuf qualifizieren (Prenzel, 2013). In diesem Zusammenhang ist die bildungswissenschaftliche und fachdidaktische Forschung zentral, um Evidenz bereitzustellen, die für das professionelle Handeln und Entscheiden genutzt werden kann (Prenzel, 2013). Aufbauend auf den Anforderungen des Lehrerberufs werden in der Expertise der von der Kultusministerkonferenz eingesetzten Kommission zur Perspektive der Lehrerbildung in Deutschland wissenschaftlich fundierte Kompetenzen für das erfolgreiche Handeln von Lehrpersonen abgeleitet (Terhart, 2000). Die spezifischen Kompetenzen umfassen die vier Bereiche Unterrichten, Erziehen, Diagnostizieren/Beurteilen/Evaluieren sowie Weiterentwicklung (Terhart, 2000) und entsprechen damit den vier Kompetenzbereichen der KMK-Standards der Lehrerbildung (2004) Unterrichten, Erziehen, Beurteilen und Innovieren, die auf Grundlage dieser Expertise (Terhart, 2000)

verabschiedet wurden. Der Bereich des Unterrichtens fokussiert auf fachbezogene Lehr-Lernprozesse, der Bereich des Erziehens auf die Möglichkeiten und Grenzen der Erziehungsaufgabe, der Bereich des Diagnostizierens auf das Beurteilen von Lernvoraussetzungen, Lernprozessen und Lernergebnissen, der Bereich des Innovierens auf die ständige Weiterentwicklung der eigenen beruflichen Kompetenz (KMK, 2004; Terhart, 2000).

Das Vorgehen bei der Entwicklung der KMK-Standards der Lehrerbildung zeigt, dass die Kontextspezifität von Kompetenzen, wie in der gewählten Arbeitsdefinition hervorgehoben, dazu beiträgt, einen Bezug zum späteren beruflichen Handeln der Lehrperson herzustellen und so Wissenschaftlichkeit und Berufsfeldorientierung zu verbinden.

2.1.1.2.3 Zusammenfassende Bewertung

Zusammenfassend ist zu konstatieren, dass die in der vorliegenden Arbeit gewählte Arbeitsdefinition von Kompetenzen als „*kontextspezifische kognitive Leistungsdispositionen*, die sich funktional auf Situationen und Anforderungen in bestimmten *Domänen* beziehen“ (Klieme & Leutner, 2006a, S.879; Hervorhebungen im Original) für die vorliegende Arbeit zweckmäßig ist. Sie ermöglicht eine klare Abgrenzung von Kompetenz zu den Konstrukten Motivation und Intelligenz und erlaubt durch die Kontextspezifität des Kompetenzbegriffs einen Bezug zum beruflichen Handlungsfeld von Lehrpersonen. Die zentrale Bedeutung des Kontextbezugs für Kompetenz wirft jedoch die Frage auf, wie weit oder eng dieser Kontext und dadurch auch die damit verbundene Kompetenz gefasst werden sollen. Sowohl eine sehr breite als auch eine sehr begrenzte Definition des Kontexts können dazu führen, dass ein Kompetenzkonstrukt aus wissenschaftlicher Perspektive an Nutzen verliert (Hartig, 2008). Deshalb stellt die Präzisierung des Kontextes, d. h. der Situationen und Anforderungen, auf die sich eine spezifische Kompetenz bezieht, eine wichtigen Schritt bei der Definition eines spezifischen Kompetenzkonstrukts dar (Hartig, 2008). Wird ein Kontext zu weit gefasst, besteht zum einen die Gefahr, dass die Grenzen zu bereits bestehenden allgemeinen Konstrukten wie z. B. Intelligenz verschwimmen (Hartig, 2008). Zum anderen ist es dann kaum möglich, konkrete Messinstrumente zur Erfassung dieser Kompetenz zu entwickeln (Hartig, 2008). Demzufolge gewinnt ein Kompetenzkonstrukt durch die Einschränkung des Kontextes an Präzision und Aussagekraft (Hartig, 2008). Allerdings sollte der Kontext auch nicht zu eng gefasst werden, um zu vermeiden, dass einfaches Sachwissen oder isolierte Fertigkeiten als Kompetenzen aufgefasst werden (Hartig, 2008). Weder

einfaches Sachwissen noch isolierte Fertigkeiten entsprechen dem hohen Grad an Komplexität, durch den sich Kompetenzen auszeichnen.

Abschließend ist mit Blick auf die Nützlichkeit der gewählten Arbeitsdefinition zu betonen, dass diese stets für den jeweiligen Anwendungskontext neu beurteilt werden muss und nicht universal gültig ist (Hartig, 2008). Gerade für die Entwicklung von Messinstrumenten, die auf theoretisch und messmethodisch fundierten Kompetenzmodellen basieren und Unterschiede sowie Entwicklungen in der Ausprägung von Kompetenzen empirisch abbilden sollen, stellt diese eingeschränkte Arbeitsdefinition eine nützliche Leitlinie dar (Jude & Klieme, 2008) – dementsprechend auch für das Ziel der vorliegenden Arbeit, die Validität eines derartigen Messinstruments zu überprüfen. Jedoch können im Rahmen anderer Anwendungen oder wissenschaftlicher Zusammenhänge andere Definitionen zielführender sein (Hartig, 2008).

2.1.2 Kompetenzerfassung

Im Rahmen der empirischen Überprüfung der Lehrerbildung ist es erforderlich, Kriterien für die angestrebten Ergebnisse der Bildungsprozesse festzulegen, was auf einer übergeordneten Ebene bereits mit der Verabschiedung der KMK-Standards der Lehrerbildung (2004) und den darin genannten Kompetenzbereichen geschehen ist. Darüber hinaus ist die Entwicklung von Messinstrumenten zentral, die es ermöglichen, Lernerfolge Studierender zu diagnostizieren und rückzumelden sowie Grundlagenwissen für eine gezielte Förderung bereitzustellen. Angesichts der ambitionierten Ziele sind die Ansprüche an die adäquate Messung von bildungsbezogenen Kompetenzen und ihren Veränderungen gestiegen (Artelt & Schneider, 2011). Allerding stehen den hohen Erwartungen an eine Kompetenzdiagnostik derzeit noch unzureichende theoretische Grundlagen wie z. B. fundierte theoretische Modelle über Struktur, Niveau und Entwicklung der Kompetenz sowie Messverfahren gegenüber (Jude & Klieme, 2008). Aufgrund dieses Missverhältnisses werden im Folgenden zunächst zentrale Herausforderungen bei der Erfassung von Kompetenzen beschrieben und daran anschließend verschiedene methodische Ansätze zur Erfassung von Kompetenzen als Ergebnisse von Bildungsprozessen diskutiert.

2.1.2.1 Zentrale Herausforderungen

Aufgrund der Kontextspezifität und Komplexität von Kompetenz kann eine adäquate Erfassung nur durch Berücksichtigung personaler und situativer Aspekte gelingen (Koeppen et al., 2008; Maag Merki & Werner, 2011). Dabei lassen sich vier zentrale Herausforderungen identifizieren (vgl. Klieme, Hartig & Rauch, 2008; Koeppen et al., 2008): Grundsätzlich muss beachtet werden, dass die Entwicklung eines adäquaten Messinstruments von (1) der beabsichtigten Nutzung der Messergebnisse abhängt. Entscheidend für die Erfassung von Kompetenz ist (2) die elaborierte theoretische Modellierung dieser, ergänzt durch (3) die Konstruktion adäquater psychometrischer Modelle. Darauf basierend müssen (4) konkrete Messinstrumente zur empirischen Erfassung von Kompetenzen entwickelt werden.

2.1.2.1.1 Nutzung der Messergebnisse

Es muss berücksichtigt werden, dass nicht ein universelles Messinstrument existiert, das unabhängig von den mit der Kompetenzerfassung verbundenen Zielen Kompetenzen adäquat erfasst (Koeppen et al., 2008; Pellegrino, Chudowsky & Glaser). Es werden beispielsweise Messinstrumente benötigt, die Entwicklungen auf Individualebene sensibel erfassen, um eine differenzierte Rückmeldung über den individuellen Lernerfolg zu geben, oder Messinstrumente, die Rückmeldungen auf einem aggregierten Level erlauben, um die Wirksamkeit von Seminaren oder Studiengängen zu überprüfen (Koeppen et al., 2008). Für eine detaillierte Diskussion dieses zentralen Aspekts der Kompetenzerfassung, der oft nicht im Bewusstsein von Entwicklern und Nutzern von Instrumenten zur Kompetenzerfassung ist, siehe Abschnitt 2.3.2.1.

2.1.2.1.2 Theoretische Modellierung von Kompetenzen

Die Entwicklung von adäquaten Messinstrumenten sollte auf ausgearbeiteten theoretischen Kompetenzmodellen basieren (Hartig & Klieme, 2006; Klieme et al., 2008; Koeppen et al., 2008). Diese Forderung gewinnt vor dem Hintergrund der Kontextspezifität von Kompetenz an besonderer Bedeutung. Denn nur durch eine hinreichend präzise theoretische Modellierung kann spezifiziert werden, in welchen Situationen sich auf welche Art und Weise inter- oder intraindividuelle Kompetenzunterschiede zeigen (Klieme & Hartig, 2007). Dieser spezifische Kontext liefert entscheidende Hinweise für die Operationalisierung des

Kompetenzkonstrukts, indem beispielsweise aus der Definition der relevanten Situationen mögliche Testinhalte abgeleitet werden können (Klieme & Hartig, 2007). Die grundlegende Bedeutung der theoretischen Modellierung von Kompetenzen wird deutlich, wenn Frey und Jung (2011) eine Abgrenzung zwischen Standards (z. B. KMK, 2004) und Kompetenzen vornehmen. Sie konstatieren, dass Standards zwar ein theoretisches Kompetenzmodell zugrunde liegen kann, diese jedoch in der Regel über Sammeln, Bewerten und Strukturieren durch Experten und Stakeholder entwickelt werden (Frey & Jung., 2011). Es ist Aufgabe der empirischen Bildungsforschung, Kompetenzen theoretisch zu modellieren und empirisch zu überprüfen. Um ein theoretisches Kompetenzmodell spezifizieren zu können, bedarf es allerdings einer Eingrenzung des Kontexts (Hartig, 2008). Demnach stellen die Kompetenzen, die in den KMK-Standards der Lehrerbildung formuliert wurden, eine übergeordnete Ebene dar, innerhalb derer Kompetenzen ausdifferenziert und Eingrenzungen des Kontextes vorgenommen werden müssen, um ein konkretes Messinstrument basierend auf einem fundierten theoretischen Kompetenzmodell entwickeln zu können.

Der Einbezug von individuellen und kontextspezifischen Komponenten bei der theoretischen Modellierung von Kompetenzen wirkt sich zum einen auf die Kompetenzstruktur und zum anderen auf die Beschreibung von Kompetenzniveaus aus (Koeppen et al., 2008). Darüber hinaus darf der Aspekt der Entwicklung nicht vernachlässigt werden (Koeppen et al., 2008). Mit Blick auf die theoretische Kompetenz können demnach drei Modelle unterschieden werden: Struktur-, Niveau- und Entwicklungsmodelle (Hartig & Klieme, 2006; Koeppen et al., 2008). Während Strukturmodelle darauf abzielen, zugrunde liegende Kompetenzstrukturen zu identifizieren und in Bezug zu unterschiedlichen Kontexten zu setzen, beschreiben Niveaumodelle spezifische Anforderungen, die durch Personen mit einem bestimmten Kompetenzniveau erfolgreich bewältigt werden (Koeppen et al., 2008). Beide Arten von Modellen beleuchten unterschiedliche Aspekte von Kompetenzen und sind im Idealfall komplementär (Koeppen et al., 2008). Zusätzlich sollte die Entwicklung von Kompetenzen theoretisch modelliert werden (Koeppen et al., 2008). Zentrale Fragestellungen sind dabei, ob Entwicklungen kontinuierlich oder sprunghaft stattfinden oder ob sich die Kompetenzstruktur für Novizen und Experten unterscheidet. Folglich gilt es festzuhalten, dass in Bezug auf Kompetenzerfassung ein hoher Bedarf an theoretischer Modellierung von Kompetenz, beginnend bei der Struktur, über unterschiedliche Niveaus bis hin zur Entwicklung besteht (Koeppen et al., 2008).

2.1.2.1.3 Konstruktion adäquater psychometrischer Modelle

In einem nächsten Schritt müssen auf Grundlage dieser theoretisch fundierten Kompetenzmodelle adäquate psychometrische Modelle entwickelt werden (Koeppen et al., 2008). Ein psychometrisches Modell dient dazu, die vermutete Beziehung zwischen Testverhalten und dahinterliegendem theoretischem Konstrukt zu beschreiben und darauf aufbauend ein adäquates Scoring-Verfahren für das Messergebnis festzulegen (Leutner, Hartig & Jude, 2008). Für eine detaillierte Beschreibung dieses komplexen Prozesses siehe Wilson (2005). Das Messergebnis sollte, bedingt durch die Kontextspezifität von Kompetenzen, in Bezug zur Bewältigung spezifischer domänen-relevanter Situationen stehen (Hartig & Klieme, 2006). Darüber hinaus sollten psychometrische Modelle abbilden können, dass aufgrund der Komplexität von Kompetenzkonstrukten mehrere Fähigkeiten für die Bewältigung dieser Anforderungen benötigt werden (Hartig & Klieme, 2006). In diesem Zusammenhang bietet die Item-Response-Theorie (IRT) die Möglichkeit, Personen- und Itemparameter auf einer gemeinsamen Skala zu messen (vgl. Abschnitt 4.1.3.1) und damit die Voraussetzung, individuelle und situative Faktoren in ein psychometrisches Modell zu integrieren (Koeppen et al., 2008). Derartige Modelle ermöglichen es, auch die Interaktion zwischen individuellen Fähigkeiten und situativen Anforderungen empirisch zu untersuchen (Koeppen et al., 2008). Aus den aufgeführten Gründen scheint die Verbindung von komplexen psychometrischen Modellen mit fundierten theoretischen Kompetenzmodellen für eine adäquate Kompetenzerfassung vielversprechend (Koeppen et al., 2008).

2.1.2.1.4 Entwicklung von konkreten Messinstrumenten zur empirischen Erfassung von Kompetenzen

Basierend auf einem theoretisch fundierten Kompetenzmodell und einem adäquaten psychometrischen Modell gilt es nun, ein konkretes Messinstrument zur Erfassung der spezifischen Kompetenz zu entwickeln (Koeppen et al., 2008). Dazu müssen konkrete Indikatoren für die spezifizierten Kompetenzdimensionen ermittelt werden (Maag Merki & Werner, 2011). Diese lassen sich nur dann aus einem elaborierten theoretischen Kompetenzmodell ableiten, wenn dieses konkret genug formuliert ist (Koeppen et al., 2008). Die Erfassung von Kompetenzen als Ergebnisse von Bildungsprozessen findet im Rahmen vielfältiger Anlässe (z. B. von Vergleichsstudien im Large-Scale-Kontext oder der Beurteilung individueller Qualifikationen im Rahmen von Eignungsfeststellungsverfahren) und unter Nutzung verschiedener Erhebungsmethoden (vgl. Abschnitt 2.1.2.2.1) statt (Koeppen et al.,

2008). Dabei besteht die besondere Herausforderung darin, dass die Ergebnisse der Kompetenzmessung Rückschlüsse auf die Bewältigung von Anforderungen im realen Berufsfeld erlauben (Klieme et al., 2008). Diesbezüglich stellen computerbasierte Erhebungsmethoden eine vielversprechende Alternative zu Verhaltensbeobachtungen in der beruflichen Umgebung dar. Durch die Möglichkeit, komplexe reale Situationen im Rahmen der Kompetenzmessung zu simulieren, können bei computerbasierten Messinstrumenten die Kontextspezifität und Komplexität von Kompetenzen berücksichtigt werden (Klieme et al., 2008). Entscheidend hierbei ist jedoch, dass der Kontext der Kompetenzerfassung von den Teilnehmenden als authentisch eingeschätzt wird (Shavelson, 2012). Es bedarf noch viel theoretischer und empirischer Forschung, um die komplexen computerbasierten Messverfahren angemessen mit theoretisch fundierten Kompetenzmodellen und adäquaten psychometrischen Modellen zu verbinden (Koeppen et al., 2008). Ein positives Beispiel dafür, dass dies gelingen kann, liefert die computerbasierte Erfassung von Problemlösekompetenz in PISA 2012 (Organisation for Economic Co-Operation and Development [OECD], 2014).

2.1.2.2 Diskussion verschiedener methodischer Ansätze zur Kompetenzerfassung im Bildungsbereich

In den letzten Jahren wurde eine Vielzahl empirischer Studien zur Erfassung professioneller Kompetenzen von (zukünftigen) Lehrpersonen durchgeführt, in denen unterschiedliche Erhebungsmethoden eingesetzt wurden (Kunter & Klusmann, 2010). Im Folgenden werden vier grundlegende methodische Ansätze zur Erfassung professioneller Kompetenz im Rahmen der Lehrerbildung dargestellt und hinsichtlich ihrer Vor- und Nachteile diskutiert: Tests, Selbsteinschätzungen, Dokumentationen und Beobachtungen (Maag Merki & Werner, 2011). Anschließend werden diese Ansätze hinsichtlich ihrer Eignung für den Einsatz im Large-Scale-Kontext bewertet.

2.1.2.2.1 Verschiedene methodische Ansätze zur Kompetenzerfassung

Im Rahmen der Kompetenzmessung haben sich vor allem im Bereich der Erfassung von Wissenskomponenten Leistungstests etabliert (Maag Merki & Werner, 2011). Prominente Beispiele hierfür sind Tests zur Erfassung von Fachwissen und fachdidaktischem Wissen von Mathematiklehrkräften im Rahmen des COAKTIV-Projekts (Brunner et al., 2006), zur

Erfassung von pädagogisch-psychologischem Wissen von Referendaren im Rahmen von COAKTIV-R (Voss et al., 2011) sowie zur Erfassung aller drei Wissensdomänen von Lehramtsstudierenden im Rahmen des T-KNOX-Projekts (Kleickmann et al., 2013). Über die Erfassung von Wissenskomponenten hinaus wird der Einsatz von Tests zur Kompetenzmessung eher kritisch gesehen (Maag Merki & Werner, 2011), da die zugrunde liegenden Indikatoren zu abstrakt und dekontextualisiert sind, um einen Bezug zum tatsächlichen Berufsfeld zu besitzen (Frey, 2006). Allerdings muss auch hinsichtlich der Erfassung von Wissen berücksichtigt werden, dass es entscheidend ist, inwieweit das Wissen zugänglich und auf eine spezifische Situation anwendbar ist, um sich auf tatsächliches Handeln auswirken zu können (Kersting, Givvin, Thompson, Santagata & Stigler, 2012).

Der methodische Ansatz, Kompetenzen (zukünftiger) Lehrpersonen über Selbsteinschätzungen zu erfassen, kann in mündliche Selbsteinschätzungen mittels Interviews und schriftliche Selbsteinschätzungen über Fragebögen differenziert werden (Maag Merki & Werner, 2011). Interviews können mit einzelnen Personen oder Gruppen sowie leitfadengestützt oder offen durchgeführt werden (Maag Merki & Werner, 2011). Allerdings sind sie mit einem hohen Zeitaufwand verbunden (Maag Merki & Werner, 2011) und bergen durch die Person des Interviewers Risiken für die Objektivität, Reliabilität und Validität (Bortz & Döring, 2006). Im Kontext der Kompetenzerfassung sind Interviews wenig verbreitet (Frey, 2006). Im Gegensatz dazu stellen Selbsteinschätzungen im Fragebogenformat die häufigste Methode zur Erfassung von Kompetenzen dar (Frey, 2006; Kunter & Klusmann, 2010; Maag Merki & Werner, 2011). Dies ist vornehmlich dadurch begründet, dass mit Hilfe von Fragebögen in kurzer Zeit und mit einem geringen finanziellen Aufwand große Stichproben erreicht werden können (Frey, 2006). Darüber hinaus wird die Auffassung vertreten, dass die Person selbst am besten Auskunft über die eigene Kompetenz geben kann (Frey, 2006). Dementgegen stehen mögliche Verzerrungen durch die retrospektive Erfassung und soziale Erwünschtheit (Maag Merki & Werner, 2011). Es gilt zu beachten, dass auch ohne verzerrte Erfassung eine Diskrepanz zwischen Selbsteinschätzungen und tatsächlichem Verhalten bestehen kann, da durch Selbsteinschätzungen lediglich Einstellungen erfasst werden (Bledow & Frese, 2009; Desimone, 2009). Des Weiteren birgt die Einschätzung von Fragebogenitems auf einer Likert-Skala die Gefahr, dass Personen unterschiedliche Vorstellungen davon haben, wie sich eine hohe oder niedrige Ausprägung konstituiert, und folglich bestimmte individuelle Antworttendenzen auftreten können (Bledow et al., 2009).

Unter dem Ansatz der Dokumentation werden Methoden zusammengefasst, die auf die Darlegung von Handlungen oder Ergebnissen fokussieren und damit als Reflexions- und Entwicklungsgrundlage dienen (Maag Merki & Werner, 2011). Beispiele im Rahmen der Kompetenzerfassung bei (zukünftigen) Lehrkräften sind das Unterrichtsprotokoll oder Portfolios (Maag Merki & Werner, 2011), die beide mit einem erhöhten Zeitaufwand verbunden sind. Entscheidend für diesen methodischen Ansatz ist es, konkrete Qualitätskriterien festzulegen, die inter- und intraindividuelle Vergleiche ermöglichen (Frey, 2006). Allerdings erweist sich die Erfüllung der Testgütekriterien grundsätzlich als problematisch (Frey, 2006).

Einen weiteren methodischen Ansatz stellt die direkte Beobachtung von Kompetenzen (zukünftiger) Lehrpersonen in „on-the-job"-Situationen dar (Frey, 2006). Dabei wird von den beobachteten Verhaltensweisen auf die Kompetenz geschlossen (Frey, 2006). Dieser methodische Ansatz zielt darauf ab, das Handeln als Kombination aus Wissen und Können (vgl. Baumert & Kunter, 2006) möglichst objektiv zu erfassen und scheint demnach prinzipiell vielversprechend für die Erfassung von Lehrerkompetenzen (Maag Merki & Werner, 2011; Terhart, 2007). Dabei ist entscheidend, dass konkrete Beobachtungskriterien festgelegt und die beobachtenden Personen hinreichend geschult werden. Der Kontext der Beobachtung (z. B. Unterrichtssituation) bleibt jedoch unkontrollierbar, was zu Einschränkungen in Bezug auf die Reliabilität führt (Maag Merki & Werner, 2011). Mit Hilfe von videogestützter Beobachtung wird der natürliche Handlungsablauf nicht gestört und zusätzlich kann durch die Möglichkeit der Wiederholung die Komplexität der Handlung erfasst werden (Maag Merki & Werner, 2011). Allerdings ist die direkte Beobachtung von Kompetenz mit einem erheblichen zeitlichen und finanziellen Aufwand verbunden (Oser, 1997).

2.1.2.2.2 Bewertung der methodischen Ansätze aus Large-Scale-Perspektive

Hinsichtlich der adäquaten Erfassung von Kompetenzen gilt es zu beachten, dass es nicht einen universell geeigneten Ansatz gibt, sondern, dass verschiedene Erfassungsmethoden unterschiedliche Geltungsbereiche beanspruchen können (Kunter & Klusmann, 2010). Demzufolge muss die Eignung eines methodischen Ansatzes zur Kompetenzerfassung mit Blick auf die intendierte Interpretation und Nutzung der Messergebnisse beurteilt werden. Im Folgenden werden die in Abschnitt 2.1.2.2.1 vorgestellten methodischen Ansätze zur

Kompetenzerfassung hinsichtlich ihres Einsatzes im Large-Scale-Kontext beurteilt. Studien zur Kompetenzerfassung im Large-Scale sind über ihr Ziel definiert, Rückschlüsse auf die Verteilung der untersuchten Kompetenzen in der Zielpopulation zu ziehen (Seidel & Prenzel, 2008). Deshalb basieren diese Studien auf großen Stichproben, die möglichst repräsentativ für die Zielpopulation sind (Seidel & Prenzel, 2008).

Geht es über die Erfassung von Wissenskomponenten hinaus, kommen für die Kompetenzerfassung in derart großen Stichproben aus ökonomischen Gesichtspunkten lediglich Selbsteinschätzungen im Fragebogenformat in Betracht (vgl. Abschnitt 2.1.2.2.1). Diese erfassen Kompetenzen allerdings nicht in ihrer vollen Komplexität (Frey, 2006) und Kontextspezifität, wie dies beispielsweise durch die videogestützte Verhaltensbeobachtung möglich ist. Verhaltensbeobachtungen wird demnach auch eine höhere Validität zugesprochen (Frey, 2006). Wiederum entsprechen diese nicht der Forderung nach standardisierten Instrumenten zur Kompetenzerfassung, um Befunde integrierend interpretieren zu können (Kunter & Klusmann, 2010). Die Forderung nach Standardisierung gewinnt gerade vor dem Hintergrund eines großflächigen Instrumenteneinsatzes zur Generierung von Aussagen auf systemischer Ebene (z. B. Wirksamkeit von Studiengängen) an Bedeutung. Mit dem Ziel, die Vorteile der Verhaltensbeobachtung mit den Vorteilen der Selbsteinschätzungen im Fragebogenformat zu kombinieren, wurden in den letzten Jahren vermehrt Messinstrumente entwickelt, die Lehrerkompetenzen verhaltensnah erfassen. Beispiele dafür stellen folgende Messinstrumente dar: Situational Judgement Tests z. B. zur Erfassung von Beratungskompetenz (Bruder, Keller, Klug & Schmitz, 2011), Text-Vignetten z. B. zur Erfassung adaptiver Lehrkompetenz (Beck et al., 2008), die Integration von Videobeispielen z. B. zur Erfassung von Kompetenzprofilen wie „Gruppenunterricht“ (Oser et al., 2010) und Computersimulationen z. B. ein simulierter Klassenraum mit virtuellen Schülerinnen und Schülern zur Erfassung von diagnostischer Kompetenz (Südkamp, Möller & Pohlmann, 2008). Diese Messinstrumente zeichnen sich dadurch aus, dass sie einen konkreten komplexen und standardisierten Kontext definieren, der es wahrscheinlich macht, dass die Personen die jeweils fokussierte Kompetenz zeigen. Es gilt jedoch zu überprüfen, inwieweit solche innovativen Messinstrumente zur verhaltensnahen Kompetenzerfassung im Large-Scale-Kontext einsetzbar sind.

Zusammenfassend ist zu konstatieren, dass im Kontext der Erfassung von Lehrerkompetenzen der Bedarf besteht, bisher gängige Selbsteinschätzungen im Fragebogenformat zu ergänzen (Kunter & Klusmann, 2010; Oser et al., 2010; Schaefers, 2002). Mit Blick auf die

Definition des Kompetenzbegriffs im Bereich der empirischen Bildungsforschung, in der Komplexität und Kontextspezifität zwei zentrale Charakteristika darstellen, sollten Instrumente entwickelt werden, die Kompetenzen kontextualisiert und verhaltensnah erfassen. Dafür ist eine fundierte theoretische Modellierung von Kompetenz, beginnend bei der Struktur, über unterschiedliche Niveaus bis hin zur Entwicklung, grundlegend. Insbesondere im Hinblick auf einen Einsatz derartiger Messinstrumente im Large-Scale-Kontext ist darüber hinaus eine Standardisierung dieser Instrumente erforderlich.

2.2 Fokus professionelle Unterrichtswahrnehmung

Die Kompetenzen, die in den KMK-Standards der Lehrerbildung (2004) definiert werden, sind im Hinblick auf die Entwicklung eines Messinstruments zur empirischen Erfassung zu weit gefasst. Deshalb wird im DFG-Projekt Observe ein spezifischer Fokus auf die professionelle Unterrichtswahrnehmung gesetzt, die im Kompetenzbereich Unterrichten der KMK-Standards der Lehrerbildung (2004) zu verorten ist. Professionell ist hier im Sinne von professionsrelevant zu verstehen. Es wird ein Bezugssystem geschaffen, welche Aspekte des Unterrichtsgeschehens zentral für das Lehren und Lernen sind und von Lehrpersonen wahrgenommen werden sollten. Im zweiten Kapitel des theoretischen Hintergrunds wird professionelle Unterrichtswahrnehmung zunächst definiert und ihre Struktur auf Basis bisheriger Forschungsbefunde modelliert. Anschließend wird detailliert auf die Erfassung professioneller Unterrichtswahrnehmung eingegangen.

2.2.1 Definition und Handlungsbezug[1]

Professionelle Unterrichtswahrnehmung beschreibt, wie Lehrpersonen komplexe Unterrichtssituationen beobachten und interpretieren (Seidel & Stürmer, in Druck; van Es & Sherin, 2002), und stellt einen wichtigen Bestandteil der Lehrerexpertise dar (Goodwin, 1994). Wie im Folgenden erläutert wird, ist professionelle Unterrichtswahrnehmung eine notwendige Voraussetzung für die Handlungskompetenz von Lehrpersonen und demnach zentral für den Professionalisierungsprozess zukünftiger Lehrpersonen.

Professionelle Unterrichtswahrnehmung beruht zum einen auf Wissen über effektives Lehren und Lernen (Seidel et al., 2010a) und stellt zum anderen einen Indikator dafür dar, inwieweit Lehramtsstudierende in der Lage sind, dieses Wissen auf konkrete Unterrichtssituationen anzuwenden (Seidel & Stürmer, in Druck; Sherin & van Es, 2009). Bei dieser Wissensanwendung werden Unterschiede in der Qualität und Vernetztheit der zugrunde liegenden Wissensstrukturen sichtbar. Denn über die Quantität erworbenen Wissens hinaus ist es entscheidend, wie dieses Wissen strukturiert und inwiefern es flexibel zugänglich ist (Kersting, Givvin, Sotelo & Stigler, 2010). Das Wissen über effektives Lehren und Lernen ist nach Shulman (1987) inhaltlich dem pädagogisch-psychologischen Wissen zuzuordnen.

[1] Dieser Abschnitt basiert zu Teilen auf Jahn, G., Stürmer, K., Seidel, T. & Prenzel, M. (in Druck). Professionelle Unterrichtswahrnehmung von Lehramtsstudierenden: Eine Scaling-up Studie des Observe-Projekts. *Zeitschrift für Entwicklungspsychologie und pädagogische Psychologie*.

In diesem spiegelt sich die Heterogenität, Komplexität und Kontextualität von Unterricht wider (Kersting, 2008). Pädagogisch-psychologisches Wissen ist fächerübergreifend essentiell für eine effektive Gestaltung von Lernumgebungen für Schülerinnen und Schüler (Voss et al., 2011) und deshalb wichtiger Bestandteil der universitären Lehrerbildung (Cochran-Smith, 2003). Wie bisherige Forschung zeigt, fällt es Lehramtsstudierenden vor dem Besuch entsprechender universitärer Lehrveranstaltung schwer, Wissen über effektives Lehren und Lernen auf videographierte Unterrichtsbeispiele anzuwenden (vgl. Santagata & Angelici, 2010; Santagata, Zannoni & Stigler, 2007; Star & Strickland, 2008; van Es & Sherin, 2002). Allerdings liegen auch Befunde dazu vor, dass durch die Vermittlung von Wissen über Lehren und Lernen (Stürmer, Könings & Seidel, 2013) sowie von Fertigkeiten in der Unterrichtsanalyse (Santagata & Guarino, 2011; Star & Strickland, 2008) sowohl das Wissen über effektives Lehren und Lernen als auch seine Anwendbarkeit auf Videobeispiele erhöht werden kann (Seidel & Stürmer, in Druck; Stürmer et al., 2013). Dies spricht dafür, dass professionelle Unterrichtswahrnehmung grundsätzlich erlernbar ist.

Die Qualität des Wissens über effektives Lehren und Lernen ist entscheidend, um im komplexen Unterrichtsgeschehen verantwortungsbewusst, problemlösend und souverän handeln zu können (Prenzel, 2009). Damit stellt die Anwendung dieses Wissens auf eine spezifische Unterrichtssituation im Rahmen der professionellen Unterrichtswahrnehmung eine wichtige Voraussetzung für das professionelle Handeln dar (Bromme, 1992; Seidel et al., 2010a). Nach Heckhausen (1989) kann das Handeln in drei Phasen unterschieden werden: präaktionale, aktionale und postaktionale Phase. Diese Differenzierung wurde in der dritten Projektphase für das unterrichtliche Handeln übernommen und postuliert, dass eine professionelle Wahrnehmung der Unterrichtssituation für alle drei Handlungsphasen relevant ist (vgl. Seidel & Prenzel, 2011). Übertragen auf das Unterrichten bezieht sich die präaktionale Phase auf die Planung der unterrichtlichen Handlung. Hier ist die professionelle Unterrichtswahrnehmung von zentraler Bedeutung, da mit dem Prozess der Wahrnehmung eine Vorbereitung des Handelns erfolgt (Bromme, 1992; Kersting et al., 2010). Denn es können nur Aspekte wie z. B. unterschiedliche Lernvoraussetzungen der Schülerinnen und Schüler in der Planung des eigenen Unterrichts berücksichtigt werden, die auch wahrgenommen und als lernrelevant erachtet werden. Die aktionale Phase fokussiert die tatsächliche Durchführung der unterrichtlichen Handlung. Hier wirkt sich die Qualität der zugrunde liegenden Wissensstrukturen auf die Strukturierung der tatsächlichen Handlung aus (Goodwin, 1994). Die postaktionale Phase nimmt die Reflexion der unterrichtlichen Hand-

lung in den Blick. Eine professionelle Wahrnehmung der Unterrichtssituation ermöglicht es, die Umsetzung der eigenen Planung differenziert zu beschreiben, das eigene Handeln vor dem Hintergrund von Wissen über effektives Lehren und Lernen zu erklären und Vorhersagen über Auswirkungen des eigenen Handelns auf Lernerfolge der Schülerinnen und Schüler vorherzusagen. Auf dieser Basis kann Unterricht angemessen reflektiert und daraus Optimierungsbedarf für das eigene Handeln abgeleitet werden (Oser et al., 2010).

2.2.2 Struktur professioneller Unterrichtswahrnehmung[2]

Detaillierter betrachtet berücksichtigt professionelle Unterrichtswahrnehmung zwei wissensbasierte Prozesse der Aufmerksamkeitssteuerung und der Informationsverarbeitung, die ineinander greifen: Noticing und Reasoning (Sherin, 2002). Noticing beschreibt die Fähigkeit, die Aufmerksamkeit auf relevante Situationen und Ereignisse im Unterrichtsgeschehen zu richten (van Es & Sherin, 2008). Im komplexen Unterrichtskontext ist es wichtig, Situationen und Ereignisse zu identifizieren, die für ein effektives Lehren und Lernen entscheidend sind (Seidel & Stürmer, in Druck). Die Situationen, auf die Lehrpersonen ihre Aufmerksamkeit richten, während sie Unterricht beobachten, gibt Aufschluss über das zugrunde liegende Wissen (Sherin, Jacobs & Philipp, 2011).

Der zweite Prozess, Reasoning, stellt die wissensbasierte Verarbeitung identifizierter Situationen und Ereignisse dar (Borko, 2004; Sherin, 2007; van Es & Sherin, 2002). Diese Fähigkeit erlaubt Rückschlüsse auf die Qualität und Vernetztheit der zugrunde liegenden Wissensstruktur und deren Anwendung auf die aktuelle Klassenzimmersituation (Borko, 2004; Seidel & Stürmer, in Druck). Bei Experten und Novizen zeigen sich qualitative Unterschiede hinsichtlich der Wahrnehmung von Unterricht (Berliner, 2001; Carter, Cushing, Sabers, Stein & Berliner, 1988), die auf die Struktur des zugrunde liegenden Wissens zurückführbar sind. Es können drei qualitativ unterschiedliche Verarbeitungsebenen unterschieden werden: eine lehr-lernrelevante Unterrichtssituation differenziert zu beschreiben, sie auf Basis wissenschaftlicher Theorien und Befunde vor dem Hintergrund von Lernwirksamkeit zu erklären und ihre Auswirkung auf den Lernerfolg der Schülerinnen und Schüler vorherzusagen. Studien (z. B. Carter et al., 1988; Seidel & Prenzel, 2007; Seidel & Stürmer, in Druck) zeigen, dass Novizen im Gegensatz zu Experten vornehmlich auf der Ebene naiven Beschreibens operieren und zu übergeneralisierenden Beurteilungen tendie-

[2] Dieser Abschnitt basiert zu Teilen auf Jahn et al. (in Druck).

ren. Daher wird angenommen, dass die Aspekte Erklären und Vorhersagen stärker vernetzte Wissensstrukturen erfordern und mit einem höheren Maß an Expertise einhergehen (Berliner, 2001; Seidel & Shavelson, 2007; Seidel & Stürmer, in Druck). Auf Grundlage eines integrierten und vernetzten Wissens um effektives Lehren und Lernen sind Lehrpersonen in der Lage, Verbindungen zwischen Wissenselementen herzustellen und damit spezifische Unterrichtssituationen differenziert zu erklären sowie Vorhersagen für weitere Lehr-Lernprozesse zu treffen. Hinsichtlich der Struktur werden in einigen Studien die drei genannten Aspekte differenziert (z. B. Seidel & Shavelson, 2007), in anderen werden die beiden Aspekte Erklären und Vorhersagen als integrierende Fähigkeit zusammengefasst, da beide Aspekte kognitive Prozesse höherer Ordnung erfordern (z. B van Es & Sherin, 2008).

2.2.3 Erfassung professioneller Unterrichtswahrnehmung

Im Folgenden wird zunächst das DFG-Projekt kurz in die Forschung zur professionellen Unterrichtswahrnehmung eingeordnet. Anschließend wird ausführlich dargestellt, wie professionelle Unterrichtswahrnehmung im Rahmen dieses Projekts erfasst wird.

2.2.3.1 Einordnung des DFG-Projekts Observe in die Forschung zur professionellen Unterrichtswahrnehmung

Das Konstrukt der professionellen Unterrichtswahrnehmung stammt aus Nordamerika. Dort wurde von Goodwin (1994) der Begriff der „professional vision“ geprägt, der von van Es und Sherin (2002) auf den Bereich der Lehrerbildung übertragen wurde. Es liegen Befunde vor, die auf einen positiven Zusammenhang zwischen professioneller Unterrichtswahrnehmung der Lehrperson und dem Lernzuwachs der Schülerinnen und Schüler hindeuten (Kersting et al., 2010; Kersting et al., 2012). Dieser Zusammenhang scheint durch die Unterrichtsqualität mediiert zu werden (Kersting et al., 2012). Im Mittelpunkt der vornehmlich qualitativen Forschung steht die gezielte Förderung der professionellen Wahrnehmung durch die Analyse von videographierten Unterrichtssequenzen. Der Fokus liegt dabei meist auf Videosequenzen, die Mathematikunterricht zeigen. Positive Effekte konnten dabei bereits durch Weiterbildungsangebote für Lehrpersonen, z. B. Video-Clubs, erzielt werden (vgl. Sherin & Han, 2004; Sherin & van Es, 2009; van Es, 2009; van Es &

Sherin, 2008). Auch für Lehramtsstudierende wurden erfolgreiche Interventionen entwickelt, um die Fähigkeit, Unterricht professionell wahrzunehmen, im Rahmen der universitären Lehrerbildung zu fördern (z. B. Santagata & Angelici, 2010; Santagata & Guarino, 2011; Star & Strickland, 2008). Im Hinblick auf die Erfassung hat Kersting (2008) einen vielversprechenden Ansatz entwickelt, in dem standardisierte Videoclips als Item-Prompts fungieren und mit einer offenen Frage verbunden werden, die auf die Analyse der beobachteten Unterrichtssituation abzielt. Dadurch wird die Anwendung integrierten Wissens auf einen konkreten Kontext erfasst. Dieser Ansatz wird in ähnlicher Art und Weise auch für andere Studien in diesem Forschungsbereich genutzt (z. B. Santagata & Angelici, 2010; Santagata & Guarino, 2011; Star & Strickland, 2008). In diesem Forschungskontext existiert jedoch kein Instrument, das professionelle Unterrichtswahrnehmung quantitativ erfasst und zum Einsatz im Large-Scale-Kontext geeignet ist.

Nachdem sich das Konzept der „professional vision“ erfolgreich im nordamerikanischen Raum etabliert hatte, um Wissenserwerbsprozesse von Studierenden im Rahmen der universitären Lehrerbildung zu beschreiben, wurde das Konzept von Seidel und Kollegen (2010) unter dem Begriff Professionelle Unterrichtswahrnehmung erfolgreich nach Deutschland übertragen. Im Rahmen des DFG-Projekts Observe wurde ein videobasiertes Online-Tool (Seidel et al., 2010b) mit dem Ziel entwickelt, professionelle Unterrichtswahrnehmung kontextualisiert und gleichzeitig standardisiert zu erfassen (Seidel et al., 2010a). Dieser innovative Ansatz wird in Deutschland mittlerweile vielfach rezipiert. Dabei wird entweder ganzheitlich pädagogisch-psychologisches Wissen fokussiert (z. B. König, Blömeke, Klein, Suhl & Busse, 2014; Plöger & Scholl, 2014) oder ein spezifischer Bereich wie Klassenführung (z. B. Gold, Förster & Holodynski, 2013; Syring et al., 2013). Neben Messinstrumenten werden auch Seminare entwickelt, um professionelle Unterrichtswahrnehmung im Rahmen der universitären Lehrerbildung zu fördern (z. B. Gold et al., 2013).

2.2.3.2 Erfassung professioneller Unterrichtswahrnehmung im Rahmen des DFG-Projekts Observe

Ziel des DFG-Projekts Observe war es, ein Messinstrument zu entwickeln, dass grundsätzlich den Anforderungen an eine Kompetenzmessung (vgl. Abschnitt 2.1.2.2.2) entspricht und darüber hinaus im Large-Scale-Kontext einsetzbar ist. Basierend auf einem fundierten

theoretischen Modell sollte professionelle Unterrichtswahrnehmung kontextualisiert und verhaltensnah, aber dennoch standardisiert erfasst werden. Im Folgenden wird zunächst auf die Umsetzung dieser drei zentralen Aspekte eingegangen bevor der konkrete Aufbau des Tools Observer beschrieben wird.

2.2.3.2.1 Theoretische Modellierung und deren Umsetzung im Instrument

Die Forderung nach einer fundierten theoretischen Modellierung wurde in Bezug auf die Entwicklung des Tools Observer wie folgt umgesetzt: Die Befunde bisheriger qualitativer Forschung (vgl. Abschnitt 2.2.2) wurden dazu genutzt, ein Kompetenzstrukturmodell professioneller Unterrichtswahrnehmung zu entwickeln. Dieses bildet die Grundlage für die Entwicklung des Instruments, mit dessen Hilfe wiederum die theoretisch angenommene Kompetenzstruktur empirisch geprüft werden sollte. Gemäß Sherin (2002) kann professionelle Unterrichtswahrnehmung in zwei ineinandergreifende Prozesse differenziert werden: Noticing und Reasoning (Sherin, 2002). Während Noticing die Fähigkeit beschreibt, die Aufmerksamkeit auf relevante Situationen im Unterricht zu richten (van Es & Sherin, 2008), bezieht sich Reasoning auf die wissensbasierte Verarbeitung bereits identifizierter Unterrichtssituationen, die für effektives Lehren und Lernen relevant sind (Borko, 2004; Sherin, 2007; van Es & Sherin, 2002). Mit dem Ziel, ein standardisiertes Messinstrument bereitzustellen, wurde der Prozess des Reasonings fokussiert, da dieser einen Indikator für die Qualität und Vernetztheit der zugrunde liegenden Wissensstruktur darstellt (Borko, 2004; Seidel & Stürmer, in Druck) und sich quantitativ erfassen lässt. Der Prozess des Noticings wurde standardisiert, um sicher zu stellen, dass bei der Bearbeitung des Instruments alle Teilnehmenden dieselben Unterrichtssituationen wissensbasiert weiterverarbeiten (vgl. Abschnitt 2.2.3.2.3). Aus Studien mit Experten und Novizen (Berliner, 2001; Carter et al., 1988), kann auf qualitative Unterschiede hinsichtlich der Verarbeitung von identifizierten Unterrichtssituationen geschlossen werden. Dementsprechend werden drei qualitativ unterschiedliche Verarbeitungsebenen postuliert: die Unterrichtssituation differenziert zu beschreiben, sie vor dem Hintergrund von Lernwirksamkeit zu erklären und ihr Auswirkungen auf den Lernerfolgs der Schülerinnen und Schüler vorherzusagen. Dies entspricht einem dreidimensionalen Strukturmodell professioneller Unterrichtswahrnehmung. Um alle drei theoretisch angenommenen Dimensionen empirisch abzubilden, wurden Rating-Items entwickelt, die darauf abzielen, die Unterrichtssituation zu beschreiben, zu erklären und vorherzusagen vgl. (Seidel et al., 2010a).

2.2.3.2.2 Kontextualisierte und verhaltensnahe Erfassung

Um professionelle Unterrichtswahrnehmung kontextualisiert zu erfassen, wurden videographierte Unterrichtssequenzen in das Messinstrument integriert (vgl. Seidel et al., 2010a). Neben den vielen Vorteilen, die die Analyse von Videobeispielen im Rahmen der universitären Lehrerbildung bietet (vgl. Santagata et al., 2007), besitzen Videos auch bezüglich der Erfassung von Kompetenz großes Potential. Sie bilden die vielschichtigen Interaktionen im Unterrichtsgeschehen in ihrer vollen Komplexität ab (Desimone, 2009) und können als Prompts genützt werden, um Wissen über effektives Lehren und Lernen zu aktivieren (Kersting, 2008). Damit geht das Messinstrument über die Erfassung deklarativen Wissens hinaus, da es erforderlich ist, das Wissen auf die spezifische Unterrichtssituationen im Videobeispiel anzuwenden. Somit nähert sich das Messinstrument einer Erfassung von Verhalten an. Gemäß dem Ansatz der Approximations of Practice (Grossman et al., 2009), in dem verschiedene Tätigkeiten nach aufsteigender Authentizität als Annäherungen an die Praxis des Unterrichtens differenziert werden, können auch bei der Kompetenzmessung verschiedene Annäherungen an die Erfassung von Verhalten unterschieden werden. Demzufolge stellt die Erfassung der Wissensanwendung auf spezifische Unterrichtsituationen eine Form der verhaltensnahen Kompetenzerfassung dar. Denn nur Wissen, das zugänglich und damit anwendbar ist, kann unterrichtliches Handeln steuern (Kersting et al., 2012).

2.2.3.2.3 Standardisierte Erfassung[3]

Mit Blick auf den Einsatz des Messinstruments im Large-Scale-Kontext, um größere Stichproben von Lehramtsstudierenden zu untersuchen und Aussagen auf systemischer Ebene (z. B. standortübergreifende Überprüfung der Wirksamkeit von Lehrveranstaltungen) tätigen zu können, bedarf es einer standardisierten Erfassung. Dieses Ziel wurde bei der Entwicklung des Tools Observer auf drei Arten umgesetzt. Zum einen wird, im Gegensatz zu bisherigen Studien, die Videos als Item-Prompts verwenden (v. a. Kersting, 2008; Kersting et al., 2010; Kersting et al., 2012), der Prozess des Reasonings mithilfe von Rating-Items im geschlossenen Antwortformat erfasst. Damit wird die Datenerhebung und -auswertung weiter standardisiert. Zum anderen wird durch die Integration von Videobeispielen ein standardisierter Kontext geschaffen. Darüber hinaus wird eine Eingrenzung des

[3] Dieser Abschnitt basiert zu Teilen auf Jahn et al. (in Druck).

Kontexts vorgenommen, da die Komplexität von Lehr- und Lernprozessen im Unterricht den Rahmen eines standardisierten Messinstruments bei weitem überschreitet. Dazu wurde ein Wissensfokus gesetzt, vor dessen Hintergrund professionelle Unterrichtswahrnehmung erfasst wird und in dessen Kontext die Ergebnisse zu interpretieren sind (Kane, 1994). An dieser Stelle ist anzumerken, dass dieser Fokus einen von vielen möglichen darstellt. Im Rahmen des Projekts wurde die Auswahl relevanten Wissens im Bereich effektiven Lehren und Lernens basierend auf Befunden der Unterrichtseffektivitätsforschung vorgenommen. Diese zeigen, dass verschiedene Unterrichtskomponenten, die nach Shulman (1987) inhaltlich dem pädagogisch-psychologischen Wissen zuzuordnen sind, entscheidend für die Unterstützung von Lernprozessen der Schülerinnen und Schüler sind (z. B. Seidel & Shavelson, 2007). Um das Messinstrument nicht zu überladen, wurden drei dieser Unterrichtskomponenten ausgewählt: Zielorientierung, Lernbegleitung und Lernatmosphäre (vgl. Seidel & Stürmer, in Druck). Diese wurden noch weiter spezifiziert und auf jeweils zwei Teilaspekte begrenzt. Zielorientierung wird als Ziel- und Anforderungsklärung definiert. Die Videoclips werden dahingehend beurteilt, ob die Lehrperson die Lernziele klärt und eine kohärente Struktur vorgibt. Lernbegleitung wird durch die Art der Fragen und des Feedbacks charakterisiert. Für die Videoclips wird eingeschätzt, inwieweit die Lehrperson qualitativ hochwertige Fragen stellt, die eigenständige Lernprozesse anregen, und Beiträge der Schülerinnen und Schüler aufgreift und fortführt. Lernatmosphäre setzt sich aus den Aspekten Humor als unterrichtliches Mittel und Ernstnehmen der Schülerinnen und Schüler zusammen. Für die Videoclips wird beurteilt, inwieweit die Lehrperson gezielt Humor einsetzt, um eine positive Lernatmosphäre zu schaffen und die Schülerinnen und Schüler ernst nimmt, indem sie respektvoll und wertschätzend mit ihnen spricht. Vor dem Hintergrund der Selbstbestimmungstheorie (Deci & Ryan, 1993) fördert das Bereitstellen dieser drei Unterrichtskomponenten grundlegende Bedingungen für Lernprozesse, wie das Erleben von Kompetenz, Autonomie und sozialer Eingebundenheit (Seidel & Stürmer, in Druck). Denn die Schülerinnen und Schüler wissen durch das Transparentmachen von Zielen und Anforderungen genau, was auf sie zukommt, befinden sich in einer positiven Lernatmosphäre, in der sie nicht bloßgestellt werden, und werden durch qualitativ hochwertige Fragen zu eigenständigem Denken angeregt und dabei durch konstruktives Feedback unterstützt. Über die Befriedigung dieser grundlegenden Bedingungen hinaus, die sich motivationsfördernd auswirken, ermöglichen diese Unterrichtskomponenten Vorhersagen über potentielle Auswirkungen auf weitere Lernprozesse. Diese Fokussierung auf das Wissen

über Zielorientierung, Lernbegleitung und Lernatmosphäre wird im Tool Observer umgesetzt, indem vorgegeben wird, worauf die Teilnehmenden in den Videobeispielen ihre Aufmerksamkeit richten sollen. Dies geschieht, indem die Videoclips in Rating-Items eingebettet sind, die pro Videoclip auf jeweils zwei der drei Unterrichtskomponenten fokussieren. Damit wird der Prozess des Noticings vorgegeben und somit standardisiert.

2.2.3.2.4 Aufbau des Tools Observer[4]

Das videobasierte Online-Tool Observer (Seidel et al., 2010b) zeichnet sich durch eine kontextualisierte und gleichzeitig standardisierte Erfassung professioneller Unterrichtswahrnehmung aus (Seidel et al., 2010a). Dazu werden videographierte Unterrichtssequenzen als item-unabhängige Prompts genutzt (Seidel & Stürmer, in Druck) und in standardisierte Rating-Items eingebettet. Das Instrument umfasst sechs zwei- bis vierminütige Videoclips, die Unterricht in der Sekundarstufe aus den Fächern Mathematik, Physik, Französisch und Geschichte zeigen. Nachdem das Wissen um Zielorientierung, Lernbegleitung und Lernatmosphäre fächerübergreifend essentiell für eine effektive Gestaltung von Lernumgebungen ist (vgl. Voss et al., 2011), wurden gezielt Videoclips mit Unterrichtssituationen verschiedener Fächer ausgewählt.

Die Videoclips wurden von Experten in einem mehrstufigen Verfahren anhand der folgenden drei Kriterien ausgewählt: Sie sollten (1) authentische Beispiele von Unterrichtspraxis darstellen und repräsentativ für die drei ausgewählten Unterrichtskomponenten Zielorientierung, Lernbegleitung und Lernatmosphäre sein, (2) Lehrerwissen aktivieren, wobei eine Balance zwischen Aktivierung und Überforderung angestrebt wurde, und (3) relevant für den Lernerfolg der Schülerinnen und Schüler sein, sei es durch positive, ambivalente oder negative Beispiele von Unterrichtspraxis (Seidel & Stürmer, in Druck). Unter Anwendung der beschriebenen Kriterien durch drei Experten der Unterrichtsforschung wurden in einem ersten Schritt aus frei zugänglichen, deutschsprachigen Videoportalen Videoclips zusammengetragen und eine Vorauswahl von 86 Clips getroffen (Seidel & Stürmer, in Druck). In einem zweiten Schritt ordneten die Experten unabhängig voneinander die Videoclips zwei der drei Unterrichtskomponenten im Hinblick auf jeweils zwei spezifische Teilaspekte zu (Zielorientierung: Zielklärung und Anforderungsklärung; Lernbegleitung: Art der Fragen und Feedback; Lernatmosphäre: Humor als unterrichtliches Mittel und Ernstnehmen der

[4] Dieser Abschnitt basiert zu Teilen auf Jahn et al. (in Druck).

Schülerinnen und Schüler) (Seidel & Stürmer, in Druck). Die Zuordnung wurde anschließend diskutiert und konsensvalidiert, woraus eine finale Auswahl von zwölf Videoclips resultierte (Seidel & Stürmer, in Druck). Um die Bearbeitungszeit auf maximal 90 Minuten zu begrenzen, wurde eine finale Version des Instruments mit sechs Videoclips (2x Physik, 1x Sprache, 2x Mathematik, 1x Geschichte) erstellt (Seidel & Stürmer, in Druck). Die Auswahl der Videoclips wurde basierend auf Befunden einer Pilotierungsstudie mit $N = 40$ Lehramtsstudierenden und einer weiteren Studie mit $N = 119$ Lehramtsstudierenden (vgl. Abschnitt 2.3.2.3) sowie mit Blick auf ein ausgewogenes Verhältnis zwischen repräsentierten Unterrichtskomponenten und -fächern vorgenommen.

Zu Beginn der Bearbeitung des Instruments erhalten die Lehramtsstudierenden eine kurze Definition der drei Unterrichtskomponenten Zielorientierung, Lernbegleitung und Lernatmosphäre. Im Anschluss sehen sie die Videoclips an und schätzen diese anhand von Multiple-Choice-Items auf einer vierstufigen Likert-Skala (‚1' trifft nicht zu/‚4' trifft zu) ein. Zusätzlich gibt es die Option, keine Angabe (k.A.) zu wählen. Die Videoclips werden durch kurze Hintergrundinformationen bezüglich Jahrgangsstufe, Unterrichtsfach und Thema der Stunde eingeleitet. Es besteht die Möglichkeit, jeden Videoclip mehrmals anzusehen. Pro Videoclip werden zwei der drei Unterrichtskomponenten eingeschätzt. Der Aufbau des Instruments ist exemplarisch an zwei Screenshots zum ersten Videoclip in

Abbildung 1 illustriert. Der erste Videoclip zeigt eine Unterrichtssequenz zu Beginn einer Physikstunde zum Thema Optik. Für die Einschätzung des Videoclips werden die zwei Unterrichtskomponenten Zielorientierung und Lernatmosphäre fokussiert. Zielorientierung wird anhand von jeweils neun Rating-Items zu den Teilaspekten Zielklärung und Anforderungsklärung eingeschätzt. Pro Teilaspekt zielen die ersten drei Rating-Items jeweils darauf ab, die gesehene Unterrichtssituation differenziert zu beschreiben (z. B. Der Lehrer ordnet das Thema in einen übergeordneten Zusammenhang ein), die nächsten drei, diese vor dem Hintergrund von Wissen über effektives Lehren und Lernen zu erklären (z. B. Die Schülerinnen und Schüler haben die Möglichkeit, ihr Vorwissen zum Thema zu aktivieren), und die letzten drei, Vorhersagen über den Lernerfolg der Schülerinnen und Schüler zu treffen (z. B. Die Schülerinnen und Schüler können ihren Lernprozess auf das Lernziel ausrichten).

Abbildung 1. Illustration des Aufbaus des Tools Observer am Beispiel von zwei Screenshots zum ersten Videoclip.

In Abbildung 2 ist schematisch dargestellt, wie sich die Rating-Items über das gesamte Instrument auf die Aspekte professioneller Unterrichtswahrnehmung, die ausgewählten Unterrichtskomponenten sowie die einzelnen Videoclips verteilen. Aus drei Rating-Items pro Teilaspekt für Beschreiben, Erklären und Vorhersagen (vgl. Beispiel Zielklärung in

Abbildung *1*) ergeben sich sechs Rating-Items pro Unterrichtskomponente. Bei drei Unterrichtskomponenten bedeutet dies, dass die Rating-Items insgesamt mit jeweils 18 Rating-Items im gleichen Maß auf die drei Aspekte professioneller Unterrichtswahrnehmung abzielen. Die Rating-Items unterscheiden sich lediglich im fokussierten Inhalt der jeweiligen Unterrichtskomponente und verteilen sich mit jeweils 18 Rating-Items gleichmäßig auf Zielorientierung, Lernbegleitung und Lernatmosphäre. Demzufolge gibt es insgesamt 54 Rating-Items, die sowohl alle Aspekte professioneller Unterrichtswahrnehmung als auch alle ausgewählten Unterrichtskomponenten abdecken. Bei insgesamt sechs Videoclips, die jeweils zwei Unterrichtskomponenten fokussieren, werden damit alle 54 Rating-Items viermal eingesetzt, woraus sich eine Gesamtzahl von 216 Items ergibt. Im Rahmen der Itementwicklung wurde festgestellt, dass die Items lokal stochastisch unabhängig sind, was eine Datenauswertung mit Analyseverfahren basierend auf der Item-Response-Theorie ermöglicht (Seidel & Stürmer, in Druck).

	Beschreiben *(18 Rating-Items)*			**Erklären** *(18 Rating-Items)*			**Vorhersagen** *(18 Rating-Items)*		
	Zielorien-tierung *(6 Rating-Items)*	Lernbe-gleitung *(6 Rating-Items)*	Lernat-mosphäre *(6 Rating-Items)*	Zielorien-tierung *(6 Rating-Items)*	Lernbe-gleitung *(6 Rating-Items)*	Lernat-mosphäre *(6 Rating-Items)*	Zielorien-tierung *(6 Rating-Items)*	Lernbe-gleitung *(6 Rating-Items)*	Lernat-mosphäre *(6 Rating-Items)*
Einsatz: 4 x 54 Rating-Items = 216 Items									
Clip 1	X		X	X		X	X		X
Clip 2	X	X		X	X		X	X	
Clip 3		X	X		X	X		X	X
Clip 4	X		X	X		X	X		X
Clip 5	X	X		X	X		X	X	
Clip 6		X	X		X	X		X	X

Abbildung 2. Überblick über die Verteilung der Rating-Items der finalen Version des videobasierten Online-Tools Observer auf die Aspekte professioneller Unterrichtswahrnehmung, die ausgewählten Unterrichtskomponenten sowie die einzelnen Videoclips.

In Ermangelung eindeutiger forschungsbasierter Qualitätskriterien von Unterricht wurde als Bezugsnorm für die Rating-Items eine Expertennorm herangezogen. Dieser Ansatz basiert auf der Annahme, dass Experten ihr gut strukturiertes und integriertes Wissen nutzen, um professionelle Aufgaben wie die Einschätzung von Unterrichtssequenzen zu bewältigen (vgl. Ericsson, Krampe & Tesch-Römer, 1993). Eine detaillierte Begründung der Wahl dieses Ansatzes ist bei Seidel und Stürmer (in Druck) nachzulesen. Die Expertennorm wurde festgelegt, indem drei Experten mit 100 bis 400 Stunden Erfahrung in der Beobachtung von Unterricht unabhängig voneinander alle 216 Items einschätzten. Die Konsistenz dieser Experten-Ratings ist mit einem durchschnittlichen Cohen's Kappa von $\kappa = .79$ zufriedenstellend (Seidel et al., 2010a). Für die Items, bei denen keine Übereinstimmung der drei Experten vorlag, wurde eine Konsensvalidierung durchgeführt. Die Übereinstimmung mit der Expertennorm wurde wie folgt umkodiert: ‚1' Expertennorm getroffen und ‚0' Expertennorm nicht getroffen. Die Wahl keine Angabe wurde als ‚0' Expertennorm nicht getroffen kodiert. Dadurch entsteht eine 0-1-Datenmatrix. In der Skalierungsstudie erwies sich das Anlegen dieses strengen Kriteriums einer Umkodierung unter Berücksichtigung der Tendenz mit ‚2' Expertennorm getroffen, ‚1' korrekte Tendenz und ‚0' Expertennorm nicht getroffen als klar überlegen (vgl. Seidel & Stürmer, in Druck). Es wurden wesentlich höhere Reliabilitäten der einzelnen Skalen erreicht und deutlich mehr Varianz erklärt (Seidel & Stürmer, in Druck).

2.2.3.2.5 Zusammenfassung

Zusammenfassend ist zu konstatieren, dass im Rahmen der Entwicklung des Tools Observer die Forderungen nach theoretischer Modellierung sowie kontextualisierter, verhaltensnaher und standardisierter Kompetenzerfassung umgesetzt wurden. Auf Grundlage eines fundierten theoretischen Modells professioneller Unterrichtswahrnehmung, das auf Befunden qualitativer Studien basiert, wurden Rating-Items entwickelt, um die theoretische Kompetenzstruktur empirisch abbilden zu können. Des Weiteren wurden videographierte Unterrichtssequenzen als Prompts genutzt (vgl. Kersting, 2008), um Wissen über effektives Lehren und Lernen zu aktivieren und auf einen spezifischen Kontext anzuwenden. Eine Standardisierung der Erfassung wurde hinsichtlich dreier Aspekte umgesetzt: (1) Durch die Integration von Videobeispielen wird ein standardisierter Kontext geschaffen, (2) die Videobeispiele sind in Rating-Items mit geschlossenem Antwortformat eingebettet und (3) für die Interpretation der Videobeispiele ist ein spezifischer Wissensfokus – Zielorientierung, Lernbegleitung und Lernat-

mosphäre – vorgegeben. Die Fähigkeit, Unterrichtssituationen vor dem Hintergrund der gezeigten Zielorientierung, Lernbegleitung und Lernatmosphäre zu beschreiben, zu erklären und vorherzusagen, verdeutlicht, in welchem Ausmaß (zukünftige) Lehrpersonen in der Lage sind, ihr Wissen über diese Unterrichtskomponenten auf konkrete Unterrichtssituationen anzuwenden (Seidel & Stürmer, in Druck). Folglich geht das video-basierte Online-Tool Observer über eine reine Erfassung deklarativen Wissens hinaus, indem es die Wissensanwendung fokussiert, was einer verhaltensnahen Kompetenzerfassung entspricht.

2.3 Anforderungen an ein Messinstrument zur Erfassung professioneller Unterrichtswahrnehmung

Das dritte Kapitel des theoretischen Hintergrunds hat zum Ziel, spezifische Anforderungen an ein Messinstrument zur Erfassung professioneller Unterrichtswahrnehmung zu definieren. Aus psychometrischer Perspektive werden an ein Messinstrument klassischerweise drei Hauptanforderungen gestellt: Objektivität, Reliabilität und Validität z. B. (Lienert & Raatz, 1998). Dementsprechend soll das Tool Observer professionelle Unterrichtswahrnehmung objektiv, reliabel und valide erfassen. Im Folgenden wird zunächst auf die Anforderungen bezüglich Objektivität und Reliabilität sowie die Gewährleistung dieser Anforderungen im Rahmen des DFG-Projekts Observe eingegangen. Der Schwerpunkt liegt anschließend auf den Anforderungen hinsichtlich der Validität und der Überprüfung dieser Anforderungen.

2.3.1 Objektivität und Reliabilität

2.3.1.1 Definition

Objektivität bezieht sich auf das Ausmaß, in dem das Messergebnis unabhängig von jeglichen Einflüssen außerhalb der untersuchten Person ist (Rost, 2004). Sie umfasst sowohl die Testdurchführung bzw. den Instrumenteneinsatz als auch die Auswertung und Interpretation (Lienert et al., 1998). Entsprechend werden Durchführungs-, Auswertungs- und Interpretationsobjektivität differenziert (Bühner, 2011). Im Rahmen der Kompetenzerfassung im bildungswissenschaftlichen Kontext wird zur Gewährleistung der Objektivität am häufigsten die Strategie angewandt, alle notwendigen Schritte im Rahmen der Durchführung, Auswertung und Interpretation zu standardisieren und zu dokumentieren (Leutner et al., 2008). Reliabilität ist definiert als das Ausmaß in dem die Unterschiede in den Messergebnissen die tatsächlichen Merkmalsunterschiede, die gemessen werden sollen, widerspiegeln (Furr & Bacharach, 2008). Damit stellt Reliabilität sozusagen eine Eigenschaft des Messergebnisses dar, die auf unterschiedliche Arten basierend auf Korrelationskoeffizienten oder Verhältnissen der Varianzen berechnet werden kann (Furr & Bacharach, 2008). Bezogen auf die Kompetenzerfassung im bildungswissenschaftlichen Kontext werden bei der Beurteilung der Reliabilität Koeffizienten, die auf der klassischen Test-Theorie basieren, der Multidimensionalität vieler Kompetenzkonstrukte nicht gerecht. Als Alternative können im Rahmen der Item-Response-Theorie Maße für den Fit verschiedener Modelle herangezogen werden (Leutner et al., 2008).

2.3.1.2 Objektivität und Reliabilität des Tools Observer

Die Durchführungsobjektivität des Tools Observer ist dadurch gewährleistet, dass das Instrument in eine Online-Plattform integriert ist, auf der zu Beginn der Bearbeitung eine standardisierte Instruktion gegeben wird. Dadurch wird der potentielle Effekt eines Testleiters minimiert und Anonymität gewährleistet (Treiblmaier, 2010), was dazu führt, dass sozial erwünschtes Antwortverhalten abnimmt (Paulhus, 1984). Die Auswertungsobjektivität ist dahingehend sichergestellt, dass für die Auswertung der geschlossenen Rating-Items eine Expertennorm als Referenz herangezogen wird. Die Entwicklung dieser Expertennorm ist unter Abschnitt 2.2.3.2.4 beschrieben. Für jeden Studierenden wird eine prozentuale Übereinstimmung mit der Expertennorm berechnet, wodurch eine objektive Interpretation der Messergebnisse erzielt wird.

Im Hinblick auf die Reliabilität wurde im Rahmen des Projekts eine Studie zur Retest-Reliabilität durchgeführt (vgl. Seidel & Stürmer, in Druck). Bei $N = 20$ Lehramtsstudierende, die das Instrument randomisiert zugeteilt im Abstand von einer, zwei oder drei Wochen zweimal bearbeiteten, zeigten sich keine signifikanten Unterschiede in der erfassten Kompetenz (Seidel & Stürmer, in Druck). Demzufolge kann angenommen werden, dass das Instrument ohne weitere Intervention professionelle Unterrichtswahrnehmung über die Zeit stabil erfasst.

Zusammenfassend ist festzuhalten, dass die Bedingungen, unter denen das Tool Observer eingesetzt, ausgewertet und interpretiert wird, auf eine hohe Objektivität schließen lassen und die bisherigen Ergebnisse der Überprüfung der Reliabilität auf eine reliable Erfassung professioneller Unterrichtswahrnehmung hindeuten.

2.3.2 Validität

Im Gegensatz zur Reliabilität bezieht sich die Validität nicht direkt auf das Messergebnis selbst, sondern auf dessen Interpretation und Nutzung (Furr & Bacharach, 2008). Denn ein Messinstrument wird meist für einen bestimmten Zweck entwickelt, der wiederum eine bestimmte Verwendung vorgibt (Kane, 2013a). Damit ist die beabsichtigte Nutzung und Interpretation des Messergebnisses entscheidend und die Validität kann nicht losgelöst davon beurteilt werden. Die beiden Konzepte Reliabilität und Validität sind konzeptuell und statistisch eng verbunden, wobei Reliabilität eine notwendige, aber keine hinreichende Bedingung für Validität darstellt (Furr & Bacharach, 2008). Die Frage nach der Validität ist ein zentrales

Thema in der Psychometrie (Furr & Bacharach, 2008) und gleichzeitig die größte Herausforderung (Kingston, 2007). Denn ohne Validierungen von Messinstrumenten sind alle Schlüsse, die aus entsprechenden Messungen gezogen werden, problematisch und nur begrenzt nützlich (Zumbo, 2007). Gerade vor dem Hintergrund, dass Ergebnisse psychometrischer Messverfahren oft als Grundlage für weitreichende Entscheidungen dienen, sei es auf individueller Ebene (z. B. Bewerberauswahl über Assessment Center) (Furr & Bacharach, 2008) oder systemischer Ebene (z. B. Vergabe von Fördergeldern für Schulprojekte oder Studiengänge), können nicht-valide Interpretationen von Messergebnissen erhebliche Konsequenzen nach sich ziehen. Deshalb ist es von besonderer Bedeutung, dass im Rahmen der Entwicklung und Evaluation von Messverfahren neben der Gewährleistung der Objektivität und Reliabilität vor allem die Überprüfung der Validität in den Blick genommen wird. Nachdem bisherige Ergebnisse dafür sprechen, dass das Tool Observer professionelle Unterrichtswahrnehmung objektiv und reliabel erfasst, liegt der Fokus der vorliegenden Arbeit auf der Überprüfung der Anforderungen an das Instrument bezüglich Validität. Dazu wird im Folgenden das Konzept der Validität theoretisch aufgearbeitet. Basierend darauf werden die Anforderungen an das Instrument hinsichtlich Validität ausdifferenziert, bisherige Überprüfungen dieser Anforderungen im Rahmen des DFG-Projekts Observe dargestellt sowie noch ausstehende Überprüfungen, die in der vorliegenden Arbeit adressiert werden, abgeleitet.

2.3.2.1 Theoretische Überlegungen zu Validität

Im Folgenden werden theoretische Überlegungen zum Konzept der Validität aufgearbeitet. Dabei wird eine Definition gegeben und das der Arbeit zugrunde liegende übergreifende Konzept von Validität erläutert. Abschließend werden der Prozess der Validierung und ein generelles Modell bildungswissenschaftlichen Testens dargestellt, das in der vorliegenden Arbeit auf das Tool Observer übertragen wird und als Grundlage für die Beschreibung des Validierungsprozesses des Instruments fungiert.

2.3.2.1.1 Definition von Validität

Das Konzept der Validität war im letzten Jahrhundert einer stetigen Veränderung unterzogen (Zumbo, 2007). Zunächst dominierte ein kriteriumbasiertes Modell, in dem der Zusammenhang des Messergebnisses mit externalen Kriterien im Vordergrund steht, mit vereinzelten Schwerpunkten auf einer inhaltsbasierten Sichtweise (Zumbo, 2007). Anschließend verschob

sich der Fokus auf ein konstruktbasiertes Modell von Validität (Zumbo, 2007), vor allem beeinflusst durch die Arbeit von Cronbach und Meehl (1955). Dieses Konzept hat sich bis heute etabliert und wurde allen voran von Messick (z. B. 1975, 1980) um eine ethisch begründete Komponente, die Konsequenz des Einsatzes und der Interpretation von Messinstrumenten, erweitert (Zumbo, 2007). Ein ausführlicher Überblick über die Entwicklung des Konzepts der Validität ist beispielsweise bei Kane (2001) zu finden. Diese Entwicklung spiegelt sich aber auch in den gemeinsamen Standards von AERA, APA und NCME wider (Tittle, 2006).

Im Gegensatz zu älteren Versionen dieser Standards wird Validität in den letzten beiden Versionen nicht mehr als Eigenschaft des Messergebnisses betrachtet, sondern bezieht sich vielmehr auf dessen Interpretation und Nutzung (Frey, 2014, März). Das bedeutet, dass Validität als das Ausmaß verstanden wird, indem Evidenz und Theorie die Interpretation eines Messergebnisses im Rahmen der empfohlenen Einsatzmöglichkeiten des Instruments unterstützen (AERA et al., 1999). Als zusätzliche Neuerung wird die Differenzierung in verschiedene Arten von Validität durch ein übergreifendes Konzept von Validität abgelöst (Frey, 2014, März). Dieses Verständnis wird im deutschsprachigen Raum jedoch noch kaum rezipiert (Frey, 2014, März). Der Prozess der Validierung bezieht sich dabei auf das Akkumulieren von Evidenz, um eine solide wissenschaftliche Grundlage für die Interpretation des Messergebnisses als Folge der vorgeschlagenen Nutzung des Messinstruments bereitzustellen (AERA et al., 1999). Diese Definition von Validität, die der vorliegenden Arbeit zugrunde gelegt wird, bringt drei wesentliche Implikationen für Validitätsaussagen von Messinstrumenten mit sich. Erstens kann ein Messinstrument selbst niemals als valide oder nicht valide bezeichnet werden, vielmehr wird beurteilt, wie valide die Interpretation und die Nutzung der Messergebnisse ist (Furr & Bacharach, 2008). Zweitens ist eine Validitätsaussage immer im Rahmen des geplanten Anwendungskontexts zu treffen und kann nicht losgelöst davon beurteilt werden (AERA et al., 1999). Drittens ist eine Validitätsaussage niemals dichotom (valide/nicht valide), sondern bewegt sich auf einem Kontinuum, auf dem der Grad der Validität beurteilt wird (Zumbo, 2007).

2.3.2.1.2 Übergreifendes Konzept von Validität

Entgegen der vor allem im deutschsprachigen Raum immer noch etablierten Dreiteilung von Validität in Inhalts-, Kriteriums- und Konstruktvalidität (z. B. Bühner, 2011) stellt das übergreifende Konzept von Validität die Konstruktvalidität in den Mittelpunkt (AERA et al., 1999; Messick, 1995). Allerdings wird unter Konstruktvalidität Evidenz basierend auf Überlegungen zu Inhalts- und Kriteriumsvalidität subsumiert (Cronbach, 1980; Messick, 1975, 1980). Der Begriff Konstruktvalidität deutet darauf hin, dass Messergebnisse als Indikatoren für dahinterstehende theoretische Konstrukte interpretiert werden, auf die durch unterschiedliche Arten von Evidenz geschlossen werden kann (AERA et al., 1999). Dadurch wird die zentrale Bedeutung einer zugrunde liegenden vertretbaren psychologischen Theorie für die Interpretation und die Nutzung von Messergebnissen betont (Furr & Bacharach, 2008). Der Begriff Konstruktvalidität kann umfassend für Validität verwendet werden, da in der gegenwärtigen Bildungsforschung alle Messergebnisse als Indikatoren für theoretische Konstrukte angesehen werden (AERA et al., 1999).

Um beurteilen zu können, inwieweit die Interpretation der Messergebnisse valide ist, können jedoch unterschiedliche Arten von Evidenz herangezogen werden. Im Vergleich zur klassischen Dreiteilung von Validität fand eine deutliche Erweiterung der Evidenzgrundlage statt (Tittle, 2006). Entsprechend den Standards (AERA et al., 1999) werden fünf Grundlagen für Evidenz unterschieden: *Inhalt, Antwortprozess, interne Struktur, Beziehungen mit anderen Variablen* und *Konsequenzen der Messung*. Dahingegen schlägt Messick (1995) eine Differenzierung in sechs Aspekte vor: *Inhalt, Substanz, Struktur, Generalisierbarkeit, Externalität* und *Konsequenzen*. Beide Differenzierungsversuche stimmen darin überein, dass Validität als ein übergreifendes Konzept verstanden wird (AERA et al., 1999; Messick, 1995). Die verschiedenen Evidenzgrundlagen betreffen verschiedene Aspekte von Validität, stellen aber keine distinkten Arten von Validität dar, die getrennt voneinander beurteilt werden (AERA et al., 1999). Folglich werden in der vorliegenden Arbeit die Begriffe Evidenzgrundlage und Aspekt synonym verwendet.

Für die vorliegende Arbeit wird ein integriertes Modell von Validität erarbeitet, dass die Differenzierungen in verschiedene Aspekte von Validität von Messick (1995) und der Standards (AERA et al., 1999) vereint. Im Folgenden wird dargestellt, wie diese Differenzierungen ineinander überführbar sind. Dabei wird entlang der sechs Aspekte von Messick (1995) vorgegangen und die fünf verschiedenen Grundlagen von Evidenz der Standards (AERA et al., 1999) werden innerhalb dieser verortet. Zusätzlich wird darauf eingegangen, in welchen As-

pekten die klassischen Arten von Validität, Inhalts- und Kriteriumsvalidität wiederzufinden sind. Durch die zentrale Stellung von Konstruktvalidität, die im übergreifenden Validitätsverständnis mit Validität gleichgesetzt wird (vgl. AERA et al., 1999), liefern alle Aspekte Evidenz zur Beurteilung der Konstruktvalidität. Dieses dargelegte integrierte Modell von Validität, das auf einem übergreifenden Verständnis von Validität mit dem Fokus auf Konstruktvalidität basiert, ist in Abbildung 3 dargestellt. Es dient in der vorliegenden Arbeit als Grundlage für die Analyse, welche Aspekte von Validität im DFG-Projekt Observe bereits adressiert wurden und welche Überprüfungen noch ausstehen.

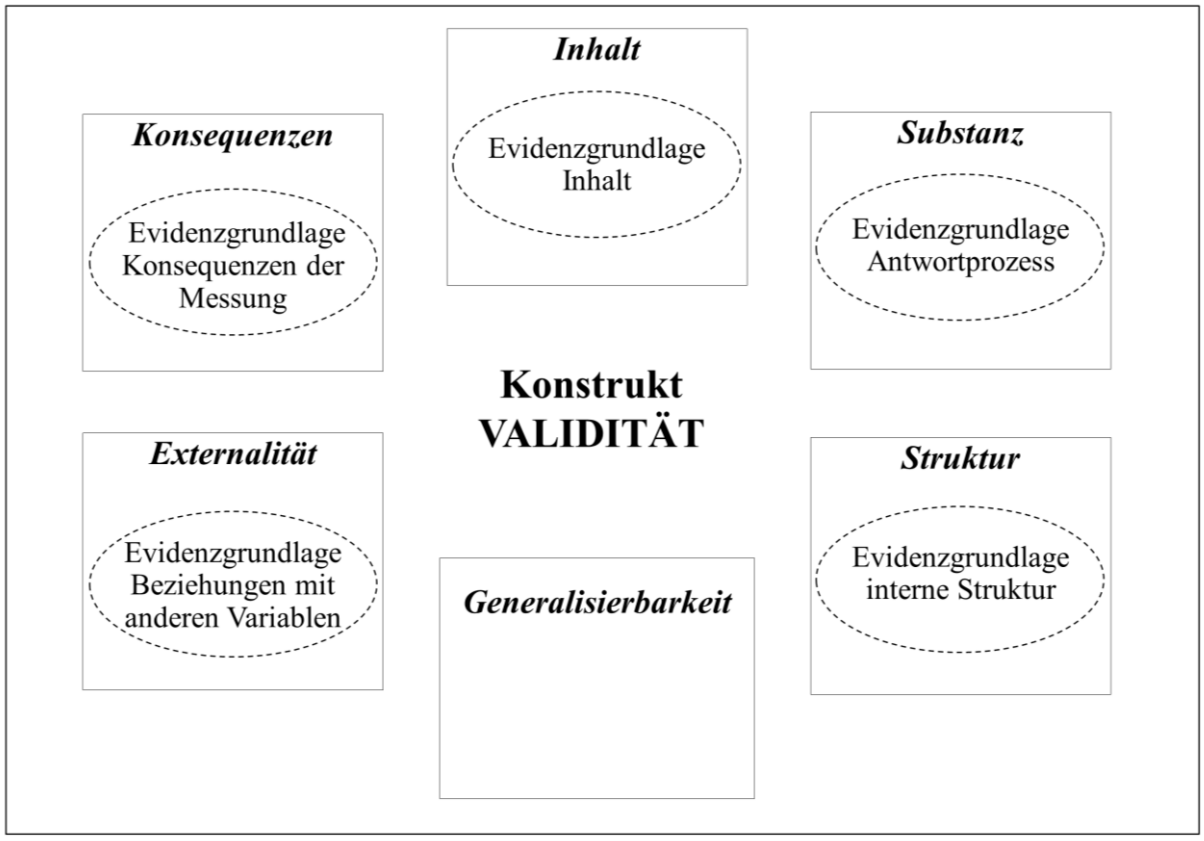

Abbildung 3. Integriertes Modell von Validität (adaptiert nach AERA et al., 1999; Messick, 1995).

Drei Validitätsaspekte bei Messick (1995) und drei Evidenzgrundlagen der Standards (AERA et al., 1999) werden mit vergleichbaren Namen bezeichnet und stimmen inhaltlich überein. Diese betreffen *Inhalt*, *Struktur* und *Konsequenzen*.

Der Validitätsaspekt *Inhalt* kann mit der Evidenzgrundlage *Testinhalt* gleichgesetzt werden. Der Inhalt des Messinstruments sollte relevant und repräsentativ für das zu messende Konstrukt sein (Messick, 1995). Diese Aussage entspricht der Formulierung in den Standards, dass die Passung zwischen Inhalt eines Messinstruments und Inhalt des zu messenden theoretischen Konstrukts fokussiert wird (AERA et al., 1999). Dies ist nicht der Fall, werden konstrukt-irrelevante Inhalte (Überrepräsentation) erfasst oder Teile des Konstrukts nicht mit dem Messinstrument erfasst (Unterrepräsentation) (Furr & Bacharach, 2008). Der Validitätsaspekt *Inhalt* stellt damit Evidenz dar, die in der klassischen Dreiteilung von Validität herangezogen wird, um die Inhaltsvalidität zu beurteilen (AERA et al., 1999). Die Problematik der Überprüfung dieses Validitätsaspekts wird immer wieder diskutiert (z. B. Bühner, 2011). Eine Möglichkeit, Evidenz zu generieren, stellen Studien mit Experten dar, da diese für einen spezifischen Forschungsbereich in der Lage sind, die Qualität der Umsetzung eines theoretischen Konstrukts mittels eines Messinstruments zu beurteilen (Furr & Bacharach, 2008).

Der Validitätsaspekt S*truktur* entspricht der Evidenzgrundlage *interne Struktur*. In beiden wird gefordert, dass die Struktur des Messinstruments mit der Struktur des zu messenden Konstrukts übereinstimmt (AERA et al., 1999; Messick, 1995). Für die Überprüfung dieses Validitätsaspekts sind sowohl ein präzises theoretisches Modell des Konstrukts als auch adäquate Methoden zur Untersuchung der Struktur des Messinstruments notwendig. Letztere hängen von der Komplexität der Struktur ab (AERA et al., 1999). Beispielsweise werden zur Überprüfung der Dimensionalität häufig Faktoranalysen durchgeführt (Furr & Bacharach, 2008).

Der Validitätsaspekt *Konsequenzen* entspricht der Evidenzgrundlage *Konsequenzen der Messung*. Beide zielen darauf ab, intendierte und nicht intendierte Implikationen der Interpretation eines Messergebnisses sowie potentielle Folgen der Benutzung des Messinstruments zu berücksichtigen (AERA et al., 1999; Messick, 1995). Derartige Konsequenzen können direkt relevant für die Validität sein, wenn die Ursache (z. B. Unterrepräsentation des theoretischen Konstrukts) auf mangelnde Validität in einem anderen Aspekt (z. B. struktureller Aspekt) zurückgeführt werden kann (AERA et al., 1999). Ist dies nicht der Fall, könnten derartige Konsequenzen zwar immer noch für Grundsatzentscheidungen hinsichtlich der Nutzung des Messinstruments von Bedeutung sein, fallen aber nicht mehr in den eigentlichen Geltungsbereich der Validität (AERA et al., 1999). Demnach gilt es, den Einsatz eines Messinstruments zu überwachen, Konsequenzen zu erfassen und nach deren potentiellen Ursachen zu suchen.

Außerdem tragen zwei Validitätsaspekte bei Messick (1995), nämlich die der *Substanz* und der *Externalität*, und zwei Evidenzgrundlagen der Standards (AERA et al., 1999) – *Antwortprozess* und *Beziehungen zu anderen Variablen* – unterschiedliche Namen, sind aber inhaltlich vergleichbar.

Der Validitätsaspekt *Substanz* ist mit der Evidenzgrundlage *Antwortprozess* vergleichbar. Zentrales Merkmal beider ist die Passung des Antwortverhaltens, das durch das Messinstrument ausgelöst wird, mit dem zu erfassenden Konstrukt (AERA et al., 1999; Messick, 1995). Messick (1995) betont, dass damit die Forderung im Kontext des Validitätsaspekts *Inhalt*, dass die Inhalte des Messinstruments repräsentativ sein müssen, auf den Bereich der Prozesse, die durch die Bearbeitung des Messinstruments ausgelöst werden, ausgeweitet wird, d. h. es müssen alle relevanten Prozesse erfasst werden. Dies kann beispielsweise in Studien überprüft werden, in denen während der Bearbeitung des Messinstruments Laut-Denken-Protokolle eingesetzt oder Augenbewegungen aufgezeichnet werden (Messick, 1995).

Der Validitätsaspekt *Externalität* ist mit der Evidenzgrundlage *Beziehungen zu anderen Variablen* vergleichbar. Innerhalb des Validitätsaspekts *Externalität* wird überprüft, in welchem Ausmaß theoretisch abgeleitete Zusammenhänge zwischen dem Messergebnis und relevanten Variablen erwartungsgemäß bestehen (Messick, 1995). Dies entspricht dem Fokus der Standards auf die Analyse der Beziehung des Messergebnisses mit anderen externalen Variablen (AERA et al., 1999). Dabei werden konvergente und divergente Zusammenhänge unterschieden (Furr & Bacharach, 2008). Erstere beruhen darauf, in welchem Maß das Messergebnis mit Konstrukten, die mit dem zu messenden Konstrukt verwandt sind, zusammenhängt. Letztere basieren darauf, in welchem Maß das Messergebnis mit Konstrukten, für die kein Zusammenhang mit dem zu messenden Konstrukt erwartet wird, nicht zusammenhängt (Furr & Bacharach, 2008). Beide Zusammenhänge können korrelativ oder experimentell untersucht werden (AERA et al., 1999). Von besonderer Bedeutung sind Zusammenhänge zwischen dem Messergebnis und einem relevanten externalen Kriterium (z. B. Studienerfolg) (Messick, 1995), die als Spezialfälle unter konvergente Zusammenhänge gefasst werden (Furr & Bacharach, 2008). Abhängig von der Übereinstimmung des Messzeitpunkts werden simultane und prädiktive Designs zur Überprüfung derartiger Test-Kriteriums-Beziehungen unterschieden (AERA et al., 1999). Von grundlegender Bedeutung bei derartigen Studien ist sowohl die Auswahl als auch die Erfassung des Kriteriums (AERA et al., 1999). Dieser Validitätsaspekt bezieht sich somit auf Evidenz, die in der klassischen Dreiteilung von Validität relevant für die Beurteilung der Kriteriumsvalidität ist (AERA et al., 1999).

Darüber hinaus unterscheidet Messick (1995) noch den Validitätsaspekt *Generalisierbarkeit*. Dabei wird untersucht, inwieweit die Eigenschaften und die Interpretation der Messergebnisse auf andere Gruppen, Settings oder Aufgaben übertragbar sind, mit dem Ziel, Grenzen der Generalisierbarkeit abzustecken (Messick, 1995). Hinsichtlich der Eigenschaften der Messergebnisse wie z. B. Auswahl von Aufgaben und Ratern überschneidet sich dieser Aspekt mit Untersuchungen zur Überprüfung der Reliabilität (Messick, 1995). Allerdings fokussieren Reliabilitätsanalysen auf die Quantifizierung von Messfehlern, die grundsätzlich als zufällig angesehen werden (AERA et al., 1999). Dahingegen werden systematische Fehler, die beispielsweise auf verschiedene Itemschwierigkeiten für einzelne Gruppen zurückzuführen sind, in klassischen Ansätzen zur Überprüfung der Reliabilität nicht analysiert und fallen in den Bereich der Validitätsprüfung. Eine Ausnahme stellt die Generalisierbarkeitstheorie dar (AERA et al., 1999), da sie für die Evaluation von Messergebnissen das Konzept der klassischen Reliabilitätstheorie um einzelne Aspekte der Validität erweitert (Shavelson & Ruiz-Primo, 2005). Der Fokus der vorliegenden Arbeit liegt jedoch auf der Überprüfung der Validität, weshalb die Generalisierbarkeitstheorie als erweiterter Ansatz zur Überprüfung der Reliabilität nicht weiter berücksichtigt wird.

Der Validitätsaspekt *Generalisierbarkeit* wird im Rahmen der Standards (AERA et al., 1999) zwar nicht explizit als eigenständige Evidenzgrundlage erwähnt, ist aber implizit darin enthalten. Bezieht sich die *Generalisierbarkeit* auf Eigenschaften der einzelnen Items eines Messinstruments, fällt dieser Validitätsaspekt in die Evidenzgrundlage *interner Struktur*, in der beispielsweise explizit erwähnt wird, dass die Vergleichbarkeit der Itemschwierigkeiten über Differential Item Functioning (DIF) überprüft werden soll (AERA et al., 1999). Betrifft die *Generalisierbarkeit* psychologische Prozesse während der Bearbeitung des Messinstruments, ist dieser Validitätsaspekt durch die Evidenzgrundlage *Antwortprozesse* abgedeckt. Darin wird explizit darauf hingewiesen, dass Unterschiede in Antwortprozessen nicht auf den Probanden limitiert sind (AERA et al., 1999). Gerade vor dem Hintergrund eines großflächigen Einsatzes des Messinstruments im Large-Scale-Kontext sollte die Überprüfung der *Generalisierbarkeit* der Interpretation von Messergebnissen unter keinen Umständen ausbleiben. Deshalb wird dieser Aspekt im integrierten Modell von Validität (vgl. Abbildung 3), explizit aufgeführt und folglich in der vorliegenden Arbeit die Nomenklatur von Messicks Differenzierung (1995) verwendet.

2.3.2.1.3 Prozess der Validierung

Im Hinblick auf den Prozess der Validierung wirkt das beschriebene theoretische Konzept von Validität (AERA et al., 1999; Messick, 1995) jedoch abstrakt und umfangreich (Kane, 2013a). Deshalb wird im Folgenden zunächst der argumentbasierte Ansatz (Kane, 1992, 2001, 2013a) als Strategie für eine systematische Validierung beschrieben und im Anschluss ein Prozessmodell dargestellt, das in der vorliegenden Arbeit dazu genutzt wird, den Validierungsprozesses des Tools Observer zu beschreiben.

Argumentbasierter Ansatz

Der argumentbasierte Ansatz wurde vornehmlich durch Kane (1992, 2001, 2013a) entwickelt und stellt eine Strategie für die systematische Validierung eines Messinstruments dar. Ziel ist es, den Prozess der Validierung zu vereinfachen, zu begrenzen und damit praktikabler zu machen (Kane, 2013a). Er basiert auf dem grundlegenden wissenschaftlichen Prinzip, dass Behauptungen durch geeignete Argumente gestützt werden sollten (Kane, 2013a). Entsprechend werden bei der Validierung zuerst Behauptungen hinsichtlich Schlussfolgerungen und Vermutungen, auf denen die vorgeschlagene Interpretation und Nutzung des Messergebnisses beruhen, formuliert (Interpretations-/Nutzungs-Argument) und diese anschließend bezüglich Kohärenz, Vollständigkeit, Plausibilität und Evidenzbasierung evaluiert (Validitätsargument) (Kane, 2013a). Dabei ist zu berücksichtigen, dass diese beiden Prozesse in der Testpraxis oft verknüpft sind und nicht sequentiell verlaufen (Kane, 2013a). Ein Messinstrument wird meist für einen bestimmten Zweck entwickelt, der wiederum zusammen mit der theoretischen Fundierung des zu messenden Konstrukts die Entwicklung des Messinstruments und die Spezifikation des Interpretations-/Nutzungs-Arguments steuert (Kane, 2013a). Das Vorgehen, das im argumentbasierten Ansatz (Kane, 2013a) postuliert wird, wird in der vorliegenden Arbeit im Hinblick auf die Validierung des Tools Observer umgesetzt. Es werden zunächst Anforderungen an das Instrument ausgehend von der intendierten Interpretation und Nutzung der Messergebnisse spezifiziert (vgl. Abschnitt 2.3.2.2).

In Bezug auf die Evaluation des Interpretations-/Nutzungs-Arguments kann die theoretische Differenzierung in verschiedene Grundlagen von Evidenz eine Hilfe sein, unterschiedliche Aspekte im Validierungsprozess zu berücksichtigen (Messick, 1995). Entsprechend dem übergreifenden Validitätsverständnis ist jedoch keine Evidenz grundsätzlich höherwertig einzuschätzen oder zu bevorzugen (AERA et al., 1999). An dieser Stelle bietet der argumentba-

sierte Ansatz der Validierung den Vorteil, dass durch die konkrete Spezifikation des Interpretations-/Nutzungs-Arguments vorgegeben ist, welche Evidenz für dessen Evaluation notwendig ist (Kane, 2013a). Es muss lediglich Evidenz für die konkreten Schlussfolgerungen und Vermutungen generiert werden, auf denen die zuvor spezifizierte Interpretation und Nutzung des Messergebnisses basieren. Durch dieses Vorgehen wird zum einen der Prozess der Validierung praktikabler, da ein Start- und ein Endpunkt vorgegeben sind (Kane, 2013a). Zum anderen wird damit aber auch erneut die starke Verknüpfung von Validität und Kontext deutlich, indem die Relevanz und Qualität einer Evidenzgrundlage zur Beurteilung der Validität durch die spezifizierte Interpretation und den beabsichtigten Einsatzkontext bestimmt wird (AERA et al., 1999; Kane, 2013a).

Ziel eines jeden Validierungsprozesses muss es sein, verschiedene Evidenzgrundlagen in eine übergreifende Validitätsaussage zu synthetisieren (Messick, 1998; Tittle, 2006). Entscheidend ist, dass die angeführten Evidenzgrundlagen die zuvor spezifizierte Interpretation des Messergebnisses verbunden mit einer bestimmten Nutzung rechtfertigen oder Gründe für die Beschränkung der Interpretation oder Nutzung anführen (Kane, 2013a; Messick, 1995). Übertragen auf die vorliegende Arbeit bedeutet dies, dass im Rahmen der Validierung des Tools Observer abhängig von den spezifizierten Anforderungen an das Instrument verschiedene Validitätsaspekte überprüft werden müssen. Auf dieser Grundlage kann eine umfassende Validitätsaussage getroffen und Möglichkeiten und Grenzen eines großflächigen Einsatzes aufgezeigt werden.

Modell zur Beschreibung eines Validierungsprozesses

In der vorliegenden Arbeit wird ein generelles Modell bildungswissenschaftlichen Testens (Hattie, Jaeger & Bond, 1999; adaptiert nach Zumbo, 2007) genutzt, um den Validierungsprozess des Tools Observer zu beschreiben (vgl. Abbildung 4). Dazu werden die einzelnen Komponenten dieses Prozessmodells auf die Entwicklung des Instruments im Rahmen des DFG-Projekts Observe übertragen. Das Modell bietet im Hinblick auf die Beschreibung eines komplexen Validierungsprozesses mehrere Vorteile. Zum einen findet sich der zyklische Charakter des Validierungsprozesses, der im argumentbasierten Ansatz der Validierung (Kane, 2013a) betont wird, graphisch aufgegriffen. Zum anderen liefern die sechs Komponenten des Models Anhaltspunkte dafür, welche Schlussfolgerungen im Sinne des argumentbasierten Ansatzes der Validierung (Kane, 2013a) im Interpretations-/Nutzungs-Argument enthalten

sind, und damit gleichzeitig auch dafür, welche Arten von Evidenz für das Validitätsargument notwendig sind. Auf diese Art und Weise wird die notwendige Struktur für die Beschreibung des Validierungsprozesses bereitgestellt. Diese geht über die Orientierungshilfe innerhalb des argumentbasierten Ansatzes der Validierung (Kane, 1992) im Sinne von vier typischen Schlussfolgerungen (z. B. das Scoring-Verfahren betreffend) hinaus und stellt somit eine wichtige Ergänzung für die Validierungspraxis dar.

Im generellen Modell bildungswissenschaftlichen Testens (Hattie et al., 1999; adaptiert nach Zumbo, 2007) wird bildungswissenschaftliches Testen als zyklischer Prozess dargestellt, um zu illustrieren, dass dieser Prozess nicht mit dem Messergebnis endet (Hattie et al., 1999). Dabei unterscheiden Hattie und Kollegen (1999) fünf aufeinander folgende Komponenten, die zentrale methodische Fragen im Kontext der Entwicklung eines Messinstruments darstellen: *konzeptuelle Messmodelle, Test- und Item-Entwicklung, Test-Administration, Test-Nutzung* und *Test-Evaluation.* Zumbo (2007) ergänzt diesen Zyklus um die Komponente der Entwicklung von *konzeptuellen theoretischen Modellen* der zu erfassenden Konstrukte. Diese zusätzliche Komponente entspricht der Forderung nach einer präzisen theoretischen Modellierung professioneller Kompetenzen als Voraussetzung für deren valide Erfassung (Klieme & Leutner, 2006a) und betont damit die zentrale Stellung der theoretischen Modellierung der zu erfassenden Konstrukte im Rahmen der Testentwicklung. Entscheidend ist, dass der Prozess des bildungswissenschaftlichen Testens nicht mit dem Messergebnis abgeschlossen ist, sondern zyklischen Charakter aufweist (Hattie et al., 1999). Die Ergebnisse der *Test-Evaluation* beeinflussen wiederum das *konzeptuelle theoretische Modell* (Zumbo, 2007) und der Kreislauf beginnt erneut.

Die erste Komponente *konzeptuelle theoretische Modelle* beinhaltet die Entwicklung und Darlegung des theoretischen Modells des zu erfassenden Konstrukts (Zumbo, 2007). Anschließend erfolgt die Spezifikation von *konzeptuellen Messmodellen* (Hattie et al., 1999). Für die Modellierung der Beziehung zwischen der theoretischen Struktur und der numerischen Repräsentation stehen unterschiedliche Messmodelle zur Auswahl wie z. B. die klassische Testtheorie oder die Item-Response-Theorie (Hattie et al., 1999). Aufbauend darauf findet die *Test- und Item-Entwicklung* statt, die wichtige Entscheidungen wie die Wahl der Item-Formate, des Scoring-Verfahrens oder der Regel, nach der Antworten numerische Werte zugewiesen werden, enthält (Hattie et al., 1999). Eine weitere Komponente des Modells stellt die Frage nach der *Test-Administration* dar, auf der Suche nach der Balance zwischen Standardisierung einerseits sowie Repräsentativität und Authentizität der Erhebungsbedingungen andererseits

(Hattie et al., 1999). Die Diskussion reicht dabei von Vor- und Nachteilen verschiedener Erhebungsbedingungen, z. B. Paper-und-Pencil versus computerbasierte Erhebungen bis zur Auswahl von Items für adaptive Testverfahren (Hattie et al., 1999). Die nächste Komponente *Test-Nutzung* trägt der Tatsache Rechnung, dass Messergebnisse stets für bestimmte Zwecke genutzt werden, z. B. für Diagnostik oder als Grundlage für bestimmte politische Entscheidungen (Hattie et al., 1999). Hier besteht teilweise die Gefahr einer Simplifizierung der Realität durch das Stützen auf Messergebnisse; darüber hinaus gewinnt die Darstellung des Messergebnisses an Bedeutung (Hattie et al., 1999). Die sechste und letzte Komponente *Test-Evaluation* umfasst im ursprünglichen Modell (Hattie et al., 1999) eine Fülle an Überprüfungen, unter anderem der Reliabilität, Validität und Dimensionalität. Im Gegensatz dazu wird in der vorliegenden Arbeit der Prozess der Validierung auf alle Komponenten des Modells ausgeweitet, da in jeder Komponente Entscheidungen getroffen werden, die sich darauf auswirken, wie valide die Interpretation der Messergebnisse ist. Diese Ausweitung entspricht der Empfehlung von Zumbo (2007) und findet weitere Unterstützung im argumentbasierten Ansatz, in dem z. B. explizit auf die Bedeutung der Wahl des Scoring-Verfahrens hingewiesen wird, das in der Komponente *Test- und Item-Entwicklung* berücksichtigt ist (Kane, 1992).

2.3.2.2 Ausdifferenzierung der Anforderungen hinsichtlich Validität

Bevor der bisherige Validierungsprozess des Tools Observer mit Hilfe des generellen Modells bildungswissenschaftlichen Testens (Hattie et al., 1999; adaptiert nach Zumbo, 2007) beschrieben wird, werden im Folgenden gemäß dem Vorgehen im argumentbasierten Ansatz von Validierung (vgl. Kane, 2013a) die Anforderungen an das Instrument bezüglich Interpretation und Nutzung der Messergebnisse ausdifferenziert.

Die Anforderungen an ein Messinstrument ergeben sich aus dem spezifischen Kontext, in dem dieses eingesetzt wird. Professionelle Unterrichtswahrnehmung, die mit dem Tool Observer erfasst wird, stellt ein Beispiel für eine professionelle Kompetenz Lehramtsstudierender dar. Professionelle Kompetenzen können als Ergebnisse von Bildungsprozessen im Rahmen des Studiums verstanden werden (vgl. Hartig, 2008). Folglich zielt die Erfassung derartiger Kompetenzen nicht nur darauf ab, Rückmeldungen auf individueller Ebene zu geben (z. B. über Lernerfolg und Entwicklung eines Studierenden), sondern auch auf systemischer Ebene (z. B. Überprüfung der Wirksamkeit von Studiengängen). Dafür ist ein übergreifender Einsatz des Instruments im Large-Scale-Kontext notwendig. Vor diesem Hintergrund lassen

sich spezifische Anforderungen bezüglich Interpretation und Nutzung der Messergebnisse an das Tool Observer zur Erfassung professioneller Unterrichtswahrnehmung ableiten.

Hinsichtlich der Interpretation der Messergebnisse des Instruments ergeben sich folgende drei Anforderungen, deren Überprüfung auf die Validitätsaspekte *Inhalt*, *Substanz* und *Struktur* (vgl. Abschnitt 2.3.2.1.2) abzielt. (1) 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. Die Überprüfung dieser Anforderungen fällt in den Validitätsaspekt *Inhalt*. (2) Zudem sollte sichergestellt werden, dass durch die Kombination aus Videoclips und Rating-Items der Prozess der professionellen Unterrichtswahrnehmung ausgelöst und erfasst wird. Diese Anforderung entspricht dem Validitätsaspekt *Substanz*. (3) Zusätzlich ist es wichtig, dass die theoretisch angenommene Struktur professioneller Unterrichtswahrnehmung durch das Tool empirisch abgebildet wird. Im Hinblick auf einen großflächigen Einsatz des Instruments ist es zentral, dass sich die theoretische Kompetenzstruktur nicht nur in einer lokalen Stichprobe empirisch abbilden lässt, sondern auch standortübergreifend. Die Überprüfung dieser Anforderungen adressiert den Validitätsaspekt *Struktur*.

Darüber hinaus leiten sich weitere zwei Anforderungen aus potenziellen Nutzungen der Messergebnisse ab, deren Überprüfung die Validitätsaspekte *Generalisierbarkeit* und *Externalität* (vgl. Abschnitt 2.3.2.1.2) betrifft. (4a) Voraussetzung für den großflächigen Einsatz des Tools Observer in der universitären Lehrerbildung als Instrument zur Erfassung professioneller Unterrichtswahrnehmung ist die Möglichkeit einer ökonomischen Kompetenzmessung im Large-Scale-Kontext. Deshalb wurde das Instrument in eine Online-Plattform integriert, um dadurch eine internetbasierte Bearbeitung zu ermöglichen. Die Bearbeitung des Tools wird teilweise aber auch in Lehrveranstaltungen integriert und als verpflichtende Gruppentestung durchgeführt, sei es aufgrund von Rekrutierungsproblemen oder um die Lehrveranstaltung zu evaluieren. Folglich ist es entscheidend, dass das Instrument über verschiedene Bearbeitungskontexte und Arten der Teilnahme hinweg professionelle Unterrichtswahrnehmung stabil erfasst. Diese Anforderung bezieht sich auf den Validitätsaspekt *Generalisierbarkeit*. (4b) Im Rahmen eines übergreifenden Einsatzes im Large-Scale-Kontext soll das Tool Observer von Studierenden unterschiedlicher Lehramtsstudiengänge bearbeitet und dazu genutzt werden, Unterschiede in der professionellen Unterrichtswahrnehmung zwischen diesen Studierenden abzubilden. Dazu muss zum einen garantiert sein, dass das Tool Observer für Studierende verschiedener Lehr-

amtsstudiengänge die Struktur professioneller Unterrichtswahrnehmung vergleichbar empirisch abgebildet, und zum anderen, dass die Itemschwierigkeiten für die verschiedenen Studierendengruppen vergleichbar sind. Ansonsten können gemessene Unterschiede in der professionellen Unterrichtswahrnehmung nicht zwingend auf Unterschiede in den Personenfähigkeiten zurückzuführen sein, sondern auch durch Eigenschaften des Instruments hervorgerufen werden. Auch diese Anforderung wird unter den Validitätsaspekt *Generalisierbarkeit* gefasst.

(5) Voraussetzung dafür, dass das Tool als Diagnoseinstrument zur Erfassung der Entwicklung professioneller Unterrichtswahrnehmung im Verlauf der universitären Lehrerbildung eingesetzt werden kann, ist ein positiver Zusammenhang zwischen dem in einer Lehrveranstaltung vermittelten Wissen über effektives Lehren und Lernen und der erfassten professionellen Unterrichtswahrnehmung. Eine weitere Möglichkeit eines großflächigen Einsatzes des Instruments ist die Nutzung zur Selbstexploration im Rahmen der Studienwahl. Dafür ist es erforderlich, dass ein Zusammenhang mit dem Studienerfolg besteht. 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. Die Überprüfung dieser Anforderungen zielt auf den Validitätsaspekt *Externalität* ab.

2.3.2.3 Bisherige Überprüfungen im Rahmen des DFG-Projekts Observe

Im Folgenden wird dargestellt, inwieweit die Anforderungen an das Tool Observer hinsichtlich der Validität der Interpretation und Nutzung der Messergebnisse (vgl. Abschnitt 2.3.2.2) im Rahmen des DFG-Projekts Observer bisher überprüft wurden. Als Grundlage für die Beschreibung des bisherigen Validierungsprozesses wird das unter Abschnitt 2.3.2.1.3 vorgestellte generelle Modell bildungswissenschaftlichen Testens (Hattie et al., 1999; adaptiert nach Zumbo, 2007) auf die Entwicklung des Instruments übertragen (vgl. Abbildung 4). Unter der entsprechenden Komponente des Prozessmodells werden Studien eingeordnet, die im Projektverlauf durchgeführt wurden und bereits einzelne Aspekte des integrierten Modells von Validität (vgl. Abbildung 3) adressieren.

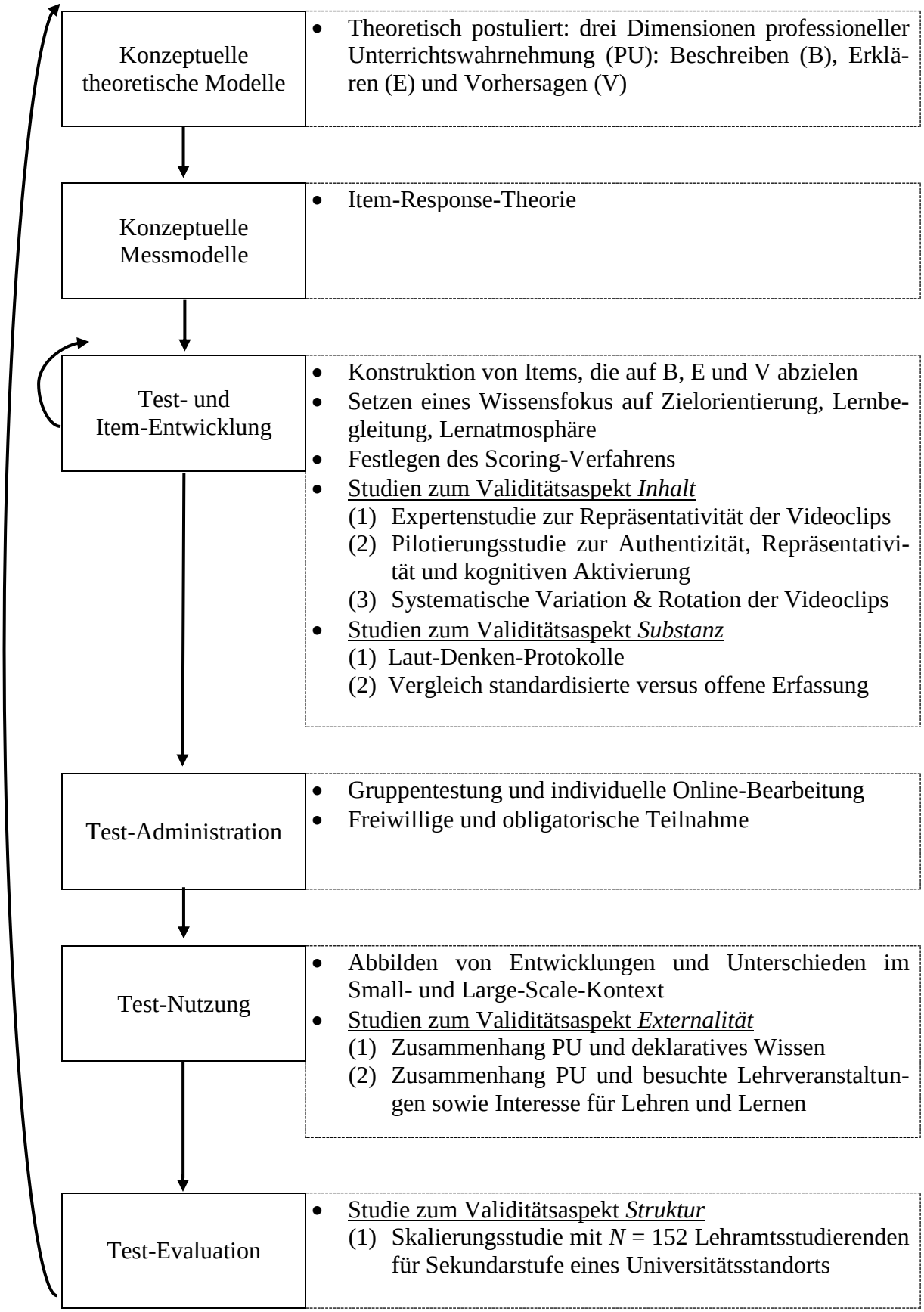

Abbildung 4. Generelles Modell bildungswissenschaftlichen Testens (Hattie et al., 1999; adaptiert nach Zumbo, 2007) angewandt auf bisherige Studien zum Tool Observer.

Im Rahmen des DFG-Projekts Observe wurde das Tool Observer entwickelt, um professionelle Unterrichtswahrnehmung Lehramtsstudierender kontextualisiert, aber gleichzeitig standardisiert zu erfassen. Dies wurde mittels kurzer videographierter Unterrichtsituationen, die in standardisierte Rating-Items eingebettet und in eine Online-Plattform integriert sind, realisiert. Hinsichtlich der Komponente *konzeptuelle theoretische Modelle* wurde im Projekt theoriegeleitet und aufbauend auf qualitativen Studien ein dreidimensionales Modell professioneller Unterrichtswahrnehmung mit den drei qualitativ unterschiedlichen Aspekten Beschreiben, Erklären und Vorhersagen postuliert (vgl. Abschnitt 2.2.2).

Bezüglich der Komponente *konzeptuelle Messmodelle* wurde als Messmodell die Item-Response-Theorie gewählt, da angenommen wird, dass die Einschätzungen der Videoclips des Tools Observer einen Indikator für die latente Variable professionelle Unterrichtswahrnehmung darstellen.

Im Rahmen der *Test- und Item-Entwicklung* wurde darauf geachtet, die gleiche Anzahl an Items zu konstruieren, die auf die Fähigkeiten abzielen, die Unterrichtssituationen zu beschreiben, zu erklären und vorherzusagen. Aufgrund der Komplexität von Lehr- und Lernprozessen im Unterricht, die die Kapazität standardisierter Messinstrumente bei Weitem übersteigt, wurde ein Wissensfokus gesetzt, vor dessen Hintergrund professionelle Unterrichtswahrnehmung erfasst wird: die Umsetzung der Unterrichtskomponenten Zielorientierung, Lernbegleitung und Lernatmosphäre (vgl. Seidel & Stürmer, in Druck). Zudem wurde ein Scoring-Verfahren festgelegt. Das Anlegen eines strenges Kriteriums (‚1' Expertennorm getroffen, ‚2' Expertennorm nicht getroffen) erwies sich einem weniger strengen Kriterium unter Berücksichtigung der Tendenz (‚2' Expertennorm getroffen, ‚1' korrekte Tendenz, ‚0' Expertennorm nicht getroffen) hinsichtlich Reliabilität und Varianz als überlegen (vgl. Seidel & Stürmer, in Druck). Im Projektverlauf wurden drei Studien durchgeführt, die den Validitätsaspekt *Inhalt* adressieren. (1) Im Rahmen einer Studie mit drei Experten der Unterrichtsforschung (vgl. Abschnitt 2.2.3.2.4) wurden aus frei zugänglichen, deutschsprachigen Videoportalen Videoclips ausgewählt, in denen jeweils zwei der drei Unterrichtskomponenten gut sichtbar sind (Seidel & Stürmer, in Druck). Daraus resultierte eine finale Auswahl von zwölf Videoclips (Seidel & Stürmer, in Druck). (2) Mit diesen zwölf Videoclips wurde eine Pilotierungsstudie durchgeführt, in der $N = 40$ Lehramtsstudierende unterschiedlicher Semester die Videoclips in Bezug auf Authentizität, Repräsentativität und kognitive Aktivierung einschätzten. Damit konnte empirisch belegt werden, dass die ausgewählten Videoclips als authentisch

und repräsentativ für die drei Unterrichtskomponenten sowie als kognitiv aktivierend wahrgenommen werden (Seidel et al., 2010a). Zusätzlich wurden die zwölf Videoclips hinsichtlich ihrer kognitiven Stimulation und Beanspruchung verglichen (Seidel et al., 2010a). Varianzanalysen zeigen, dass weder über alle Videoclips hinweg noch getrennt nach Unterrichtsfach und repräsentierten Unterrichtskomponenten signifikante Unterschiede in der Wahrnehmung der Videoclips bestehen (Seidel et al., 2010a). (3) Die Repräsentativität der ausgewählten Videoclips für die drei Unterrichtskomponenten wurde in einer weiteren Studie bestätigt, in der $N = 119$ Lehramtsstudierende randomisiert auf zwei Testversionen mit je sechs Videos, die im Hinblick auf Unterrichtsfach und die repräsentierten Unterrichtskomponenten systematisch variiert und rotiert wurden, verteilt wurden (Seidel & Stürmer, in Druck). Diese drei Studien liefern Evidenz dafür, dass die Anforderungen an das Tool Observer, dass die eingesetzten Videoclips als authentische Unterrichtssituationen wahrgenommen werden und repräsentativ für die fokussierten Unterrichtskomponenten Zielorientierung, Lernbegleitung und Lernatmosphäre sind, erfüllt sind.

Darüber hinaus wurden im Rahmen der *Test- und Item-Entwicklung* zwei Studien durchgeführt, die auf die Überprüfung des Validitätsaspekts der *Substanz* abzielen. (1) In der Pilotierungsstudie wurden $N = 40$ Lehramtsstudierende unterschiedlicher Semester während der Bearbeitung des Tools Observer zusätzlich darum gebeten, laut zu denken (Stürmer, Seidel & Blomberg, 2010). Die Laut-Denken-Protokolle verdeutlichen, dass die Studierenden die Videoclips als aktivierend und die standardisierten Rating-Items als unterstützend für die Analyse der Unterrichtssituation wahrnehmen (Stürmer et al., 2010). (2) In einer zweiten Studie mit $N = 109$ Lehramtsstudierenden wurde die Erfassung professioneller Unterrichtswahrnehmung über standardisierte Rating-Items für einen Videoclip mit einer qualitativen Erfassung über ein offenes Antwortformat kombiniert (Schäfer & Seidel, akzeptiert). Die offenen Antworten wurden hinsichtlich der Verwendung von Fachkonzepten und der Übereinstimmung dieser mit einer Experteneinschätzung beurteilt (Schäfer & Seidel, akzeptiert). Dazu wurde analog zur Expertennorm als Referenz für die Qualität der Bearbeitung des Instruments eine Experteneinschätzung von drei Experten mit fünf bis zehn Jahren Erfahrung in der Lehrerbildung und Beobachtung von Unterricht generiert (Schäfer & Seidel, akzeptiert). Die positiven Korrelationen der beiden Messungen sprechen dafür, dass der Prozess professioneller Unterrichtswahrnehmung mit den standardisierten Rating-Items adäquat erfasst wird (Schäfer & Seidel, akzeptiert). Die Ergebnisse dieser beiden Studien deuten darauf hin, dass den Anforde-

rungen an das Instrument hinsichtlich des Auslösens und Erfassens des Prozesses der professionellen Unterrichtswahrnehmung entsprochen wird.

Im Hinblick auf die *Test-Administration* werden mit dem Tool Observer Gruppentestungen an Universitäten im Rahmen von Lehrveranstaltungen oder individuelle Bearbeitungen online von zuhause durchgeführt. Die Teilnahme ist teilweise freiwillig, teilweise obligatorischer Teil einer Lehrveranstaltung.

In Bezug auf die *Test-Nutzung* ist für das Instrument das Abbilden von Unterschieden und Entwicklungen der professionellen Unterrichtswahrnehmung Lehramtsstudierender im Large-Scale-Kontext sowie der Einsatz als Instrument zur Selbstexploration im Rahmen der Studienwahl beabsichtigt. Im Projektverlauf wurden zwei Studien durchgeführt, die der Überprüfung des Validitätsaspekts *Externalität* zuzuordnen sind und den Zusammenhang zwischen professioneller Unterrichtswahrnehmung und dem in einer Lehrveranstaltung vermittelten Wissen über effektives Lehren und Lernen fokussieren. (1) In einer Studie mit $N = 53$ Lehramtsstudierenden in fortgeschrittenen Semestern wurden zu Beginn und am Ende einer einsemestrigen Lehrveranstaltung das deklarative Wissen über lernwirksame Unterrichtskomponenten mit einem Multiple-Choice-Test und professionelle Unterrichtswahrnehmung mit dem Tool Observer erfasst (Stürmer et al., 2013). Varianzanalysen belegen einen signifikanten Zuwachs sowohl für das deklarative Wissen als auch für die professionelle Unterrichtswahrnehmung (Stürmer et al., 2013). (2) Darüber hinaus wurde eine Studie mit $N = 55$ Lehramtsstudierenden zur Überprüfung des Zusammenhangs zwischen der professionellen Unterrichtswahrnehmung Lehramtsstudierender und inhaltsspezifischen Lehrveranstaltungen sowie Interesse durchgeführt (Stürmer, Könings & Seidel, eingereicht). Multiple Regressionsanalysen unter Kontrolle des Semesters zeigen, dass sowohl die Anzahl der besuchten Lehrveranstaltungen zum Thema “Lehren und Lernen” als auch das Interesse an derartigen Inhalten positiv mit der professionellen Unterrichtswahrnehmung Lehramtsstudierender zusammenhängen (Stürmer et al., eingereicht). Die höchste Varianzaufklärung wird dabei für den Aspekt Vorhersagen erreicht (Stürmer et al., eingereicht). Die Befunde dieser beiden Studien weisen darauf hin, dass der Anforderung bezüglich eines Zusammenhangs zwischen professioneller Unterrichtswahrnehmung und dem Erwerb von inhaltsspezifischem Wissen im Rahmen der universitären Lehrerbildung Genüge geleistet wird.

Im Kontext der *Test-Evaluation* wurde im Projektverlauf eine Studie durchgeführt, die den Validitätsaspekt *Struktur* fokussiert. In einer Skalierungsstudie (vgl. Seidel & Stürmer, in Druck) mit $N = 152$ Lehramtsstudierenden eines Universitätsstandorts wurde die Struktur pro-

fessioneller Unterrichtswahrnehmung empirisch überprüft, indem das theoretisch postulierte dreidimensionale Modell professioneller Unterrichtswahrnehmung (Beschreiben, Erklären und Vorhersagen) mit einem eindimensionalen Modell (professionelle Unterrichtswahrnehmung als eine Gesamtfähigkeit) und einem zweidimensionalen Modell (Beschreiben und Integrieren) verglichen wurde. Ein Vergleich der Skalenindizes macht deutlich, dass alle drei Modelle professionelle Unterrichtswahrnehmung reliabel erfassen, jedoch das dreidimensionale Modell am meisten Varianz erklärt (Seidel & Stürmer, in Druck). Zudem weist das dreidimensionale Modell den besten Modellfit (signifikanter Likelihood-Quotienten-Test und kleinstes Bayessches Informationskriterium (BIC)) auf (Seidel & Stürmer, in Druck). Detailliertere Analysen mittels bivariater Korrelationen der Personenfähigkeiten zeigen, dass Beschreiben, Erklären und Vorhersagen interkorrelieren und stark mit der Gesamtfähigkeit professionelle Unterrichtswahrnehmung zusammenhängen (Seidel & Stürmer, in Druck). Die Befunde der Skalierungsstudie liefern erste Evidenz dafür, dass die Anforderung an das Tool Observer, die theoretisch angenommene Struktur professioneller Unterrichtswahrnehmung empirisch abzubilden, erfüllt wird.

2.3.2.4 Ausstehende Überprüfungen vor dem Hintergrund eines übergreifenden Einsatzes des Instruments im Large-Scale-Kontext

Im Projektverlauf wurden bisher insgesamt vier der sechs Aspekte des integrierten Modells von Validität (vgl. Abbildung 3) adressiert: *Inhalt*, *Substanz*, *Struktur* und *Externalität*. Es wurden sechs Studien durchgeführt, die Anforderungen an die Validität des Tools Observer überprüfen, die sich aus der Interpretation der Messergebnisse ableiten. Bezüglich des Validitätsaspekts *Inhalt* ist zu konstatieren, dass Evidenz dafür generiert wurde, dass die Anforderungen, authentische Videoclips einzusetzen, die repräsentativ für die drei fokussierten Unterrichtskomponenten sind, erfüllt werden (vgl. Seidel et al., 2010a; Seidel & Stürmer, in Druck). Im Hinblick auf den Validitätsaspekt *Substanz* wurde Evidenz dafür geliefert, dass den Anforderungen entsprochen wird, den Prozess der professionellen Unterrichtswahrnehmung auszulösen und zu erfassen (vgl. Schäfer & Seidel, akzeptiert; Stürmer et al., 2010). Darüber hinaus liegt hinsichtlich des Validitätsaspekts *Struktur* erste Evidenz dafür vor, dass das Ziel, die theoretisch angenommene Struktur professioneller Unterrichtswahrnehmung mit dem Instrument empirisch abzubilden, erreicht wird (vgl. Seidel & Stürmer, in Druck). Zur Überprüfung der Anforderungen an die Validität des Instruments, die sich aus der Nutzung der Messergebnisse ergeben, wurden im Projektverlauf bisher zwei Studien durchgeführt.

Diese bieten in Bezug auf den Validitätsaspekt *Externalität* Evidenz dafür, dass der Anforderung hinsichtlich eines Zusammenhangs zwischen professioneller Unterrichtswahrnehmung und dem Erwerb von inhaltsspezifischem Wissen im Rahmen der universitären Lehrerbildung entsprochen wird (Stürmer et al., 2013, eingereicht).

Die dargestellten Studien liefern zwar hinreichend Evidenz dafür, dass die Messergebnisse des Tools im Hinblick auf Rückmeldungen auf individueller Ebene (z. B. Lernerfolg und Entwicklung Studierender) valide interpretiert und genutzt werden können. Eine Interpretation und Nutzung der Messergebnisse auf systemischer Ebene (z. B. standortübergreifende Wirksamkeit der Überprüfung von Lehrveranstaltungen) erfordert jedoch einen großflächigen Einsatz des Instruments. Vor dem Hintergrund einer Ausweitung des Einsatzes des Tools auf einen Large-Scale-Kontext gilt es zwei spezifische Anforderungen zu überprüfen, die tiefergehend die Validitätsaspekte *Struktur* und *Externalität* adressieren, und zwei Anforderungen, die auf den Validitätsaspekt *Generalisierbarkeit* abzielen. Diese vier Anforderungen werden im Folgenden ausführlich beschrieben. Darauf aufbauend wird das Ziel der vorliegenden Arbeit, das Tool Observer in Hinblick auf einen großflächigen Einsatz zu validieren, in vier Teilziele ausdifferenziert.

2.3.2.4.1 Validitätsaspekt Struktur

Wie in Abschnitt 2.3.2.3 beschrieben, konnte eine Skalierungsstudie (Seidel & Stürmer, in Druck) bereits erste Evidenz dafür liefern, dass mit dem Tool Observer die theoretisch angenommene Struktur professioneller Unterrichtswahrnehmung mit den drei Aspekten Beschreiben, Erklären und Vorhersagen empirisch abgebildet werden kann. In dieser Studie bearbeiteten $N = 152$ Lehramtsstudierende für Sekundarstufe eines Universitätsstandortes das Instrument. Diese Stichprobe weist jedoch im Wesentlichen drei Limitationen auf: (1) Die lokale Eingrenzung der Stichprobe auf einen Universitätsstandort ist potentiell problematisch, da aufgrund der äußerst heterogenen Studienstrukturlandschaft der Lehrerbildung in Deutschland (Bauer et al., 2010; Blömeke et al., 2009; Keuffer, 2010) offen bleibt, ob die Befunde der Skalierungsstudie (Seidel & Stürmer, in Druck) auf andere Universitätsstandorte generalisierbar sind. In der Heterogenität der deutschen Lehrerbildung zeichnet sich zwar eine Tendenz in Richtung gestufter Studiengänge mit Bachelor- und Masterabschlüssen ab (Winter, 2008), allerdings halten einige Bundesländer, teils parallel dazu, an bisherigen Formen wie dem Staatsexamen fest (Terhart, 2008). Weiterhin existieren in den Bachelorstudiengängen Kon-

zeptionen mit einer lehramtsspezifischen oder einer polyvalenten Ausrichtung. Beide zeichnen sich durch verschiedene Schwerpunktsetzungen aus, die wiederum mit unterschiedlichen, verpflichtenden Studienanteilen in den Bildungswissenschaften einhergehen, wobei die Bandbreite innerhalb der Ausrichtungen stark variiert (Bauer et al., 2011). (2) Einen weiteren limitierenden Faktor bezüglich der Generalisierbarkeit der Befunde der Skalierungsstudie (Seidel & Stürmer, in Druck) stellt die Begrenzung der Stichprobe auf Lehramtsstudierende für Sekundarstufe dar. (3) Darüber hinaus bestehen weitere Einschränkungen aufgrund der geringen Stichprobengröße, die Verzerrungen im Rückschluss auf die Zielpopulation begünstigt (Field, 2009).

Aufgrund der angeführten Limitationen bezüglich der Zusammensetzung und Größe der Stichprobe scheint es angebracht, die Befunde der Skalierungsstudie (Seidel & Stürmer, in Druck) an einer großen und im Hinblick auf Universitätsstandorte und Studiengänge heterogenen Stichprobe zu replizieren. Mit einer derartig angelegten Studie kann in ausreichendem Umfang Evidenz dafür generiert werden, dass das Tool Observer die Anforderung erfüllt, die theoretisch angenommene Struktur professioneller Unterrichtswahrnehmung empirisch abzubilden. Demzufolge verfolgt die vorliegende Arbeit das erste Teilziel, den Validitätsaspekt *Struktur* weiter zu adressieren und zu überprüfen, ob die theoretisch angenommene Struktur professioneller Unterrichtswahrnehmung mit den drei Aspekten Beschreiben, Erklären und Vorhersagen ebenso in einer großen und heterogenen Stichprobe von Lehramtsstudierenden mit dem Tool Observer erfasst wird.

2.3.2.4.2 Validitätsaspekt Generalisierbarkeit über verschiedene Erhebungsbedingungen[5]

Das Tool Observer wird aktuell unter verschiedenen Erhebungsbedingungen eingesetzt. Abhängig von der intendierten Nutzung findet die Bearbeitung des Instruments in Gruppentestungen an Universitäten im Rahmen von Lehrveranstaltungen oder individuell online von zuhause statt. In Bezug auf Effekte der Bedingungen internetbasierter Testverfahren auf die Kompetenzerfassung besteht noch großer Forschungsbedarf (Jurecka, 2008). Die Möglichkeit einer Online-Bearbeitung bietet gerade im Hinblick auf einen Einsatz im Large-Scale-Kontext erhebliche Vorteile, da auf diese Art und Weise schnell große und heterogene Zielgruppen erreicht werden können. Allerdings können Online-Erhebungen auch mit Risiken verbunden sein. Insbesondere bei internetbasierten Lehr-Lernszenarien stellt die Abbruchquote ein po-

[5] Dieser Abschnitt basiert zu Teilen auf Jahn et al. (in Druck).

tentielles Problem dar: Die Hemmschwelle, die Bearbeitung abzubrechen ist in diesem Nutzungskontext im Vergleich zu face-to-face Szenarien deutlich höher (Zumbach & Reimann, 2001). Hohe Abbruchquoten zusammen mit der Gefahr der Mehrfacheingabe können zu einer Vorselektion bzw. Verzerrung der Stichprobe führen und damit die Validität der Interpretation der Messergebnisse gefährden (Birnbaum, 2004; Treiblmaier, 2010). Neben der Variation des Bearbeitungskontexts zwischen Gruppentestung und individueller Online-Bearbeitung unterscheidet sich auch die Art der Teilnahme. Zum Teil bearbeiten die Studierenden das Instrument freiwillig. Doch gerade bei Erhebungen mit Lehramtsstudierenden gestaltet sich die Rekrutierung oft problematisch. Eine ausreichend große Stichprobe kann teilweise nur erzielt werden, wenn die Teilnahme an der Erhebung in eine Lehrveranstaltung integriert wird und somit für den Erwerb von Leistungspunkten verpflichtend ist. Im Hinblick auf die Variation der Art der Teilnahme muss berücksichtigt werden, dass die wahrgenommene Autonomie bei der Entscheidung über eine Teilnahme einen Faktor darstellt, der sich positiv auf die Motivation und die Qualität von Wissensprozessen auswirkt (Prenzel, Seidel & Drechsel, 2004). Vor diesem Hintergrund wird deutlich, dass hinsichtlich der *Generalisierbarkeit* der Nutzung der Messergebnisse über verschiedene Erhebungsbedingungen hinweg Forschungsbedarf besteht. Zentrale Aspekte, die durch unterschiedliche Erhebungsbedingungen beeinflusst werden können, stellen auf Ebene der Bearbeitung des Messinstruments die Abbruchquote und die Bearbeitungszeit dar sowie auf Ebene der Kompetenzmessung das Messergebnis.

Abbruchquote und Bearbeitungszeit können als Indikatoren für die Akzeptanz eines Messinstruments herangezogen werden, geben aber zusätzlich noch Hinweise auf Umfang und Intensität der Bearbeitung. Die Auswertung elektronischer Protokolle der Computernutzung, sogenannter Logfile-Analysen, die zum Standardwerkzeug in der Hypermedia-Forschung zählen (Priemer, 2004), geben Aufschluss über die Bearbeitungszeiten unter den verschiedenen Erhebungsbedingungen. Abbruchquoten fungieren als Kriterium zur Evaluation der Akzeptanz und des Erfolgs von Trainings- oder Lernangeboten. Es liegen Befunde zu höheren Abbruchquoten bei Online-Erhebungen im Vergleich zu Labor-Erhebungen vor (vgl. Birnbaum, 2004). Dementsprechend ist zu erwarten, dass Studierende während einer Online-Bearbeitung häufiger abbrechen als während einer Gruppentestung. Hierbei sind technische Herausforderungen im Rahmen einer Online-Bearbeitung des Tools Observer zu berücksichtigen. Beispielsweise könnte die Integration von Videoclips erhöhte Ladezeiten verursachen. Zusätzlich ist während einer Online-Bearbeitung die Selbstregulationsfähigkeit der Studierenden hinsichtlich der Entscheidungen für die Dauer oder den Abbruch der Bearbeitung stärker gefor-

dert. Dies kommt unter Umständen besonders unter der Bedingung einer freiwilligen Bearbeitung des Tools zum Tragen.

Neben potentiellen Einflüssen unterschiedlicher Erhebungsbedingungen auf Aspekte der Bearbeitung wie Abbruchquote und Bearbeitungszeit stellt sich die Frage, ob sich etwaige Unterschiede in der Motivation der Studierenden bedingt durch mangelnde Wahlmöglichkeit der Teilnahme auf das Ergebnis der Kompetenzmessung auswirken. Zum Beispiel konnte Seidel (2003) einen Zusammenhang zwischen der Qualität der Motivation und der Elaborationstiefe nachweisen. Diese Befunde deuten darauf hin, dass ein unterschiedlicher Grad an Motivation der Studierenden eine unterschiedlich vertiefte Bearbeitung des Instruments bedingen kann.

Wie ausgeführt, sind potentielle Effekte unterschiedlicher Erhebungsbedingungen auf die Bearbeitung des Tools Observer und die damit verbundene Kompetenzerfassung denkbar. Um die Messergebnisse valide interpretieren und nutzen zu können, ist es jedoch von enormer Bedeutung, dass das Instrument der Anforderung entspricht, über verschiedene Erhebungsbedingungen und Arten der Teilnahme hinweg professionelle Unterrichtswahrnehmung Lehramtsstudierender stabil zu erfassen. Demzufolge zielt die vorliegende Arbeit im Rahmen eines zweiten Teilziels darauf ab, die *Generalisierbarkeit* des Tools Observer über die genannten Erhebungsbedingungen hinweg zu überprüfen und zu untersuchen, inwieweit ein Zusammenhang zwischen verschiedenen Erhebungsbedingungen und der Bearbeitung des Tools (Abbruchquote und Bearbeitungszeit) sowie der erfassten professionellen Unterrichtswahrnehmung existiert.

2.3.2.4.3 Validitätsaspekt Generalisierbarkeit über verschiedene Lehramtsstudienrichtungen

Langfristig soll das Tool Observer in verschiedenen Lehramtsstudiengängen eingesetzt werden und Unterschiede über Studiengänge hinweg abbilden. Im bisherigen Projektverlauf wurde das Instrument jedoch ausschließlich mit Lehramtsstudierenden der Sekundarstufe erprobt. Potentielle Herausforderungen eines übergreifenden Einsatzes des Tools sind zum einen in der Zielgruppe und zum anderen im Aufbau des Instruments begründet. Studien mit Studierenden verschiedener Lehramtsstudiengänge konnten immer wieder Unterschiede zwischen diesen Studierendengruppen aufzeigen. Beispielsweise unterscheiden sich Studierende verschiedener Lehramtsstudiengänge bereits zu Studienbeginn hinsichtlich kognitiver Voraussetzungen, Persönlichkeitsmerkmale (Klusmann, Trautwein, Lüdtke, Kunter & Baumert, 2009)

und der Motivation für das Lehramtsstudium (Retelsdorf & Möller, 2012) sowie im Verlauf ihres Studiums hinsichtlich ihres Fachwissens und ihres fachdidaktischem Wissen (Kleickmann & Anders, 2011). Demzufolge ist offen, inwieweit die Ergebnisse aus der Erprobung des Instruments mit Lehramtsstudierenden für Sekundarstufe auf Studierende anderer Lehramtsstudiengänge übertragbar sind. Zudem zeigen die Videoclips, die in das Instrument integriert sind, Unterrichtssequenzen aus der Sekundarstufe. Einerseits ist das Wissen über effektives Lehren und Lernen, wie die im Instrument fokussierten Unterrichtskomponenten Zielorientierung, Lernbegleitung und Lernatmosphäre, zwar schulartübergreifend relevant für die Unterstützung des Lernprozesses auf Seite der Schülerinnen und Schülerinnen (Voss et al., 2011). Andererseits kann jedoch nicht ausgeschlossen werden, dass die Diskrepanz zwischen späterem Handlungsfeld und im Video gezeigtem Unterricht die Kompetenzerfassung beeinflusst.

Vor diesem Hintergrund gilt es, die Anforderung an das Instrument zu überprüfen, professionelle Unterrichtswahrnehmung bei Studierenden unterschiedlicher Lehramtsstudiengänge vergleichbar zu erfassen. Zum einen muss geprüft werden, inwieweit das Instrument die Struktur professioneller Unterrichtswahrnehmung für Studierende verschiedener Lehramtsstudiengänge vergleichbar empirisch abbildet. Darüber hinaus muss sichergestellt werden, dass Studierende verschiedener Lehramtsstudiengänge bei der Bearbeitung weder benachteiligt noch bevorzugt werden. Demnach besteht das dritte Teilziel der vorliegenden Arbeit darin, die *Generalisierbarkeit* des Tools Observer über verschiedene Lehramtsstudiengänge hinweg zu überprüfen und zu untersuchen, inwieweit die Struktur professioneller Unterrichtswahrnehmung und die Itemschwierigkeiten vergleichbar sind.

2.3.2.4.4 Validitätsaspekt Externalität

Wie in Abschnitt 2.3.2.3 beschrieben, konnten zwei Studien bereits zeigen, dass ein Zusammenhang zwischen professioneller Unterrichtswahrnehmung und Wissen über effektives Lehren und Lernen besteht (vgl. Stürmer et al., 2013, eingereicht). Um die Erfassung professioneller Unterrichtswahrnehmung im Verlauf der universitären Lehrerbildung über das Abbilden von integrierten Wissensstrukturen hinaus zu rechtfertigen sowie im Hinblick auf einen Einsatz des Instruments zur Selbstexploration im Rahmen der Studienwahl, ist ein Zusammenhang mit dem Studienerfolg sowie späterem professionellen Handeln im Unterricht erforderlich. Als Indikatoren für den Studienerfolg werden häufig objektive Merkmale wie die

Studienabschlussnote und die Studiendauer sowie subjektive Merkmale wie die Studienzufriedenheit und das Belastungserleben während des Studiums herangezogen (z. B. Blömeke, 2009). Späteres professionelles Handeln wird häufig über den Berufserfolg operationalisiert mittels objektiver Indikatoren wie Berufsstatus sowie subjektiver Indikatoren wie Berufszufriedenheit und Belastungserleben (z. B. Blömeke, 2009). Eine Annäherung kann aber auch über Variablen erfolgen, die Erfolg im Studium oder im Beruf vorhersagen. Neben kognitiven Voraussetzungen der Studierenden, erfasst über die Abiturnote (vgl. Trapmann, Hell, Weigand & Schuler, 2007) oder Studierfähigkeitstests (vgl. Gold & Souvignier, 2008), werden auch Voraussetzungen wie die Motivation oder Persönlichkeit als Prädiktoren herangezogen (vgl. Künsting & Lipowsky, 2011). Auf dieser Grundlage sind Self-Assessments aufgebaut, die zur Überprüfung der Studien- und Berufseignung von angehenden Lehramtsstudierenden eingesetzt werden und persönliche Voraussetzungen der Bewerberinnen und Bewerber erfassen. Aktuell etablierte Verfahren in Deutschland sind „Fit für den Lehrerberuf“ (FIT; Herlt & Schaarschmidt, 2007) und Career Counselling for Teachers (CCT; Nieskens & Müller, 2009). Beide setzen sich aus verschiedenen Interessens- und Persönlichkeitsskalen zusammen, die einen hohen Überschneidungsbereich aufweisen (Köller, Klusmann, Retelsdorf & Möller, 2012). Zwar muss berücksichtigt werden, dass der Aufbau professioneller Kompetenz und damit professionellen Handelns im Verlauf des Studiums und Berufs nicht nur von persönlichen Voraussetzungen, sondern auch von angebotenen Lerngelegenheiten abhängt (vgl. Kunter et al., 2011). Allerdings beeinflussen persönliche Voraussetzungen wiederum die Nutzung dieser Lerngelegenheiten (Kunter et al., 2011). In diesem Zusammenhang, und auch in Bezug auf die Wahrnehmung von Belastungen und Zufriedenheit im Beruf, wird der Studienwahlmotivation (Kombination aus Skalen zu Interesse und Einstellungen) (z. B. Rothland, 2011; Watt & Richardson, 2008) und Persönlichkeitsmerkmalen (z. B. Czerwenka & Nölle, 2011; Tönjes, Dickhäuser & Kröner, 2008) Bedeutung zugeschrieben. Darüber hinaus wird die Studienwahlmotivation auch als relevant für die Qualität des Unterrichtens angesehen (Künsting & Lipowsky, 2011). Hinsichtlich der tatsächlichen Vorhersagekraft beider Konstrukte besteht allerdings noch Forschungsbedarf (Billich-Knapp, Künsting & Lipowsky, 2012; Rothland, 2011). Es liegen Befunde vor, die auf Unterschiede hinsichtlich Studienwahlmotivation und Persönlichkeit zwischen Studierenden verschiedener Lehramtsstudiengänge hinweisen. Fachliches Interesse ist bei Lehramtsstudierenden für Gymnasium stärker ausgeprägt, wohingegen pädagogisches Interesse (z. B. Brookhart & Freeman, 1992; Retelsdorf et al., 2012) sowie Interesse an den bildungswissenschaftlichen Studieninhalten (z. B. Rösler, Zimmermann,

Bauer, Möller & Köller, 2013) stärker bei Studierenden anderer Lehramtsstudiengänge vorhanden sind. Diese weisen zudem höhere Ausprägungen in den Persönlichkeitsmerkmalen Verträglichkeit und Extraversion auf (Klusmann et al., 2009). Jedoch deuten Befunde anderer Studien (z. B. Billich-Knapp et al., 2012; Foerster, 2008) darauf hin, dass auch Studierende eines Lehramtsstudiengangs nicht als homogene Gruppe angesehen werden können, sondern sich in Subgruppen mit verschiedenen motivationalen Voraussetzungen unterteilen.

Im Kontext der Überprüfung des Validitätsaspekts der Externalität des Instruments wurden spezifische Prädiktoren für Studien- und Berufserfolg ausgewählt, für die ein Zusammenhang mit professioneller Unterrichtswahrnehmung vermutet wird: pädagogisches Interesse, Interesse an Bildungswissenschaften, das Persönlichkeitsmerkmal Verträglichkeit sowie eine lernbegleitungsorientiere Vorstellung von Lehren. Im Folgenden wird die Auswahl dieser Prädiktoren begründet.

Das Wissen über lernwirksame Unterrichtskomponenten ist nach Shulman (1987) dem pädagogisch-psychologischen Wissen zuzuordnen. Der Aufbau dieses Wissens findet insbesondere in den bildungswissenschaftlichen Studienanteilen statt (Kunina-Habenicht et al., 2012). In Anbetracht der Kontextspezifität von Interesse (Krapp & Prenzel, 2011) sind für die Nutzung der angebotenen Lerngelegenheiten das generelle pädagogisches Interesse, das an pädagogischen Alltagssituationen orientiert ist, sowie das spezifische Interesse an den Bildungswissenschaften, das differenzierter und stärker an Theorien zum Lehren und Lernen orientiert ist, relevant. Es liegen Befunde über einen positiven Zusammenhang zwischen Interesse und Informationsverarbeitungsprozessen vor, zu denen die professionelle Unterrichtswahrnehmung zählt (Rozendaal, Minnaert & Boekaerts, 2003). Darüber hinaus konnte bei Lehramtsstudierenden bereits ein positiver Zusammenhang zwischen professioneller Unterrichtswahrnehmung und dem Interesse an Lehren und Lernen gezeigt werden (Stürmer et al., eingereicht).

Aus den Persönlichkeitsmerkmalen, die grundsätzlich eng mit Kompetenzen verbunden sind (Spinath, 2012), wurde gemäß der Klassifikation des Fünf-Faktoren Modells (McCrae & Costa, 2008) Verträglichkeit ausgewählt. Verträglichkeit schlägt sich besonders im Mitgefühl und Verständnis für andere nieder (McCrae & Costa, 2008). Generell wird Persönlichkeitsmerkmalen ein Einfluss auf die Informationsverarbeitung zugeschrieben (Humphreys & Revelle, 1984). Speziell im Hinblick auf den Prozess der Beobachtung und Interpretation von Unterrichtssequenzen in Bezug auf die Effektivität des Handelns der Lehrperson könnte sich eine hohe Verträglichkeit positiv auswirken.

Neben Interesse und Persönlichkeit stellen berufsbezogene Überzeugungen von Lehrkräften einen bedeutsamen Aspekt professioneller Kompetenz dar (Baumert & Kunter, 2006). Ihnen wird eine relevante Rolle für die Qualität beruflichen Handelns zugeschrieben (Reusser, Pauli & Elmer, 2011), unter anderem durch die Beeinflussung der Wahrnehmung und Interpretation von Unterrichtssituationen (vgl. Pajares, 1992). Der Schwerpunkt der Forschung liegt dabei auf Überzeugungen zu Lehr- und Lernprozessen (Reusser et al., 2011). Es wird theoretisch argumentiert (Brookhart & Freeman, 1992) und empirisch bestätigt, dass Lehramtsstudierende hinsichtlich ihrer Überzeugungen zu Lehr- und Lernprozessen eine heterogene Gruppe darstellen (Bauer & Drechsel, 2010, März; Drechsel, 2001). Weiter wird angenommen, dass sich eben diese auf die Nutzung der angebotenen Lerngelegenheiten und die Lernaktivität im Studium auswirken (Bauer et al., 2010, März; Schommer, 1993). Demzufolge könnte die Konzeption von Lehren und Lernen einen Zusammenhang mit professioneller Unterrichtswahrnehmung aufweisen und sich auf das spätere unterrichtliche Handeln auswirken. Nachdem bei der Erfassung professioneller Unterrichtswahrnehmung mit dem Tool Observer das Handeln der Lehrperson im Vordergrund steht, wird auf die Überzeugung von Lehren fokussiert. Trotz inkonsistenter Befundlage lässt sich ein Trend in Richtung positiver Effekte konstruktivistischer Überzeugungen auf lern- und motivationsrelevante Merkmale der Unterrichtsgestaltung ausmachen (Reusser et al., 2011).

Zur Überprüfung des Validitätsaspekts der *Externalität* des Tools Observer wird in der vorliegenden Arbeit ein personenzentrierter Ansatz gewählt, der in der aktuellen empirischen Forschung immer mehr an Bedeutung gewinnt (z. B. Bergman & Andersson, 2010; van Eye & Spiel, 2010). Dessen Grundannahme besteht darin, dass durch die aggregierte Analyse auf Variablenebene der Variabilität der Population nicht Rechnung getragen wird (van Eye & Spiel, 2010). Denn für den Fall der Existenz heterogener Subgruppen repräsentieren deskriptive Statistiken der Gesamtstichprobe die Population nicht adäquat (Pastor, Barron, Miller & Davis, 2007). Aus diesem Grund wird als Analyseeinheit die Person gewählt mit dem Ziel, Subgruppen mit ähnlichen Mustern zu identifizieren (Collins & Lanza, 2010). Der personenzentrierte Ansatz setzt sich aus einer personen-orientierten theoretischen Grundlage und einer personen-orientierten Methodik zusammen (Sterba & Bauer, 2010). Im Besonderen bieten modellbasierte Verfahren, wie latente Klassen- oder Profil-Analysen, im Rahmen dieses Ansatzes Vorteile (vgl. Sterba & Bauer, 2010). Diese können zur Identifikation von deutlich abtrennbaren Subgruppen von Personen innerhalb einer Stichprobe angewandt werden (Gollwitzer, 2012). In der Forschung zu allgemeinen und berufsspezifischen Voraussetzungen von

Lehramtsstudierenden wurde der personenzentrierte Ansatz beispielsweise erfolgreich für die Studienwahlmotivation (Billich-Knapp et al., 2012) und die Einstellungen zur Berufswahl (Watt & Richardson, 2008) genutzt. Die Befunde dieser Studien sowie die Zusammensetzung der Stichprobe aus Studierenden unterschiedlicher Lehramtsstudiengänge sprechen dafür, dass nicht zwingend von einer homogenen Population ausgegangen werden kann. Gemäß van Eye (van Eye, 2006) sollte bei einer derartigen Ausgangslage ein personenzentriertes Verfahren gewählt werden, um zu überprüfen, ob sich heterogene Subgruppen identifizieren lassen.

Vor diesem Hintergrund verfolgt die vorliegende Arbeit das vierte Teilziel, den Validitätsaspekt *Externalität* des Tools Observer weiter zu prüfen und in einem ersten Schritt zu untersuchen, inwieweit sich die Lehramtsstudierenden auf Basis der ausgewählten Variablen in Subgruppen mit bestimmten Profilen einordnen lassen, und in einem zweiten Schritt den Zusammenhang zur professionellen Unterrichtswahrnehmung zu überprüfen.

2.3.2.4.5 Zusammenfassung der ausstehenden Überprüfungen

Insgesamt werden in der vorliegenden Arbeit vier Anforderungen an das Tool Observer im Hinblick auf einen übergreifenden Einsatz des Instruments im Large-Scale-Kontext überprüft. Dabei werden insgesamt drei verschiedene Validitätsaspekte adressiert. Erstens ist aufgrund von Einschränkungen der bisherigen Stichprobe noch nicht in ausreichenden Umfang Evidenz dafür generiert worden, dass das Instrument die Anforderung erfüllt, die theoretisch angenommene Struktur professioneller Unterrichtswahrnehmung empirisch abzubilden. Deshalb wird der Validitätsaspekt *Struktur* weiter untersucht. Zweitens ist es von zentraler Bedeutung, dass das Instrument der Anforderung entspricht, über verschiedene Erhebungsbedingungen und Arten der Teilnahme hinweg professionelle Unterrichtswahrnehmung Lehramtsstudierender stabil zu erfassen. Demzufolge wird die *Generalisierbarkeit* über die genannten Erhebungsbedingungen hinweg geprüft. Drittens wird der Einsatz des Instruments in verschiedenen Lehramtsstudiengängen ausgelotet. In diesem Zuge wird im Kontext der *Generalisierbarkeit* erforscht, inwieweit das Instrument professionelle Unterrichtswahrnehmung bei Studierenden unterschiedlicher Lehramtsstudiengänge strukturell vergleichbar erfasst und keine Studierendengruppe bei der Bearbeitung benachteiligt oder bevorzugt. Viertens ist die Beziehung zwischen professioneller Unterrichtswahrnehmung und Variablen, die den Studien- und Berufserfolg vorhersagen, von Interesse. Daher wird mit Hilfe eines personenzentrierten Ansatzes der Validitätsaspekt *Externalität* untersucht.

3 FORSCHUNGSFRAGEN

Vor dem Hintergrund einer Ausweitung des Einsatzes des Tools Observer auf einen Large-Scale-Kontext (vgl. Abschnitt 2.3.2.4) vorliegenden Arbeit, die Validität der Interpretation und Nutzung der Messergebnisse für diesen Einsatzkontext zu überprüfen. Dafür werden drei Validitätsaspekte genauer in den Blick genommen: *Struktur*, *Generalisierbarkeit* und *Externalität*. Jeweils eine Forschungsfrage adressiert die Validitätsaspekte *Struktur* und *Externalität* und zwei Forschungsfragen den Validitätsaspekt *Generalisierbarkeit*. Im Folgenden werden die vier Forschungsfragen sowie die korrespondierenden Hypothesen dargelegt.

3.1 Validitätsaspekt *Struktur*[6]

Die erste Forschungsfrage zielt darauf ab, den Validitätsaspekt *Struktur* weiter zu überprüfen. Daraus lässt sich folgende Frage ableiten:

(1) Lässt sich die theoretisch angenommene Struktur professioneller Unterrichtswahrnehmung mit dem Tool Observer an einer großen und – in Bezug auf Universitätsstandorte und Studiengänge – heterogenen Stichprobe von Lehramtsstudierenden empirisch abbilden?

Es wird erwartet, dass sich die theoretisch angenommene Struktur professioneller Unterrichtswahrnehmung auch bei einer heterogenen Stichprobe als eine Gesamtfähigkeit mit den drei Aspekten Beschreiben, Erklären und Vorhersagen empirisch abbilden lässt.

3.2 Validitätsaspekt *Generalisierbarkeit*

Im Zentrum der zweiten und dritten Forschungsfrage steht der Validitätsaspekt *Generalisierbarke*it. In der zweiten Forschungsfrage liegt der Fokus auf unterschiedlichen Erhebungsbedingungen, in der dritten Forschungsfrage werden unterschiedliche Lehramtsstudiengänge in den Blick genommen.

[6] Dieser Abschnitt basiert zu Teilen auf Jahn et al. (in Druck).

3.2.1 *Generalisierbarkeit* über unterschiedliche Erhebungsbedingungen hinweg[7]

Hinsichtlich der Generalisierbarkeit über unterschiedliche Erhebungsbedingungen hinweg wirf folgende Forschungsfrage überprüft:

(2) Gibt es einen Zusammenhang zwischen unterschiedlichen Erhebungsbedingungen und der Kompetenzerfassung durch das Tool Observer?

a) Gibt es einen Zusammenhang zwischen den beiden Erhebungsbedingungen Bearbeitungskontext („Online"/„On-site") und Art der Teilnahme („Freiwillig"/ „Obligatorisch") und der Abbruchquote sowie der Bearbeitungsdauer?

Wie in Abschnitt 2.3.2.4.2 dargelegt, ist eine höhere Abbruchquote bei einer „frei willigen Teilnahme" vor allem unter der Bedingung „Online" zu erwarten und eine durchschnittlich längere Bearbeitungszeit unter der Bedingung „Online".

b) Gibt es einen Zusammenhang zwischen den beiden Erhebungsbedingungen Bearbeitungskontext („Online"/„On-site") und Art der Teilnahme („Freiwillig"/ „Obligatorisch") und der professionellen Unterrichtswahrnehmung Lehramtsstudierender?

Es wird kein Zusammenhang zwischen den beiden Erhebungsbedingungen (Bearbeitungskontext und Art der Teilnahme) und der professionellen Unterrichtswahrnehmung Lehramtsstudierender erwartet (vgl. Abschnitt 2.3.2.4.2).

3.2.2 *Generalisierbarkeit* über unterschiedliche Lehramtsstudiengänge hinweg[8]

Bezüglich der *Generalisierbarkeit* über unterschiedliche Erhebungsbedingungen hinweg wird folgende Forschungsfrage fokussiert:

(3) Sind die Struktur professioneller Unterrichtswahrnehmung sowie die Itemschwierigkeiten für unterschiedliche Studierendengruppen (Lehramt für Primarstufe, Sekundarstufe und Berufliche Schulen) vergleichbar?

Es wird eine vergleichbare Kompetenzstruktur für Studierende unterschiedlicher Lehramtsstudiengänge (Lehramt für Primarstufe, Sekundarstufe und Berufliche Schulen)

[7] Dieser Abschnitt basiert zu Teilen auf Jahn, G., Prenzel, M., Stürmer, K. & Seidel, T. (2011). Varianten einer computergestützten Erhebung von Lehrerkompetenzen: Untersuchungen zu Anwendungen des Tools „Observer". *Unterrichtswissenschaft, 39*, 136-153.

[8] Dieser Abschnitt basiert zu Teilen auf Jahn et al. (in Druck).

erwartet. Hinsichtlich der Itemschwierigkeiten werden keine Auffälligkeiten für Lehramtsstudierende der Sekundarstufe erwartet, da das Instrument ursprünglich für diesen Bereich entwickelt wurde (z. B. Auswahl der Videoclips). Aufgrund mangelnder Erkenntnisse zur Anwendung bei Studierenden anderer Lehramtsstudiengänge lassen sich dafür keine konkreten Hypothesen ableiten (vgl. Abschnitt 2.3.2.4.3).

3.3 Validitätsaspekt *Externalität*

Im Rahmen der vierten Forschungsfrage wird der Validitätsaspekt *Externalität* überprüft. Als externale Kriterien werden Variablen verwendet, die typischerweise im Rahmen von Verfahren zur Eignungsfeststellung als Prädiktoren für Studien- und Berufserfolg erfasst werden. Konkret wird folgende Forschungsfrage untersucht:

(4) Gibt es einen Zusammenhang zwischen professioneller Unterrichtswahrnehmung und Variablen, die typischerweise als Prädiktoren für Studien- und Berufserfolg herangezogen werden?

a) Lassen sich Lehramtsstudierende in Subgruppen mit bestimmten Profilen (der Variablen pädagogisches Interesse, Interesse an Bildungswissenschaften, lernbegleitungsorientierter Lehrbegriff und Verträglichkeit) einordnen?

Wie in Abschnitt 2.3.2.4.4 diskutiert, wird erwartet, dass sich Subgruppen von Lehramtsstudierenden mit bestimmten Profilen identifizieren lassen.

b) Gibt es einen Zusammenhang zwischen der Zugehörigkeit zu diesen Profilen und der professionellen Unterrichtswahrnehmung?

Bezüglich der einzelnen Variablen, die in die Profilbildung eingehen, pädagogisches Interesse, Interesse an Bildungswissenschaften, lernbegleitungsorientierter Lehrbegriff und Verträglichkeit, wird, wie in Abschnitt 2.3.2.4.4 elaboriert, ein positiver Zusammenhang mit professioneller Unterrichtswahrnehmung erwartet.

4 EMPIRISCHE ÜBERPRÜFUNG DES VALIDITÄTSASPEKTS *STRUKTUR*

Im Rahmen der ersten Forschungsfrage wird der Validitätsaspekt *Struktur* überprüft. Hierbei wird untersucht, inwieweit sich die theoretisch angenommene Struktur professioneller Unterrichtswahrnehmung mit dem Instrument an einer großen und – in Bezug auf Universitätsstandorte und Studiengänge – heterogenen Stichprobe von Lehramtsstudierenden abbilden lässt. Dazu wird zunächst das methodische Vorgehen beschrieben. Anschließend werden die Ergebnisse dargestellt und diskutiert.

4.1 Methodisches Vorgehen

Im Folgenden wird auf Stichprobe und Datenerhebung, Messinstrumente und Auswertungsmethoden eingegangen.

4.1.1 Stichprobe und Datenerhebung[9]

Die Stichprobe zur Überprüfung der ersten Forschungsfrage besteht aus $N = 1029$ Lehramtsstudierenden (61.4 % weiblich), die durchschnittlich im dritten Fachsemester ($M = 2.66$, $SD = .79$) Fachsemester studieren. Sie stellt eine heterogene Mischung aus Lehramtsstudierenden unterschiedlicher Lehramtsstudienrichtungen (Lehramt für Primarstufe, Sekundarstufe I, Sekundarstufe II und Berufliche Schulen) dar, die an Universitäten mit unterschiedlichen Studienstrukturen (traditionelles Staatsexamen, modularisiertes Staatsexamen, polyvalenter Bachelor und Lehramts-Bachelor) studieren.

Diese Stichprobe mit $N = 1029$ Lehramtsstudierenden wird im weiteren Verlauf der Arbeit als Scaling-up-Stichprobe bezeichnet, um sie klar von der Stichprobe der Skalierungsstudie (Seidel & Stürmer, in Druck) mit einer deutlich kleineren und homogeneren Gruppe von Lehramtsstudierenden abzugrenzen. Dies ist von besonderer Relevanz, da die Ergebnisse der empirischen Überprüfung der Kompetenzstruktur aus beiden Stichproben im Rahmen der Untersuchung der ersten Forschungsfrage gegenübergestellt und verglichen werden. Einen Überblick über die deskriptive Statistik und Verteilung der Scaling-up-Stichprobe auf Lehramtsstudiengänge und Studienstrukturen liefert Tabelle 1.

[9] Dieser Abschnitt basiert zu Teilen auf Jahn et al. (in Druck).

Tabelle 1

Deskriptive Statistik der Scaling-up-Stichprobe

	Scaling-up-Stichprobe (N = 1029)	
	M	*SD*
Alter	22.56	3.19
Semester Erstfach	2.66	0.79
Semester Zweitfach	2.55	0.83
	Anteil (in %)	
Geschlecht weiblich	61.4	
Studienstruktur		
Lehramts-Bachelor	27.8	
Polyvalenter Bachelor	27.9	
Modularisiertes Staatsexamen	37.6	
Traditionelles Staatsexamen	6.6	
Lehramtsstudiengänge		
Berufliche Schulen	16.6	
Sekundarstufe II	47.8	
Sekundarstufe I	17.4	
Primarstufe	16.1	
Andere	2.0	

Die Datenerhebung wurde deutschlandweit im Sommersemester 2010 und im Wintersemester 2010/2011 durchgeführt. Dabei wurden Daten an jenen 13 Universitäten erhoben, die auch am Panel zum Lehramtsstudium (PaLea; Bauer et al., 2010) teilnehmen. Zusätzlich wurde das Instrument an drei weiteren interessierten Universitäten eingesetzt. Die Datenerhebung fand teilweise eingebettet in Lehrveranstaltungen als Gruppentestung an der Universität statt, teilweise bearbeiteten die Studierenden das Instrument freiwillig individuell von zuhause. Diese Variation der Erhebungsbedingungen kann jedoch vernachlässigt werden, da sie keinen Effekt auf die Erfassung professioneller Unterrichtswahrnehmung durch das Tool Observer hat. Diese Generalisierbarkeit der Messwerte über verschiedene Erhebungsbedingungen hinweg wird

ausführlich im Rahmen der zweiten Forschungsfrage (vgl. Kapitel 5) untersucht. Alle Studierenden erhielten für die Bearbeitung des Instruments einen Gutschein im Wert von 15 Euro. Insgesamt liegen damit Daten von Lehramtsstudierenden aus 16 deutschen Universitäten vor. Die einzelnen Universitätsstandorte werden in der vorliegenden Arbeit anonymisiert dargestellt. Die Verteilung der $N = 1029$ Lehramtsstudierenden auf die verschiedenen Universitäten ist Tabelle 2 zu entnehmen. Die Studierenden verteilen sich relativ gleichmäßig auf die einzelnen Universitäten, wobei zwei Universitätsstandorte mit weniger als einem Prozent der Studierenden in der Gesamtstichprobe vertreten sind.

Tabelle 2

Verteilung der Lehramtsstudierenden der Scaling-up-Stichprobe (N = 1029) auf die einzelnen deutschen Universitäten

	Universitätsstandorte															
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Anteil (in %)	9.3	11.1	9.6	0.8	6.6	5.2	6.5	11.1	1.0	2.8	0.6	16.3	13.1	1.4	3.1	1.5

4.1.2 Messinstrument

Zur Erfassung professioneller Unterrichtswahrnehmung wird das videobasierte Online-Tool Observer (Seidel et al., 2010b) eingesetzt. Eine detaillierte Beschreibung des Instruments findet sich in Abschnitt 2.2.3.2.4.

4.1.3 Auswertungsmethoden

Die Analysen im Rahmen der ersten Forschungsfrage basieren auf der Item-Response-Theorie. Es wird angenommen, dass die Einschätzungen der Videoclips des Tools Observer einen Indikator für die latente Variable professionelle Unterrichtswahrnehmung darstellen. Bevor die Analyseverfahren beschrieben werden, wird auf theoretische Grundlagen der Item-Response-Theorie und deren Anwendung im Rahmen der Überprüfung von psychometrischen Eigenschaften von Messinstrumenten eingegangen.

4.1.3.1 Theoretische Grundlagen zur Auswertungsmethode

Die Item-Response-Theorie, auch probabilistische Testtheorie oder Latent-Trait-Theorie genannt, wird häufig als konkurrierende Theorie zur klassischen Testtheorie, auch als Messfehlertheorie bezeichnet, dargestellt (z. B. Molenaar, 1995). Auf eine Diskussion der Vor- und Nachteile wird an dieser Stelle verzichtet, da in der vorliegenden Arbeit beide Testtheorien vielmehr als komplementäre Ansätze betrachtet werden (vgl. Rost, 2004, 1999). In der klassischen Testtheorie wird angenommen, dass sich der beobachtete Messwert aus einem wahren Wert und einem Messfehler zusammensetzt (de Ayala, 2009), die Existenz eines Messwertes wird dabei vorausgesetzt (Rost, 2004). Im Gegensatz zu dieser fehlerbehafteten direkten Messung des im Test erfassten Konstrukts wird in der Item-Response-Theorie explizit eine indirekte Messung postuliert (Rauch & Hartig, 2012). Die zentralen Annahmen der Item-Response-Theorie beziehen sich darauf, wie die beobachteten Messwerte eines Instruments von der zu messenden latenten Variable abhängen (Rost, 2004). Die Wahrscheinlichkeit, dass eine bestimmte Person ein bestimmtes Item löst, kann als eine Funktion der Personenfähigkeit und eines oder mehrerer Itemparameter beschrieben werden (Molenaar, 1995). Personen- und Itemparameter werden dabei auf einer gemeinsamen Skala verortet (de Ayala, 2009), können aber separat geschätzt werden (Embretson & Reise, 2000). Dadurch ist es möglich, den Einfluss und die Interaktion von individuellen und situativen Faktoren auf den Messwert empirisch zu untersuchen (Koeppen et al., 2008). Dies ist für die Erfassung von Kompetenzen von besonderer Bedeutung (vgl. Abschnitt 2.1.2.1.3).

Innerhalb der Item-Response-Theorie können verschiedene Messmodelle für dichotome Items unterschieden werden: das Ein-Parameter-Logistische Modell (1PL-Modell) sowie dessen Erweiterungen das Zwei-Parameter-Logistische Modell (2PL-Modell) und das Drei-Parameter-Logistische Modell (3PL-Modell). Während im 1PL-Modell die Lösungswahrscheinlichkeit neben der Personenfähigkeit ausschließlich von der Itemschwierigkeit abhängt, wird im 2PL-Modell zusätzlich die unterschiedliche Trennschärfe der Items modelliert und im 3PL-Modell als weitere Ergänzung ein Rateparameter berücksichtigt (Bühner, 2011). Die Bezeichnung Rasch-Modell wird in der Literatur häufig synonym zum Begriff 1PL-Modell verwendet, da beide Modelle unterschiedliche Itemschwierigkeiten bei konstanter Trennschärfe modellieren (de Ayala, 2009). Streng genommen unterscheiden sich die beiden Modelle aber darin, dass diese Konstante im Rasch-Modell auf 1 festgesetzt ist und für das 1PL-Modell beliebig gewählt werden kann (de Ayala, 2009). Nachdem diese Differenzierung für

die vorliegende Arbeit nicht von Bedeutung ist, werden im Folgenden beide Begriffe synonym verwendet.

Die Item-Response-Theorie wird typischerweise angewandt, um die psychometrischen Eigenschaften von Tests, insbesondere die Validität, zu überprüfen (Furr & Bacharach, 2008). Dazu wird meist der Itemfit analysiert. Dieser ist ein wichtiger Kennwert, der ausdrückt, inwieweit die erhobenen Daten den erwarteten Werten unter Annahme des (Rasch-)Modells entsprechen (Bond & Fox, 2007) und damit dieses Testmodell der Datenauswertung zu Grunde gelegt werden kann. Denn für die Interpretation und Nutzung der Messwerte ist es entscheidend, dass das angenommene Modell auf die Daten passt (Rost, 2004). Es existieren verschiedene Indizes zur Beurteilung des Itemfits, die größtenteils residuenbasierten oder likelihoodbasierten Fitmaßen zuzuordnen sind (Rost, 2004). In der vorliegenden Arbeit wurde ein residuenbasierter Fit-Index zur Beurteilung des Itemfits herangezogen, der Mean Square (MNSQ), der bei zusätzlicher Gewichtung nach der Varianz der Itemantworten weniger sensibel für Ausreißer ist (Wu & Adams, 2007). Dieser kann mit der Software ConQuest (Wu, Adams & Wilson, 1998) berechnet werden. Der Erwartungswert des gewichteten Mean Squares (wMNSQ) ist 1 und entspricht dem perfekten Modellfit. Werte < 1 weisen auf einen Overfit hin, was bedeutet, dass die Daten zu deterministisch sind und eher den Antwortmustern einer Guttman-Skala ähneln (Bond & Fox, 2007). Dahingegen deuten Werte > 1 auf einen Underfit hin, bei welchem sich eher willkürliche Antwortmuster zeigen und die Daten von denen unter Annahme des Rasch-Modells erwarteten abweichen (Bond & Fox, 2007). In der Testpraxis ist immer eine gewisse Variation in den Antwortmustern gegeben. Deshalb empfehlen Bond und Fox (2007) einen Wertebereich für einen akzeptablen Modellfit von $0.75 \leq \text{wMNSQ} \leq 1.30$, der je nach Kontext angepasst werden sollte (z. B. $0.80 \leq \text{wMNSQ} \leq 1.20$ bei Schulleistungsstudien). Zusätzlich wird in ConQuest (Wu et al., 1998) ein weiterer Parameter zur Überprüfung der Qualität der Items berechnet, die Itemtrennschärfe gemäß der klassischen Testtheorie. Diese entspricht der Korrelation des Items mit dem gesamten Testergebnis, exklusive des jeweiligen Items (Rost, 2004). Die Trennschärfe beschreibt das Ausmaß, mit dem Items in der Lage sind, zwischen Personen mit hoher und niedriger Merkmalsausprägung zu differenzieren (Furr & Bacharach, 2008). Eine Trennschärfe nahe 0 deutet darauf hin, dass die Lösung des Items nicht mit dem zu messenden Konstrukt zusammenhängt, und eine negative Trennschärfe bedeutet sogar, dass ein inverser Zusammenhang zwischen der Itemlösung und dem zu messenden Konstrukt besteht (Furr & Bacharach, 2008). Demnach sind Items mit einer posi-

tiven Trennschärfe für die Testkonstruktion erstrebenswert (Furr & Bacharach, 2008), während Items mit einer negativen Trennschärfe in der Regel ungeeignet sind (Bühner, 2011).

4.1.3.2 Analyseverfahren[10]

Zur Überprüfung des Validitätsaspekts *Struktur* wurde die Kompetenzstruktur professioneller Unterrichtswahrnehmung untersucht und die Ergebnisse denen der Skalierungsstudie (Seidel & Stürmer, in Druck) gegenübergestellt. Dabei wurde methodisch analog zur Skalierungsstudie (Seidel & Stürmer, in Druck) vorgegangen. Für die Analyse der 0-1-Datenmatrix wurde ein Rasch-Modell zu Grunde gelegt. Folglich werden alle Items durch einen Itemparameter beschrieben und eine Vergleichbarkeit der Itemtrennschärfe wird postuliert.

Mithilfe der Software ConQuest (Wu et al., 1998) wurden in einem ersten Schritt der Itemfit und die Trennschärfe der $N = 216$ Items analysiert und mit denen der Skalierungsstudie (Seidel & Stürmer, in Druck) verglichen. Zur Beurteilung des Itemfits liefert ConQuest (Wu et al., 1998) als Fit-Index den wMNSQ (vgl. Abschnitt 4.1.3.1). Als Ausschlusskriterium wurde analog zur Skalierungsstudie (Seidel & Stürmer, in Druck) der von Bond und Fox (2007) vorgeschlagene Wertebereich für einen akzeptablen Modellfit herangezogen (vgl. Abschnitt 4.1.3.1): Items mit einem $\text{wMNSQ} \leq 0.75$ wurden aufgrund eines Overfits ausgeschlossen und Items mit einem $\text{wMNSQ} \geq 1.30$ aufgrund eines Underfits. Es zeigte sich, dass die Items, die bereits in der Skalierungsstudie (Seidel & Stürmer, in Druck) aufgrund eines schlechten Modellfits ausgeschlossen wurden, in der Scaling-up-Studie vergleichbar schlechte Werte aufwiesen. Demnach wurde in einem zweiten Schritt eine Reduktion der Items auf den finalen Itempool der Skalierungsstudie (Seidel & Stürmer, in Druck) von $n = 112$ Items vorgenommen und eine erneute Skalierung durchgeführt. Weitere $n = 6$ Items wurden aufgrund negativer Trennschärfe ausgeschlossen. Die verbleibenden $n = 106$ Items weisen zufriedenstellende Itemfit-Werte auf und verteilen sich ausgewogen auf die drei Aspekte Beschreiben ($n = 38$), Erklären ($n = 36$) und Vorhersagen ($n = 32$). Auf diesem Itempool basieren die nachfolgenden Analysen.

Zur empirischen Überprüfung der theoretisch angenommenen Kompetenzstruktur wurden unterschiedliche (mehr-)dimensionale Rasch-Modelle berechnet. Theoretisch postuliert wird das dreidimensionale Modell professioneller Unterrichtswahrnehmung mit den drei Aspekten Beschreiben, Erklären und Vorhersagen. Dagegen wurden folgende Modelle geprüft: ein ein-

[10] Dieser Abschnitt basiert zu Teilen auf Jahn et al. (in Druck).

dimensionales Modell mit professioneller Unterrichtswahrnehmung als einer Gesamtfähigkeit sowie ein zweidimensionales Modell mit Beschreiben und Integrieren (Erklären und Vorhersagen werden als zwei ineinander integrierte Aspekte angesehen). Für diesen Modellvergleich wurden sowohl Skalenindizes (EAP/PV-Reliabilitäten und erklärte Varianz) verglichen als auch Likelihood-Quotienten-Tests basierend auf der Final Deviance (entspricht der Log-Likelihood) berechnet. Aufgrund der Stichprobenabhängigkeit der χ^2-verteilten Teststatistik wurden diese ausschließlich, entsprechend der gängigen Forschung (z. B. Leutner, Fleischer, Wirth, Greiff & Funke Joachim, 2012), in Kombination mit dem Bayesschen Informationskriterium (BIC) interpretiert. Alle Modellvergleiche wurden denen der Skalierungsstudie (Seidel & Stürmer, in Druck) gegenübergestellt.

Darüber hinaus wurde eine detaillierte Analyse der strukturellen Zusammenhänge durchgeführt. Dafür wurden bivariate Korrelationen nach Pearson für die Personenfähigkeiten der professionellen Unterrichtswahrnehmung und ihrer drei Aspekte Beschreiben, Erklären und Vorhersagen berechnet. Die Personenfähigkeiten wurden dafür über die prozentuale Übereinstimmung mit der Expertennorm berechnet.

Der Umfang fehlender Werte ist mit durchschnittlich 0.5 % zu vernachlässigen.

4.2 Ergebnisse[11]

Zur Überprüfung des Validitätsaspekts *Struktur* im Rahmen der ersten Forschungsfrage wurde untersucht, inwieweit sich die theoretisch angenommene Struktur professioneller Unterrichtswahrnehmung mit dem Tool Observer an einer großen und – in Bezug auf Universitätsstandorte und Studiengänge – heterogenen Stichprobe von Lehramtsstudierenden empirisch abbilden lässt. Dazu wurde die Kompetenzstruktur in der Scaling-up-Studie überprüft und mit den Befunden der Skalierungsstudie (Seidel & Stürmer, in Druck) verglichen. Im Folgenden werden die Ergebnisse der Modellvergleiche sowie die Analyse struktureller Zusammenhänge vorgestellt.

4.2.1 Modellvergleiche

In einem ersten Schritt wurden IRT-Skalierungen für ein ein-, zwei- und dreidimensionales Modell professioneller Unterrichtswahrnehmung durchgeführt und die Skalenkennwerte verglichen. Die Schätzungen für EAP/PV-Reliabilitäten und Varianzen beruhen dabei jeweils auf unidimensionalen Schätzungen. Einen Überblick über die EAP/PV-Reliabilitäten und Varianzen der drei Modelle bietet Tabelle 3 für die Scaling-up-Studie im Vergleich zur Skalierungsstudie (Seidel & Stürmer, in Druck). Hinsichtlich einer reliablen Messung professioneller Unterrichtswahrnehmung zeigen alle drei Modelle in beiden Studien zufriedenstellende Kennwerte ($r \geq .81$). Allerdings erreicht das eindimensionale Modell sowohl in der Scaling-up-Studie ($r = .94$) als auch in der Skalierungsstudie (Seidel & Stürmer, in Druck) ($r = .96$) die höchste EAP/PV-Reliabilität. Die höchste Varianz, als Indikator für die Trennschärfe der Items (vgl. Leutner et al., 2012), weist in der Scaling-up-Studie ($0.63 \leq \sigma^2 \leq 1.89$) sowie in der Skalierungsstudie ($0.80 \leq \sigma^2 \leq 2.14$) jeweils das dreidimensionale Modell auf. Beim Vergleich der beiden Studien fällt jedoch auf, dass alle Modelle in der Scaling-up-Studie in der Tendenz niedrigere EAP/PV-Reliabilitäten und Varianzen erreichen als in der Skalierungsstudie (Seidel & Stürmer, in Druck).

[11] Dieser Abschnitt basiert zu Teilen auf Jahn et al. (in Druck).

Tabelle 3

Skalenkennwerte des ein-, zwei- und dreidimensionalen Modells professioneller Unterrichtswahrnehmung in der Scaling-up-Studie im Vergleich zur Skalierungsstudie (Seidel & Stürmer, in Druck)[12]

	1-dim	2-dim Modell	3-dim	1-dim	2-dim Modell	3-dim
	Scaling-up-Studie (N = 1029)			*Skalierungsstudie (N = 152)*		
Reliabilität	.94			.96		
Beschreiben		.81	.81		.90	.90
Erklären		.93	.83		.90	.91
Vorhersagen			.87			.97
Varianz	0.93			1.24		
Beschreiben		0.63	0.63		0.80	0.80
Erklären		1.29	0.91		1.70	1.33
Vorhersagen			1.89			2.14

Anmerkung. EAP/PV-Reliabilitäten und Varianzen basieren auf unidimensionalen Schätzern.

In einem zweiten Schritt wurden globalere Kennwerte für den Modellfit herangezogen: BIC, Deviance und die Anzahl der geschätzten Parameter. Wie Tabelle 4 zu entnehmen ist, weist das dreidimensionale Modell in der Scaling-up-Studie vergleichsweise den geringsten BIC auf. Auch der Likelihood-Quotienten-Test basierend auf der Deviance spricht dafür, dass das dreidimensionale Modell die Daten signifikant besser abbildet als das ein- oder zweidimensionale Modell. Damit liefern diese globaleren Kennwerte für den Modellfit für die Scaling-up-Studie ein vergleichbares Bild wie in der Skalierungsstudie (Seidel & Stürmer, in Druck) (vgl. Tabelle 4).

Zusammenfassend ist festzuhalten, dass der Vergleich des ein-, zwei- und dreidimensionalen Modells professioneller Unterrichtswahrnehmung bezüglich Skalenkennwerte sowie globalerer Kennwerte für den Modellfit analog zur Skalierungsstudie (Seidel & Stürmer, in Druck) zugunsten des dreidimensionalen Modells ausfällt.

[12] adaptiert nach Jahn et al. (in Druck).

Tabelle 4

Modellvergleich des ein-, zwei- und dreidimensionalen Modells professioneller Unterrichtswahrnehmung für die Scaling-up-Studie im Vergleich zur Skalierungsstudie (Seidel & Stürmer, in Druck)[13]

	Deviance	Parameter	ΔDeviance (*df*)	BIC
Scaling-up-Studie (N = 1029)				
dreidimensional	119802	112	-	120578
zweidimensional	120101	109	299** (3)	120857
eindimensional	120724	107	922** (5)	121466
Skalierungsstudie (N = 152)				
dreidimensional	16874	118	-	17466
zweidimensional	16898	115	24** (3)	17475
eindimensional	17002	113	128** (5)	17569

Anmerkungen. Baseline für den Modellvergleich ist das dreidimensionale Modell. *Deviance:* -2log[LR] der Modellschätzung. Δ*Deviance:* χ^2-verteilte Teststatistik des Likelihood-Quotienten-Tests, Differenz der geschätzten Parameter als Freiheitsgrade *(df).* ** $p < .01$.

[13] adaptiert nach Jahn et al. (in Druck).

4.2.2 Analyse struktureller Zusammenhänge

Zur weiteren Analyse der strukturellen Zusammenhänge innerhalb der Kompetenzstruktur professioneller Unterrichtswahrnehmung wurden bivariate Korrelationen nach Pearson zwischen der Gesamtfähigkeit und den Fähigkeitsaspekten Beschreiben, Erklären und Vorhersagen der Personen berechnet. Eine Gegenüberstellung der mittleren Personenfähigkeiten sowie deren Interkorrelationen von Scaling-up-Studie und Skalierungsstudie ist Tabelle 5 zu entnehmen. Die durchschnittlichen Personenfähigkeiten für professionelle Unterrichtswahrnehmung sowie Beschreiben, Erklären und Vorhersagen in der Scaling-up-Stichprobe entsprechen mit einer Übereinstimmung mit der Expertennorm von 37 – 44 % einem mittleren Fähigkeitsniveau. Damit rangieren sie in einem ähnlichen Bereich wie in der Skalierungsstudie mit einer durchschnittlichen Übereinstimmung mit der Expertennorm von 31 – 44 % (Seidel & Stürmer, in Druck). Für die Korrelationen zwischen Gesamtfähigkeit und Aspekten professioneller Unterrichtswahrnehmung zeigt sich in Scaling-up-Studie und Skalierungsstudie ein konformes Bild. Die Personenfähigkeiten Beschreiben, Erklären und Vorhersagen korrelieren zum einen hoch mit der Gesamtpersonenfähigkeit professioneller Unterrichtswahrnehmung ($r \geq .89$ bzw. $r \geq .92$). Zum anderen zeigen sich hohe systematische Interkorrelationen zwischen den Personenfähigkeiten der drei Aspekte, wobei die Korrelation zwischen Erklären und Vorhersagen jeweils am höchsten ausfällt ($r = .87$ bzw. $r = .89$). Die Interkorrelationen zwischen den Personenfähigkeiten der einzelnen Aspekte übersteigen jedoch in keinem Fall die Korrelationen mit der Gesamtpersonenfähigkeit.

Tabelle 5

Mittlere Personenfähigkeiten und deren bivariate Interkorrelationen nach Pearson im Vergleich zur Skalierungsstudie (Seidel & Stürmer, in Druck)[14]

Skala	*M*	*SD*	PU	Beschreiben	Erklären
PU	.41	.18	_		
Beschreiben	.44	.17	.89**	_	
Erklären	.37	.18	.95**	. 78**	_
Vorhersagen	.43	.24	.94**	.73**	.87**
PU	.37	.18	_		
Beschreiben	.44	.18	.92**	_	
Erklären	.31	.18	.95**	.82**	_
Vorhersagen	.35	.22	.94**	.77**	.89**

Anmerkungen. PU = Professionelle Unterrichtswahrnehmung. Werte entsprechen der prozentualen Übereinstimmung mit der Expertennorm. $p < .01$.

4.2.3 Zusammenfassung der Ergebnisse

In der Scaling-up-Studie bildet das theoretisch postulierte dreidimensionale Modell professioneller Unterrichtswahrnehmung die Daten besser ab als das ein- oder zweidimensionale Modell. Die detaillierte Analyse der strukturellen Zusammenhänge mithilfe bivariater Korrelationen macht deutlich, dass Beschreiben, Erklären und Vorhersagen aufgrund hoher Korrelationen mit der Gesamtfähigkeit professionelle Unterrichtswahrnehmung keine getrennten Dimensionen, sondern vielmehr Aspekte professioneller Unterrichtswahrnehmung darstellen. Die Befunde der Scaling-up-Studie sind vergleichbar mit denen der Skalierungsstudie (Seidel & Stürmer, in Druck). Folglich ist zu konstatieren, dass das Tool Observer – entsprechend der Hypothese – die theoretisch angenommene Struktur professioneller Unterrichtswahrnehmung als eine Gesamtfähigkeit, differenziert in die drei Aspekte Beschreiben, Erklären und Vorhersagen, in einer großen und – in Bezug auf Universitätsstandorte und Studiengänge – heterogenen Stichprobe von Lehramtsstudierenden empirisch abbildet.

[14] vergleiche Jahn et al. (in Druck).

4.3 Diskussion[15]

Vor dem Hintergrund einer Ausweitung des Einsatzes des Tools Observer auf einen Large-Scale-Kontext wurde im Rahmen der ersten Forschungsfrage der Validitätsaspekt *Struktur* weiter überprüft. Es wurde untersucht, inwieweit sich die theoretisch angenommene Struktur professioneller Unterrichtswahrnehmung mit dem Tool Observer an einer großen und – in Bezug auf Universitätsstandorte und Studiengänge – heterogenen Stichprobe von Lehramtsstudierenden empirisch abbilden lässt. Die Kompetenzstruktur professioneller Unterrichtswahrnehmung wurde an einer Stichprobe von $N = 1029$ Lehramtsstudierenden aus 16 deutschen Universitätsstandorten und vier Studiengängen (Lehramt für Primarstufe, Sekundarstufe I, Sekundarstufe II und Berufliche Schulen) empirisch überprüft und mit den Befunden der Skalierungsstudie (Seidel & Stürmer, in Druck) verglichen. Im Folgenden werden zunächst zentrale Befunde zusammengefasst und inhaltlich diskutiert. Anschließend wird das methodische Vorgehen bewertet und weitere Implikationen werden dargestellt.

4.3.1 Zusammenfassung und inhaltliche Diskussion zentraler Befunde

Die erste Forschungsfrage wurde mithilfe von Modellvergleichen sowie Analysen struktureller Zusammenhänge untersucht. Zur empirischen Überprüfung der theoretisch angenommenen Kompetenzstruktur wurde das theoretisch postulierte dreidimensionale Modell professioneller Unterrichtswahrnehmung mit den drei Aspekten Beschreiben, Erklären und Vorhersagen gegen ein eindimensionales Modell (professionelle Unterrichtswahrnehmung als eine Gesamtfähigkeit) sowie ein zweidimensionales Modell (Beschreiben und Integrieren) geprüft und den Befunden der Skalierungsstudie (Seidel & Stürmer, in Druck) gegenübergestellt. Wie bei der Skalierungsstudie (Seidel & Stürmer, in Druck) weisen die Skalen aller drei Modelle zufriedenstellende Kennwerte (Reliabilität und Varianz) auf, jedoch fällt auf, dass die Kennwerte in der Scaling-up-Studie durchschnittlich niedriger ausfallen. Bei einer Betrachtung getrennt nach Studierendengruppen (vgl. Kapitel 6, Tabelle 12) wird deutlich, dass die durchschnittlich niedrigeren Skalenkennwerte insbesondere auf die Studierendengruppe Lehramt für Berufliche Schulen zurückzuführen sind. Eine Diskussion dieses Befunds ist in Abschnitt 6.3.1 zu finden.

[15] Dieser Abschnitt basiert zu Teilen auf Jahn et al. (in Druck).

Analog zur Skalierungsstudie (Seidel & Stürmer, in Druck) zeigten Vergleiche des ein-, zwei- und dreidimensionalen Modells bezüglich Skalenkennwerte sowie globalerer Kennwerte für den Modellfit (BIC und Likelihood-Quotienten-Test), dass das dreidimensionale Modell die Daten am besten abbildet. Die detaillierte Analyse der strukturellen Zusammenhänge mithilfe bivariater Korrelationen macht deutlich, dass Beschreiben, Erklären und Vorhersagen aufgrund hoher Korrelationen mit der Gesamtfähigkeit professionelle Unterrichtswahrnehmung keine getrennten Dimensionen, sondern vielmehr Aspekte professioneller Unterrichtswahrnehmung darstellen. Damit ist festzuhalten, dass – wie erwartet – das Tool Observer die theoretisch angenommene Struktur professioneller Unterrichtswahrnehmung als eine Gesamtfähigkeit, differenziert in die drei Aspekte Beschreiben, Erklären und Vorhersagen, in einer großen und – in Bezug auf Universitätsstandorte und Studiengänge – heterogenen Stichprobe von Lehramtsstudierenden empirisch abbildet.

4.3.2 Methodische Überlegungen

Im Rahmen der Überprüfung der ersten Forschungsfrage konnten die theoretisch angenommene Struktur professioneller Unterrichtswahrnehmung an einer großen und heterogenen Stichprobe empirisch abgebildet werden und somit die Befunde der Skalierungsstudie (Seidel & Stürmer, in Druck) erfolgreich repliziert werden. Trotz der Größe und Heterogenität der Stichprobe muss auf die beiden folgenden methodischen Einschränkungen hingewiesen werden: (1) Homogenität der Stichprobe hinsichtlich der absolvierten Studienzeit und (2) gemeinsame Skalierung über alle Lehramtsstudiengänge.

Erstens ist die Scaling-up-Stichprobe zwar in Bezug auf Universitätsstandorte und Lehramtsstudiengänge heterogen, jedoch nicht in Bezug auf die Studiendauer. Alle Studierenden befanden sich zum Zeitpunkt der Datenerhebung in der Anfangsphase ihres Studiums, maximal im dritten Fachsemester. In der Skalierungsstudie (Seidel & Stürmer, in Druck) zeigte sich jedoch bereits, dass das Instrument die Kompetenzstruktur auch in einer Stichprobe mit Studierenden unterschiedlicher Fachsemester (M = 4.64, SD = 3.89) empirisch abbildet.

Zweitens besteht die Scaling-up-Stichprobe zwar aus Studierenden unterschiedlicher Lehramtsstudiengänge, jedoch studieren circa zwei Drittel davon Lehramt für Sekundarstufe, was dem Lehramtsstudiengang aller Studierenden der Skalierungsstudie entspricht. Nachdem die Skalierung über alle Lehramtsstudiengänge hinweg erfolgte, fallen die Studierendengruppen Lehramt Primarstufe und Berufliche Schulen wenig ins Gewicht. Somit bleibt offen, ob das

Tool Observer die Kompetenzstruktur auch in einer Stichprobe mit einer ausgewogenen Verteilung der Studierenden auf die verschiedenen Lehramtsstudiengänge empirisch abbildet. Dieser Frage wird im Rahmen der dritten Forschungsfrage nachgegangen, die auf die Überprüfung des Validitätsaspekts *Generalisierbarkeit* über verschiedene Lehramtsstudiengänge hinweg abzielt (vgl. Kapitel 6). Dazu werden getrennte Einzelskalierungen für Studierende eines Lehramtsstudiengangs durchgeführt.

4.3.3 Implikationen

Trotz der erwähnten methodischen Einschränkungen ergeben sich theoretische und praktische Implikationen aus den Befunden der Überprüfung des Validitätsaspekts *Struktur* im Rahmen der ersten Forschungsfrage. Zum einen ist die empirische Prüfung des Kompetenzstrukturmodells professioneller Unterrichtswahrnehmung in einer großen Stichprobe von theoretischer Relevanz. Damit wird der Forderung, Kompetenzen ausreichend theoretisch zu modellieren (Jude & Klieme, 2008) und empirisch zu überprüfen (Koeppen et al., 2008), nachgekommen. Ein Instrument, das auf einem theoretisch fundierten und empirisch geprüften Kompetenzstrukturmodell beruht, ermöglicht eine detaillierte Untersuchung von Entwicklungsverläufen. Darauf aufbauend können Aussagen über die Wirksamkeit eines Seminars oder Studiengangs getroffen werden. Darüber hinaus können die detaillierten Untersuchungen von Entwicklungsverläufen dazu genutzt werden, um ein Modell der Kompetenzentwicklung zu erarbeiten. Hinsichtlich der Modellierung der Entwicklung von Kompetenzen besteht generell noch erheblicher Forschungsbedarf (Koeppen et al., 2008), obwohl die Art und Weise der Förderung von Kompetenzen im Rahmen der universitären Lehrerbildung ein zentrale Frage darstellt (Frey & Jung, 2011). Für eine effektive Förderung professioneller Unterrichtswahrnehmung könnte ein Kompetenzentwicklungsmodell nützlich sein. Denn falls sich einzelne Aspekte professioneller Unterrichtswahrnehmung unterschiedlich schnell entwickeln, könnten gezielt Lerngelegenheiten für einzelne Aspekte zu unterschiedlichen Zeitpunkten und gegebenenfalls in einer bestimmten Reihenfolge angeboten werden.

Zum anderen erwies sich das Tool Observer als ein in einer Panelstudie ökonomisch handhabbares Instrument, das standortübergreifend eingesetzt werden kann. Damit wurde erste Evidenz dafür geliefert, dass sich das Instrument für einen Einsatz im Large-Scale-Kontext eignet. Es muss allerdings berücksichtigt werden, dass das Instrument unter unterschiedlichen Bedingungen bearbeitet wurde (z. B. Gruppentestung an der Universität oder individuelle

Bearbeitung von zuhause). Deshalb wird im Rahmen der zweiten Forschungsfrage überprüft, inwieweit ein Zusammenhang zwischen unterschiedlichen Erhebungsbedingungen und der Kompetenzerfassung besteht (vgl. Kapitel 5).

5 ÜBERPRÜFUNG DES VALIDITÄTSASPEKTS *GENERALISIERBARKEIT* ÜBER VERSCHIEDENE ERHEBUNGSBEDINGUNGEN HINWEG

Im Rahmen der zweiten Forschungsfrage wird der Validitätsaspekt *Generalisierbarkeit* in den Blick genommen, und zwar spezifisch die *Generalisierbarkeit* über unterschiedliche Erhebungsbedingungen hinweg. In der ersten Teilfrage wird untersucht, inwieweit ein Zusammenhang zwischen den Erhebungsbedingungen Bearbeitungskontext („Online/„On-site“) und Art der Teilnahme („Freiwillig“/„Obligatorisch“) einerseits und der Abbruchquote sowie der Bearbeitungszeit andererseits besteht. In der zweiten Teilfrage wird überprüft, inwieweit die Erhebungsbedingen mit der professionellen Unterrichtswahrnehmung der Lehramtsstudierenden zusammenhängen. Dazu wird zunächst das methodische Vorgehen beschrieben. Anschließend werden die Ergebnisse dargestellt und diskutiert.

5.1 Methodisches Vorgehen[16]

Im Folgenden wird auf die Zusammensetzung der Stichprobe, das Studiendesign, die eingesetzten Messinstrumente sowie die Analyseverfahren zur Datenauswertung eingegangen.

5.1.1 Stichprobe und Studiendesign

Die Untersuchungsstichprobe besteht aus N = 387 Lehramtsstudierenden (66.1 % weiblich) unterschiedlicher Lehramtsstudiengänge und Universitätsstandorte. Diese sind durchschnittlich 22 Jahre alt (M = 22.34, SD = 3.23) und studieren in unterschiedlichen Semestern, wobei sie sich durchschnittlich im vierten Semester (M = 3.96, SD = 2.39) befinden. Tabelle 6 bietet einen Überblick über die deskriptive Statistik der Stichprobe.

[16] Dieser Abschnitt basiert zu Teilen auf Jahn et al. (2011).

Tabelle 6

Deskriptive Statistik der Stichprobe[17]

	Stichprobe (N = 387)	
	M	*SD*
Alter	22.34	3.23
Hochschulsemester	3.96	2.39
	Anteil (in %)	
Geschlecht weiblich	66.1	
Lehramtsstudienrichtung		
Berufliche Schulen	29.5	
Sekundarstufe II	54.0	
Sekundarstufe I	11.2	
Primarstufe	5.4	
Universitätsstandorte		
Standort 1	39.5	
Standort 2	42.4	
PaLea	18.1	

Anmerkung. PaLea = Studierende unterschiedlicher Universitätsstandorte, die am Panel zum Lehramtsstudium (PaLea; Bauer et al., 2010) teilnahmen.

Hinsichtlich des Studiendesigns ist anzumerken, dass insbesondere im Kontext von Datenerhebungen mit Lehramtsstudierenden immer wieder Rekrutierungsprobleme auftreten. Die Durchführung von Studien ist davon abhängig, dass entweder Dozentinnen und Dozenten sich dazu bereit erklären, die Studie in ihre Lehrveranstaltung zu integrieren und zur Voraussetzung für den Erwerb von Leistungspunkten zu machen, oder dass Studierende freiwillig an derartigen Studien teilnehmen. Aufgrund dieser Gegebenheiten im Feld konnte in der vorliegenden Studie keine randomisierte Zuteilung der teilnehmenden Lehramtsstudierenden zu den einzelnen Erhebungsbedingungen realisiert werden. Im Rahmen der Datenerhebung wurden der Bearbeitungskontext und die Art der Teilnahme variiert: Das Tool Observer konnte „Online“ von zuhause oder „On-site“ als Gruppentestung an der Universität bearbeitet werden.

[17] Die einzelnen Universitätsstandorte werden entsprechend dem Vorgehen der anderen drei Forschungsfragen anonymisiert.

Die Bearbeitung erfolgte entweder „Freiwillig“ oder „Obligatorisch“ eingebettet in eine Lehrveranstaltung. Die Datenerhebung fand an zwei Universitätsstandorten im Wintersemester 2009/2010 bzw. im Sommersemester 2010. Die Bearbeitung des Tools Observer war an beiden Standorten in eine Lehrveranstaltung der örtlichen Lehrerbildung integriert („Obligatorisch“) und fand am ersten Standort „Online“ und am zweiten Standort im Rahmen von Gruppentestungen („On-site“) statt. Zusätzlich bestand am zweiten Standort für Studierende des beruflichen Lehramts im zweiten Semester unabhängig von einer Lehrveranstaltung die Möglichkeit, auf freiwilliger Basis („Freiwillig“) an einer Gruppentestung („On-site“) teilzunehmen. Darüber hinaus wurde für die vorliegende Studie eine Teilstichprobe der Scaling-up-Studie genutzt, die am deutschlandweiten Panel zum Lehramtsstudium (PaLea; Bauer et al., 2010) teilnahm und im Sommersemester 2010 das Instrument „Freiwillig“ und „Online“ von zuhause bearbeitete. Die Zusammensetzung der Stichprobe in Bezug auf die verschiedenen Universitätsstandorte ist Tabelle 6 zu entnehmen. Die detaillierte Verteilung der N = 387 Lehramtsstudierenden auf die unterschiedlichen Erhebungsbedingungen ist in Tabelle 7 dargestellt.

Tabelle 7

Verteilung der Stichprobe auf die beiden Erhebungsbedingungen Bearbeitungskontext und Art der Teilnahme[18]

	Bearbeitungskontext		
Art der Teilnahme	Online	On-site	Gesamt
Freiwillig	70	111	181
Obligatorisch	153	53	206
Gesamt	223	164	387

5.1.2 Messinstrumente

Im Folgenden wird darauf eingegangen, wie im Rahmen der zweiten Forschungsfrage nach der *Generalisierbarkeit* über verschiedene Erhebungsbedingungen hinweg (a) die Abbruchquote und die Bearbeitungszeit operationalisiert wurden und wie (b) professionelle Unterrichtswahrnehmung erfasst wurde.

[18] adaptiert nach Jahn et al. (2011).

5.1.2.1 Abbruchquote und Bearbeitungszeit

Das Tool Observer ist in eine Online-Erhebungsplattform integriert. Bei der Bearbeitung des Instruments werden automatisch Logfiles mitabgespeichert. Anhand dieser ist abzulesen, an welcher Stelle die Bearbeitung abgebrochen wurde. Zusätzlich wird die Gesamtbearbeitungszeit in Sekunden automatisch miterfasst. Die gespeicherte Zeit bezieht sich auf den Zeitraum zwischen Aufrufen der ersten Seite und Schließen der letzten Seite des Instruments.

5.1.2.2 Professionelle Unterrichtswahrnehmung

Professionelle Unterrichtswahrnehmung wird mit dem videobasierten Online-Tool Observer (Seidel et al., 2010b) erfasst. Eine detaillierte Beschreibung des Instruments findet sich in Abschnitt 2.2.3.2.4. Aus der Einschätzung von 216 Rating-Items resultiert ein hoch reliabler Gesamtscore für professionelle Unterrichtswahrnehmung (Cronbach's $\alpha = .90$), der der prozentualen Übereinstimmung mit der Expertennorm entspricht und sich zwischen 0 und 1 bewegt.

5.1.3 Auswertungsmethoden

Die Auswertungsmethoden gliedern sich in zwei Teilfragen. Im Rahmen der ersten Teilfrage wurden die Abbruchquote und die Bearbeitungszeit analysiert. Für die Analyse der Abbruchquote wurden in einem ersten Schritt alle gespeicherten Logfiles anhand folgender Kriterien klassifiziert: vollständige Bearbeitung, unvollständige Bearbeitung mit anschließendem erfolgreichen Bearbeitungsversuch und unvollständige Bearbeitung ohne weiteren erfolgreichen Versuch. Zusätzlich wurde eine Ausdifferenzierung der letzten Gruppe bezüglich des Zeitpunkts des Abbruchs vorgenommen. In einem zweiten Schritt wurden die Abbrüche getrennt nach den verschiedenen Erhebungsbedingungen analysiert. Die Bearbeitungszeit bezieht sich auf alle vollständigen Bearbeitungen und wurde zur besseren Verständlichkeit von Sekunden in Minuten umcodiert.

Im Rahmen der zweiten Teilfrage wurde eine zweifaktorielle, univariate Varianzanalyse durchgeführt, um die Effekte der Erhebungsbedingungen auf die Bearbeitungszeit zu überprüfen.

5.2 Ergebnisse[19]

Die zweite Forschungsfrage zielt darauf ab, die *Generalisierbarkeit* über verschiedene Erhebungsbedingungen hinweg zu überprüfen. Die Befunde der Analyse der Abbruchquote und die Effekte der Erhebungsbedingungen auf die Bearbeitungszeit im Rahmen der ersten Teilfrage sowie auf die professionelle Unterrichtswahrnehmung im Rahmen der zweiten Teilfrage werden im Folgenden beschrieben.

5.2.1 Abbruchquote und Bearbeitungszeit

Neben den $N = 387$ Lehramtsstudierenden, die das Instrument im Rahmen dieser Studie bearbeiteten, gab es $n = 53$ Studierende, die das Instrument unvollständig bearbeiteten. Eine detaillierte Analyse dieser Gruppe ist in Abbildung *5* dargestellt. Es zeigt sich, dass von diesen $n = 53$ Studierenden 28.8 % bei einem weiteren Versuch die Bearbeitung erfolgreich beendeten. Damit ergeben sich $n = 38$ Abbrüche, was einer Abbruchquote von 8.6 %. entspricht. 57.9 % der Abbrechenden beendeten die Bearbeitung vor dem ersten Clip, 18.4 % direkt danach und 23.7 % zu einem späteren Zeitpunkt im Instrument.

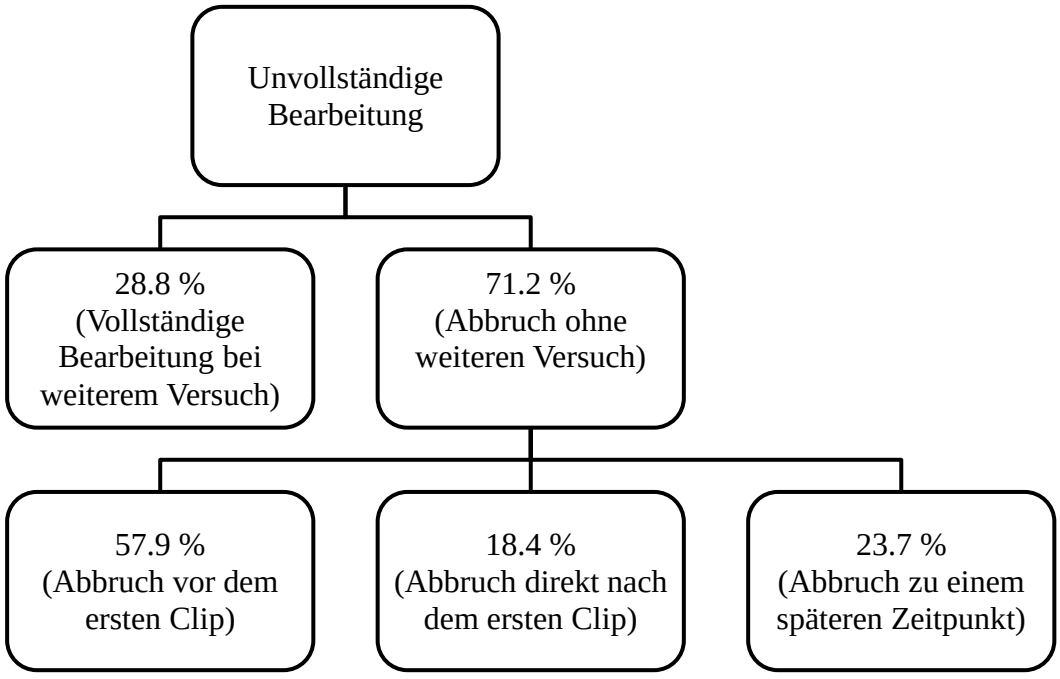

Abbildung 5. Detailanalyse der Abbrüche.[20]

[19] Dieser Abschnitt basiert zu Teilen auf Jahn et al. (2011).
[20] vergleiche Jahn et al. (2011).

Nimmt man die Abbruchquoten getrennt nach den Erhebungsbedingungen Bearbeitungskontext und Art der Teilnahme in den Blick, ergibt sich folgendes Bild (vgl.
Tabelle 8): Unter der Bedingung „Online“ gab es mehr Abbrüche als unter der Bedingung „On-site“. Am häufigsten brachen Studierende in der Kombination „Freiwillig“ und „Online“ ab. Innerhalb der Gruppe der Studierenden, die das Instrument „Obligatorisch“ bearbeiteten, kam es unabhängig von der Erhebungsbedingung zu keinen Abbrüchen. Auf Basis dieses Befunds können technische Schwierigkeiten in der Bearbeitung des Tools Observer weitgehend ausgeschlossen werden.

Tabelle 8

Anzahl der Abbrüche unter den beiden Erhebungsbedingungen Bearbeitungskontext und Art der Teilnahme[21]

	Bearbeitungskontext		
Art der Teilnahme	Online	On-site	Gesamt
Freiwillig	34	4	38
Obligatorisch	0	0	0
Gesamt	34	4	38

Bezüglich des Zusammenhangs der Erhebungsbedingungen und der Bearbeitungszeit zeigt die zweifaktorielle univariate Varianzanalyse folgendes Bild (vgl. Tabelle 9). Der Bearbeitungskontext hat einen signifikanten Effekt auf die Bearbeitungszeit ($F(1, 352) = 53.98$, $p < .001$, $\eta_p^2 = .13$). Studierende unter der Bedingung „Online“ benötigten signifikant länger für die Bearbeitung des Instruments als Studierende unter der Bedingung „On-site“ ($\Delta M = 18.54$, $SD = 2.52$, $p < .001$). Mit einer Effektstärke von 13 % ist dieser Effekt als substantiell einzustufen. Die Art der Teilnahme zeigt keinen signifikanten Effekt auf die Bearbeitungszeit ($F(1, 352) = 3.64$, $p = .06$, $\eta_p^2 = .01$). Allerdings tritt ein Interaktionseffekt zwischen den beiden Erhebungsbedingungen auf ($F(1, 352) = 15.95$, $p < .001$, $\eta_p^2 = .04$). Die höchste Bearbeitungszeit weisen Studierende auf, die das Instrument „Obligatorisch“ und „Online“ bearbeiteten. Unter der Kombination „Obligatorisch“ und „On-site“ benötigten die Studierenden ver-

[21] adaptiert nach Jahn et al. (2011).

hältnismäßig die geringste Zeit. Allerdings ist dieser Interaktionseffekt aufgrund der geringen Effektstärke zu vernachlässigen.

Über alle Erhebungsbedingungen hinweg benötigen die Studierenden durchschnittlich $M = 72.18$ ($SD = 24.97$) Minuten für die Bearbeitung des Instruments. Zusammen mit den Bearbeitungszeiten unter den einzelnen Kombinationen von Erhebungsbedingungen (vgl. Tabelle 9) deutet dies darauf hin, dass das Instrument im Rahmen der vorgesehenen maximalen Bearbeitungszeit von 90 Minuten bearbeitet werden kann.

Tabelle 9

Effekte der beiden Erhebungsbedingungen Bearbeitungskontext und Art der Teilnahme auf die Bearbeitungszeit[22]

		M	*SD*	*F*	*df*	*p*	η_p^2
Bedingung 1				53.98	1, 352	< .001	.13
Online		81.40	29.29				
On-site		61.39	11.63				
Bedingung 2				3.64	1, 352	.06	.01
Freiwillig		66.22	16.68				
Obligatorisch		78.02	29.92				
Bedingung 1 x	Bedingung 2			15.95	1, 352	< .001	.04
Online	Freiwillig	71.56	21.34				
	Obligatorisch	86.45	31.52				
On-site	Freiwillig	63.09	12.28				
	Obligatorisch	57.83	9.27				

[22] vergleiche Jahn et al. (2011).

Zusammenfassend ergibt sich folgendes Bild: Die Abbruchquote beträgt 8.6 %, wobei ausschließlich bei freiwilliger Bearbeitung und vor allem im Bearbeitungskontext „Online“ abgebrochen wurde. Im Hinblick auf die Bearbeitungszeit zeigt sich ein signifikanter, substantieller Effekt des Bearbeitungskontexts dahingehend, dass die Bearbeitungszeiten „Online“ signifikant länger sind.

5.2.2 Professionelle Unterrichtswahrnehmung

Zur Überprüfung des Zusammenhangs zwischen verschiedenen Erhebungsbedingungen und professioneller Unterrichtswahrnehmung wurde eine zweifaktorielle, univariate Varianzanalyse berechnet. Es zeigen sich weder signifikante Effekte des Bearbeitungskontextes ($F(1, 382) = 0.38$, $p = .54$, $\eta_p^2 < .01$) noch der Art der Teilnahme ($F(1, 382) = 1.69$, $p = .19$, $\eta_p^2 < .01$) auf die professionelle Unterrichtswahrnehmung von Lehramtsstudierenden. Unabhängig davon, ob die Studierenden das Instrument „Online“ oder „On-site“ beziehungsweise „Freiwillig“ oder „Obligatorisch“ bearbeiteten, bewegen sie sich hinsichtlich der Übereinstimmung mit der Expertennorm von 35 – 38 % auf einem niedrigem bis mittleren Fähigkeitsniveau. Allerdings gibt es einen signifikanten Interaktionseffekt zwischen den Erhebungsbedingungen ($F(1, 382) = 14.09$, $p = < .001$, $\eta_p^2 = .04$). Studierende, die das Tool Observer „Freiwillig“ und „Online“ oder „Obligatorisch“ und „On-site“ bearbeiteten, weisen eine höhere professionelle Unterrichtswahrnehmung auf als Studierende der anderen beiden Erhebungskombinationen. Jedoch ist die Stärke des Interaktionseffekts als gering einzustufen. Die Ergebnisse der zweifaktoriellen, univariaten Varianzanalyse sind zusammen mit den Mittelwerten und Standardabweichungen professioneller Unterrichtswahrnehmung in den verschiedenen Erhebungsbedingungen in Tabelle 10 dargestellt.

Zusammenfassend ist festzuhalten, dass die Erhebungsbedingungen Bearbeitungskontext und Art der Teilnahme keine signifikanten Haupteffekte auf die Erfassung professioneller Unterrichtswahrnehmung haben.

Tabelle 10

Effekte der Erhebungsbedingungen Bearbeitungskontext und Art der Teilnahme auf die professionelle Unterrichtswahrnehmung der Lehramtsstudierenden[23]

			M	*SD*	*F*	*df*	*p*	η_p^2
Bedingung 1					.38	1, 382	.54	< .01
Online			.35	.15				
On-site			.37	.16				
Bedingung 2					1.69	1, 382	.19	< .01
Freiwillig			.38	.16				
Obligatorisch			.35	.15				
Bedingung 1	x	Bedingung 2			14.09	1, 382	< .001	.04
Online		Freiwillig	.41	.15				
		Obligatorisch	.33	.14				
On-site		Freiwillig	.36	.16				
		Obligatorisch	.40	.16				

[23] vergleiche Jahn et al. (2011).

5.2.3 Zusammenfassung der Ergebnisse

Die Anzahl der Abbrüche variiert über die Erhebungsbedingungen Bearbeitungskontext und Art der Teilnahme hinweg in die erwartete Richtung: Die höchste Abbruchquote findet sich in der Kombination der Bedingungen „Freiwillig“ und „Online“. Gemäß der Hypothese hat der Bearbeitungskontext einen signifikanten, substantiellen Haupteffekt auf die Bearbeitungszeit (längere Bearbeitungszeit unter der Bedingung „Online“), aber nicht die Art der Teilnahme. Demzufolge besteht ein Zusammenhang zwischen beiden Erhebungsbedingungen und der Abbruchquote sowie zwischen dem Bearbeitungskontext und der Bearbeitungszeit.

Hypothesenkonform zeigen sich für keine der beiden Erhebungsbedingungen signifikante Haupteffekte auf die Erfassung professioneller Unterrichtswahrnehmung durch das Instrument. Folglich besteht kein Zusammenhang zwischen den Erhebungsbedingungen Bearbeitungskontext sowie Art der Teilnahme und der professionellen Unterrichtswahrnehmung der Lehramtsstudierenden.

5.3 Diskussion[24]

Vor dem Hintergrund eines großflächigen Einsatzes des Tools Observer wurde das Instrument aus ökonomischen Gesichtspunkten in eine Online-Plattform integriert, um eine internetbasierte Bearbeitung zu ermöglichen. Die Bearbeitung des Tools wird teilweise nicht nur individuell von zuhause, sondern auch im Rahmen von Lehrveranstaltungen als verpflichtende Gruppentestung durchgeführt, sei es aufgrund von Rekrutierungsproblemen oder um die Lehrveranstaltung zu evaluieren. Folglich wurde im Rahmen der zweiten Forschungsfrage der Validitätsaspekt der *Generalisierbarkeit* über verschiedene Erhebungsbedingungen hinweg überprüft. An einer Stichprobe von insgesamt N = 387 Lehramtsstudierenden wurden zwei Teilfragen untersucht. In der ersten Teilfrage wurde überprüft, inwieweit ein Zusammenhang zwischen den Erhebungsbedingungen Bearbeitungskontext („Online/„On-site“) und Art der Teilnahme („Freiwillig“/„Obligatorisch“) einerseits und der Abbruchquote sowie der Bearbeitungszeit andererseits besteht. In der zweiten Teilfrage wurde überprüft, inwieweit die Erhebungsbedingen mit der professionellen Unterrichtswahrnehmung der Lehramtsstudierenden zusammenhängen. Im Folgenden werden zunächst zentrale Befunde zusammengefasst und inhaltlich diskutiert. Anschließend wird das methodische Vorgehen bewertet und weitere Implikationen werden dargestellt.

5.3.1 Zusammenfassung und inhaltliche Diskussion zentraler Befunde

Im Rahmen der ersten Teilfrage wurde der Zusammenhang zwischen den verschiedenen Erhebungsbedingungen und der Abbruchquote sowie der Bearbeitungszeit untersucht. Die Anzahl der Abbrüche variiert über die Erhebungsbedingungen Bearbeitungskontext und Art der Teilnahme hinweg wie erwartet: Die höchste Abbruchquote findet sich in der Kombination der Bedingungen „Freiwillig“ und „Online“. Die Verteilung der Abbrüche entspricht den Befunden in der Medienforschung (Birnbaum, 2004; Zumbach & Reimann, 2001). Eine Abbruchquote von insgesamt 8.6 % erscheint bei einer Erhebung dieser Größenordnung akzeptabel. Damit kann die Gefahr einer Vorselektion bzw. Verzerrung der Stichprobe und damit der Validität der Interpretation der Messergebnisse bedingt durch eine hohe Zahl an Abbrüchen als gering eingeschätzt werden (vgl. Birnbaum, 2004; Treiblmaier, 2010). Hinsichtlich der Bearbeitungszeit zeigt eine zweifaktorielle, univariate Varianzanalyse wie erwartet einen

[24] Dieser Abschnitt basiert zu Teilen auf Jahn et al. (2011).

signifikanten substantiellen Haupteffekt für den Bearbeitungskontext, jedoch nicht für die Art der Teilnahme. Unter der Bedingung „Online“ benötigten die Studierenden durchschnittlich mehr Zeit. Die längeren Bearbeitungszeiten deuten auf eine höhere Selbstregulation durch die Studierenden hin, die sich in der Entscheidung bezüglich der Bearbeitungsdauer zeigt. Beispielsweise bietet das Instrument die Möglichkeit, die Videoclips vor der Einschätzung der Rating-Items mehrfach anzusehen. Eventuell wird von dieser Möglichkeit in einer Gruppentestung, in der sich die Studierenden aneinander orientieren, weniger Gebrauch gemacht.

Für die zweite Teilfrage wurde der Zusammenhang zwischen den verschiedenen Erhebungsbedingungen und der Erfassung professioneller Unterrichtswahrnehmung mithilfe einer zweifaktoriellen, univariaten Varianzanalyse geprüft. Es zeigen sich keine signifikanten Haupteffekte der einzelnen Erhebungsbedingungen auf die Kompetenzmessung. Jedoch liegt ein schwacher Interaktionseffekt vor: Studierende unter der Kombination der Bedingungen „Freiwillig“ und „Online“ sowie „Obligatorisch“ und „On-site“ schnitten besser ab als Studierende unter anderen Bedingungskombinationen. Dieser Interaktionseffekt ist zwar aufgrund der geringen Effektstärke in dieser Studie zu vernachlässigen, sollte aber in einer weiteren Studie überprüft werden. Denn zum einen könnte es sein, dass sich der wahrgenommene Grad an Autonomie während der Bearbeitung positiv auf die Motivation und die Qualität von Wissensprozessen und damit auf die Bearbeitung des Instruments auswirkt (vgl. Prenzel et al., 2004). Wahrscheinlich nehmen die Studierenden in der Kombination der Bedingungen „Freiwillig“ und „Online“ ohne den kontrollierten Rahmen einer Gruppentestung den höchsten Grad an Autonomie wahr. Zum anderen ist es aber auch möglich, dass die vergleichsweise hohen Abbruchquoten in dieser Bedingungskombination doch zu einer gewissen Vorselektion führen. Eventuell beenden vor allem Studierende die Bearbeitung, die ein gewisses Interesse und Vorerfahrungen im Bereich pädagogisch-psychologischer Kompetenzen mitbringen. Darüber hinaus wäre es denkbar, dass das gute Abschneiden der Studierenden unter der Kombination der Bedingungen „Obligatorisch“ und „On-site“ mit dem kontrollierten Rahmen einer Gruppentestung zusammenhängt. Zwar senkt der hohe Verpflichtungsgrad die Motivation, jedoch kann dies durch die Rahmenbedingungen einer Gruppentestung, in der es einen vorgesehenen Zeitrahmen und kaum Ablenkungen oder Möglichkeiten zur Fremdbeschäftigung gibt, ausgeglichen werden.

5.3.2 Methodische Überlegungen

Im Rahmen der Überprüfung der zweiten Forschungsfrage zeigten sich hypothesenkonforme Ergebnisse. Es gibt zwar signifikante, substantielle Zusammenhänge zwischen den verschiedenen Erhebungsbedingungen und der Abbruchquote sowie der Bearbeitungszeit, jedoch nicht mit Blick auf die Erfassung professioneller Unterrichtswahrnehmung. Allerdings müssen folgende Einschränkungen der Stichprobe berücksichtigt werden, die dafür sprechen, die Befunde als erste Hinweise zu interpretieren, die weiterer Prüfung bedürfen: (1) keine randomisierte Zuteilung zu den Erhebungsbedingungen und (2) Koppelung von Erhebungsbedingung und Hochschulstandort.

Erstens ist einschränkend anzuführen, dass die Zuteilung der Studierenden zu den einzelnen Erhebungsbedingungen nicht randomisiert stattfand. Dadurch ist die Vergleichbarkeit der einzelnen Gruppen nicht garantiert (Field & Hole, 2003). Zweitens muss darauf hingewiesen werden, dass in der Untersuchungsstichprobe Erhebungsbedingung und Hochschulstandort konfundiert sind. Daher erlaubt es das vorliegende Design nicht, mögliche Effekte der Hochschulstandorte systematisch zu prüfen. Jedoch gilt es zu bedenken, dass im Bereich der Forschung zu Entwicklungen von Lehramtsstudierenden bislang keine Befunde zur Berücksichtigung von Bedingungsfaktoren und zu Unterschieden in den individuellen Voraussetzungen der Studierenden (z. B. Motivation) abhängig vom Hochschulstandort vorliegen. Deshalb wird vermutet, dass keine systematischen Verzerrungen der Stichprobe im Hinblick auf individuelle Voraussetzungen der Studierenden vorliegen. Dennoch sollten die Befunde der vorliegenden Arbeit an einer Stichprobe mit randomisierter Zuteilung, in der Erhebungsbedingung und Hochschulstandort nicht konfundiert sind, überprüft werden.

Darüber hinaus bleibt offen, inwieweit die Struktur professioneller Unterrichtswahrnehmung über die einzelnen Erhebungsbedingungen hinweg vergleichbar durch das Instrument abgebildet wird. Aufgrund der geringen Stichprobengröße pro Kombination von Erhebungsbedingung waren getrennte Einzelskalierungen nicht möglich. Dies sollte in einer weiteren Studie mit mindestens $N = 150$ Lehramtsstudierenden pro Bedingung (vgl. Linacre, 1994; Wright, 1996) umgesetzt werden.

5.3.3 Implikationen

Trotz der diskutierten methodischen Einschränkungen wurde im Rahmen der Überprüfung der zweiten Forschungsfrage Evidenz dafür gefunden, dass die Messergebnisse des Tools Observer über verschiedene Erhebungsbedingungen hinweg valide interpretiert werden können. Damit wird der Forderung entsprochen, die Effekte verschiedener Bedingungen von Online-Bearbeitungen zu untersuchen (Jurecka, 2008). Praktisch gesehen bedeutet das, dass mit dem Tool Observer ein Instrument vorliegt, das stabil über verschiedene Erhebungsbedingungen hinweg eingesetzt werden kann. Durch die Möglichkeit der Online-Bearbeitung von zuhause können in kurzer Zeit viele Personen das Instrument bearbeiten. Darüber hinaus kann durch die Einbettung der Erhebung als obligatorische Gruppentestung im Rahmen von Lehrveranstaltungen eine hohe Teilnehmerzahl gesichert und damit das Problem der Rekrutierung minimiert werden. Diese beiden Aspekte stellen wichtige Voraussetzungen für den Einsatz im Large-Scale-Kontext dar. Dadurch eröffnet sich ein breites Spektrum an Einsatzmöglichkeiten, angefangen bei der Überprüfung der Wirksamkeit von Lehrveranstaltungen bis hin zum Einsatz als Selbstexplorationsverfahren im Rahmen der Studienwahl.

6 ÜBERPRÜFUNG DES VALIDITÄTSASPEKTS *GENERALISIERBARKEIT* ÜBER UNTERSCHIEDLICHE LEHRAMTSSTUDIENGÄNGE HINWEG

Die dritte Forschungsfrage fokussiert erneut den Validitätsaspekt *Generalisierbarkeit*, aber spezifisch die *Generalisierbarkeit* über unterschiedliche Lehramtsstudiengänge hinweg. Es wird untersucht, inwieweit die Struktur professioneller Unterrichtswahrnehmung sowie die Itemschwierigkeiten für unterschiedliche Studierendengruppen (Lehramt für Primarstufe, Sekundarstufe und Berufliche Schulen) vergleichbar sind. Dazu wird zunächst das methodische Vorgehen beschrieben. Anschließend werden die Ergebnisse dargestellt und diskutiert.

6.1 Methodisches Vorgehen

Im Folgenden wird auf die Zusammensetzung der Stichprobe, das Messinstrument sowie die Auswertungsmethode eingegangen.

6.1.1 Stichprobe

Zur Überprüfung der dritten Forschungsfrage wurde die Scaling-up-Stichprobe der ersten Forschungsfrage mit $N = 1029$ Lehramtsstudierenden nach Lehramtsstudiengängen aufgeteilt. Es wurde eine Unterteilung in drei Studierendengruppen vorgenommen: Lehramt für Primarstufe ($n = 166$), Lehramt für Sekundarstufe ($n = 671$) und Lehramt für Berufliche Schulen ($n = 171$). Die restlichen $n = 21$ Studierenden gaben an, ein anderes Lehramt zu studieren und wurden aus den weiteren Analysen ausgeschlossen. Ein Überblick über die Zusammensetzung der einzelnen Studierendengruppen bezüglich Alter, Semester und Geschlecht sowie die Verteilung auf deren 16 Universitätsstandorte sind Tabelle 11 zu entnehmen.

Die drei Studierendengruppen wurden mittels Chi-Quadrat-Tests und ANOVAs verglichen Die Studierenden unterscheiden sich signifikant hinsichtlich der Geschlechterverteilung ($\chi^2(2) = 27.96$, $p < .001$, Cramer's $\nu = .17$). Die Gruppe der Studierenden Lehramt Primarstufe weist mit 77.7 % den höchsten Frauenanteil auf. Dieser Anteil war jedoch zu erwarten, zieht man als Referenz beispielsweise die Zahlen des statistischen Bundesamts heran, nach dem im Schuljahr 2012/13 88.2 % aller Grundschullehrpersonen weiblich waren (Statistisches Bundesamt, 2013). Auch bezüglich des Alters ($F(2,1003) = 39.10$, $p < .001$, $\eta^2 = .07$) zeigen sich signifikante Unterschiede zwischen den Studierendengruppen. Die Gruppe der Studierenden

Lehramt Berufliche Schulen sind mit durchschnittlich $M = 24.38$ ($SD = 3.76$) Jahren erwartungsgemäß signifikant älter ($p = .02$) als die Studierenden der anderen beiden Gruppen. Dieser Befund kann darauf zurückgeführt werden, dass viele Studierende dieser Gruppe bereits eine Berufsausbildung abgeschlossen haben. Bezüglich der Semesteranzahl liegen für das Zweitfach ($F(2,911) = 0.05$, $p = .95$, $\eta^2 = .00$) keine signifikanten Unterschiede zwischen den Studierendengruppen vor, aber für das Erstfach ($F(2,989) = 11.90$, $p < .001$, $\eta^2 = .02$). Allerdings kann dieser Unterschied aufgrund der geringen Effektstärke vernachlässigt werden. Die Studierenden der verschiedenen Lehramtsstudiengänge verteilen sich signifikant unterschiedlich auf die einzelnen Universitätsstandorte ($\chi^2(30) = 599.25$, $p < .001$, Cramer's $\nu = .55$). Pro Studierendengruppe sind nicht alle 16 Universitätsstandorte vertreten, da einige Lehramtsstudiengänge an einigen Universitäten nicht angeboten werden bzw. die Teilnehmerzahl in der kompletten Scaling-up-Stichprobe bereits weniger als 1 % beträgt. Die Studierenden Lehramt Primarstufe und Sekundarstufe verteilen sich vergleichsweise ausgewogen auf unterschiedliche Standorte, wohingegen in der Gruppe Lehramt Berufliche Schule 79.6 % aus den Standorten G und L stammen.

Tabelle 11

Deskriptive Statistik der Studierendengruppen verschiedener Lehramtsstudiengänge

		Primarstufe (*N* = 166)		Sekundarstufe (*N* = 671)		Berufliche Schulen (*N* = 171)	
		M	*SD*	*M*	*SD*	*M*	*SD*
Alter		22.70	3.79	22.04	2.82	24.38	3.76
Semester Erstfach		2.68	0.72	2.71	0.85	2.38	0.57
Semester Zweitfach		2.54	0.76	2.55	0.86	2.52	0.72
		Anteil (in %)					
Geschlecht weiblich		77.7		55.9		64.9	
Universität	A	18.1		9.8		/	
	B	12.0		13.6		2.3	
	C	13.3		11.2		1.2	
	D	0.6		/		4.1	
	E	24.1		3.3		3.5	
	F	/		6.6		5.3	
	G	4.8		4.2		17.0	
	H	0.6		16.4		1.8	
	I	0.6		1.2		0.6	
	J	/		4.3		/	
	K	2.4		0.1		0.6	
	L	/		10.6		62.6	
	M	19.9		12.5		0.6	
	N	/		1.9		0.6	
	O	3.0		4.0		/	
	P	0.6		1.5		1.2	

6.1.2 Messinstrument

Professionelle Unterrichtswahrnehmung wird mit dem videobasierten Online-Tool Observer (Seidel et al., 2010b) erfasst. Eine detaillierte Beschreibung des Instruments findet sich in Abschnitt 2.2.3.2.4. Aus der Einschätzung von insgesamt 216 Rating-Items resultieren reliable Scores für professionelle Unterrichtswahrnehmung (Cronbach's $\alpha = .91$) und die drei Aspekte Beschreiben (Cronbach's $\alpha = .75$), Erklären (Cronbach's $\alpha = .81$) und Vorhersagen (Cronbach's $\alpha = .81$). Diese entsprechen jeweils der prozentualen Übereinstimmung mit der Expertennorm und bewegen sich zwischen 0 und 1.

6.1.3 Auswertungsmethoden

Ein wichtiger Aspekt bei der Überprüfung der Validität der Interpretation und Nutzung von Messwerten, speziell im Hinblick auf die Untersuchung von Unterschieden verschiedener Personengruppen, ist neben der Vergleichbarkeit der Kompetenzstruktur die Analyse von Differential Item Functioning (DIF) (Penfield & Camilli, 2007). Deshalb wird zunächst überblicksartig die theoretische Konzeption von DIF dargestellt und die Entdeckung von Items mit DIF mithilfe von IRT-basierten Verfahren kurz erläutert. Anschließend werden die Analyseverfahren zum Vergleich der Kompetenzstruktur und der Itemschwierigkeiten beschrieben.

6.1.3.1 Theoretische Grundlagen der Auswertungsmethode

Differential Item Functioning liegt, ganz grundlegend gesprochen, dann vor, wenn ein Item für verschiedene Personengruppen unterschiedlich „funktioniert" (Penfield & Camilli, 2007). Im Kontext eines Rasch-Modells auf Grundlage von dichotomen Daten bedeutet das, dass sich die bedingte Lösungswahrscheinlichkeit eines Items, gegeben derselben Personenfähigkeit, zwischen zwei Gruppen von Personen unterscheidet (de Ayala, 2009). Damit ist die Invarianz von Itemparametern über verschiedene Personengruppen hinweg nicht gegeben (Bond & Fox, 2007). Die Annahme von Invarianz für Personen- und Itemparameter ergibt sich jedoch aus dem Rasch-Kriterium der spezifischen Objektivität (Embretson & Reise, 2000). Demnach sind sowohl Personenfähigkeiten unabhängig von den bearbeiteten Items als auch Itemschwierigkeiten unabhängig davon, welche Personen sie bearbeiten (Fischer, 1995). Dies bedeutet, dass die Lösungswahrscheinlichkeit eines Items in unterschiedlichen Personengruppen vollständig durch Personenfähigkeit und Itemschwierigkeit erklärt wird und somit bei

diesem Item kein DIF vorliegt, der bestimmte Personengruppen benachteiligt oder begünstigt (vgl. Adams & Carstensen, 2002). Die Existenz von substantiellem DIF stellt damit eine Verletzung der Modellannahmen des Rasch-Modells dar (Angoff, 1993).

Bei der Aufdeckung von Items mit DIF und der Interpretation der Effektstärke bieten IRT-basierte Verfahren gegenüber Verfahren basierend auf der klassischen Testtheorie mehrere Vorteile, unter anderem, dass Gruppenunterschiede in den Personenfähigkeiten kontrolliert werden (Camilli & Shepard, 1994). Es existieren verschiedene Softwareprogramme, mit denen derartige DIF-Analysen durchgeführt werden können. Zur Beurteilung des Auftretens von DIF bei einzelnen Items sollte neben einem statistischen Signifikanztest jedoch immer auch die Effektstärke berücksichtigt werden (Monahan, McHorney, Stump & Perkins, 2007), da das statistische Analyseverfahren abhängig von der Stichprobengröße ist (Wang, 2000). Ein statistischer Signifikanztest ist besonders bei kleinen Stichproben von Bedeutung, um zufällige Unterschiede in den Itemschwierigkeiten nicht als substantiellen DIF zu interpretieren (Monahan et al., 2007). Genauso ist die Interpretation der Effektstärke bei großen Stichproben unerlässlich, um vernachlässigbare Unterschiede in den Itemschwierigkeiten nicht als substantiellen DIF zu interpretieren (Monahan et al., 2007). Bei einer ausreichend großen Stichprobe weisen die Mehrzahl der Items DIF auf (Wu & Adams, 2007).

Zur Interpretation der Effektstärke können unterschiedliche Maße herangezogen werden. Die verbreitetsten Effektgrößen basieren entweder auf der Differenz der Lösungswahrscheinlichkeit, auf dem Verhältnis der erklärten Varianz oder auf dem „Log of the Odds-Ratio“ (DeMars, 2011). Für die vorliegende Arbeit sind Effektgrößen basierend auf dem „Log of the Odds-Ratio“ relevant. Als „Odds“ oder Wettchancen wird dabei das Verhältnis aus Lösungswahrscheinlichkeit und Gegenwahrscheinlichkeit bezeichnet. Die Grundlage für diese Effektgrößen bildet der Logarithmus des Verhältnisses der Wettchancen, auch Logit-Werte genannt. Zu dieser Kategorie von Effektgrößen zählt neben der Mantel-Haenszel-Statistik und dem Haupteffekt der Gruppenzugehörigkeit bei der Logistischen Regression auch die Transformation der Differenz der Itemschwierigkeit bei IRT-Modellen (DeMars, 2011). Nachdem in der vorliegenden Arbeit die statistischen DIF-Analysen auf IRT-Modellen basieren, wurde die Differenz der Itemschwierigkeiten als Effektgröße gewählt.

Für die Interpretation der Effektstärke dichotomer Items hat der Educational Testing Service (ETS) eine Klassifikation aufgestellt, nach der DIF in vernachlässigbar, moderat und groß eingeteilt wird (Zieky, 1993). Auf Ebene des Instruments kann von substantiellen DIF-Effekten gesprochen werden, wenn 25 % oder mehr aller Items im Instrument moderaten oder

großen DIF aufweisen (Penfield & Algina, 2006). In diesem Fall ist die Validität des Instruments für die untersuchten Subgruppen eingeschränkt. Unterschiede in den Messwerten zwischen diesen Personengruppen können nicht eindeutig auf die Fähigkeiten der Personen zurückgeführt werden, da die Itemschwierigkeiten für die unterschiedlichen Personengruppen nicht vergleichbar sind. Dies unterstreicht, dass die Validität eines Tests nicht generalisierend Geltung hat, sondern immer im Kontext von spezifischen Personengruppen betrachtet werden muss (Angoff, 1993). Sowohl Items mit negativem DIF (Benachteiligung) als auch Items mit positiven DIF (Bevorzugung) sollten überarbeitet oder aus dem Messinstrument ausgeschlossen (Angoff, 1993) und die statistischen DIF-Analysen wiederholt werden. Statistische DIF-Analysen können demnach im Rahmen der Testentwicklung dazu genutzt werden, die Items eines Messinstruments zu verbessern (O'Neill & McPeek, 1993) oder das Einsatzgebiet eines Messinstruments auf spezifische Personengruppen zu erweitern. Allerdings liefern die statistischen DIF-Analysen keine Hinweise auf Ursachen für DIF (Penfield & Camilli, 2007).

6.1.3.2 Analyseverfahren[25]

Die dritte Forschungsfrage wurde mittels zweier Analysen untersucht. Zur Überprüfung der Vergleichbarkeit der Kompetenzstruktur professioneller Unterrichtswahrnehmung wurden auf Basis von $n = 106$ Items, die dem finalen Itempool der Skalierung im Rahmen der ersten Forschungsfrage entsprechen (vgl. Abschnitt 4.1.3.2), für alle drei Studierendengruppen getrennte Einzelskalierungen vorgenommen. Dabei wurde analog zum methodischen Vorgehen der ersten Forschungsfrage (vgl. Abschnitt 4.1.3.2) vorgegangen. Innerhalb jeder Studierendengruppe wurde ein ein-, zwei- und dreidimensionales Rasch-Modell berechnet. Gegen das theoretisch postulierte dreidimensionale Modell professioneller Unterrichtswahrnehmung (Beschreiben, Erklären und Vorhersagen) wurden ein zweidimensionale Modell (Beschreiben und Integrieren) und ein eindimensionales Modell (professionelle Unterrichtswahrnehmung als Gesamtfähigkeit) geprüft. Für diese Modellprüfung wurden Skalenindizes (EAP/PV-Reliabilitäten und erklärte Varianz) sowie Likelihood-Quotienten-Tests (basierend auf der Final Deviance) und BICs berücksichtigt. Die Modellvergleiche innerhalb der drei Studierendengruppen wurden abschließend gegenübergestellt.

Zur Überprüfung der Vergleichbarkeit der Itemschwierigkeiten wurden mit Blick auf die drei Studierendengruppen (Lehramt für Primarstufe, Sekundarstufe und Berufliche Schulen) DIF-

[25] Dieser Abschnitt basiert zu Teilen auf Jahn et al. (in Druck).

Analysen basierend auf dem finalen Itempool der Scaling-up-Studie (n = 106 Items) durchgeführt. Mithilfe der Software ConQuest (Wu et al., 1998) wurde eine Skalierung durch ein Multifacetten-Modell vorgenommen, indem in das Standardmodell ein zusätzlicher Interaktionsterm „Item x Facette“ integriert wurde. Als Facette fungiert dabei die kategoriale Gruppenvariable „Lehramtsstudiengang“. Es wurden Gruppenunterschiede in den Personenfähigkeiten kontrolliert sowie der mittlere DIF über alle Items auf 0 festgelegt. Als Referenzgruppe dient die Gruppe der Studierenden Lehramt Sekundarstufe, da das Instrument Videoclips aus der Sekundarstufe enthält und im bisherigen Projektverlauf für diese Zielgruppe erprobt wurde. Zur Interpretation der Effektstärke wird die Differenz der Itemschwierigkeiten herangezogen. Diese Logit-skalierte Effektgröße wurde analog zum ETS-Klassifikationsschema umgerechnet (DeMars, 2011). Dementsprechend wird für dichotome Items die Höhe von DIF für die Logit-Werte wie folgt in vernachlässigbar, moderat und groß klassifiziert: vernachlässigbar bei $|DIF| < 0.43$ oder nicht signifikant > 0, moderat bei $0.43 \leq |DIF| \leq 0.64$ und $|DIF|$ signifikant > 0, groß bei $|DIF| > 0.64$ und signifikant > 0 (vgl. z. B. Penfield & Camilli, 2007). Wenn 25 % oder mehr aller Items eines Messinstruments moderaten oder großen DIF aufweisen, deutet dies auf substantielle DIF-Effekte auf Ebene des Messinstruments hin (Penfield & Algina, 2006).

Der Umfang fehlender Werte ist mit durchschnittlich 0.5 % zu vernachlässigen.

6.2 Ergebnisse[26]

Im Rahmen der dritten Forschungsfrage wurde im Hinblick auf die Überprüfung der *Generalisierbarkeit* über verschiedene Lehramtsstudiengänge hinweg untersucht, inwieweit die Struktur professioneller Unterrichtswahrnehmung sowie die Itemschwierigkeiten für unterschiedliche Studierendengruppen vergleichbar sind. Dazu wurde in den drei Studierendengruppen (Lehramt für Primarstufe, Sekundarstufe und Berufliche Schulen) einerseits die Kompetenzstruktur verglichen, andererseits wurden die Itemschwierigkeiten auf DIF hin untersucht. Die Befunde der Vergleiche der Kompetenzstruktur sowie der Itemschwierigkeiten sind im Folgenden dargestellt.

6.2.1 Vergleich der Struktur professioneller Unterrichtswahrnehmung

Basierend auf den Einzelskalierungen wurden zum Vergleich der Struktur professioneller Unterrichtswahrnehmung, in einem ersten Schritt die Skalenkennwerte des ein-, zwei- und dreidimensionalen Rasch-Modells innerhalb jeder Studierendengruppe (Lehramt Primarstufe, Sekundarstufe und Berufliche Schulen) verglichen. Die Schätzungen für EAP/PV-Reliabilitäten und Varianzen für die drei Studierendengruppen, die jeweils auf unidimensionalen Schätzungen beruhen, sind in Tabelle 12 dargestellt. In allen Studierendengruppen zeigt das eindimensionale Modell stets die höchste EAP/PV-Reliabilität ($.93 \leq r \leq .95$), wobei alle Modelle durchgehend hohe EAP/PV-Reliabilitäten ($r \geq .78$) erreichen. Die höchste Varianz als Indikator für Trennschärfe der Items (vgl. Leutner et al., 2012) weist dagegen in allen Lehramtsstudiengängen das dreidimensionale Modell auf. Allerdings fällt beim Vergleich der drei Studierendengruppen auf, dass die Skalenkennwerte der Gruppe Lehramt Berufliche Schulen in der Tendenz geringer ausfallen als die der Gruppen Lehramt Primarstufe und Sekundarstufe.

In einem zweiten Schritt wurden das ein- zwei- und dreidimensionale Rasch-Modell mithilfe von BICs und Likelihood-Quotienten-Tests verglichen. Wie Tabelle 13 zu entnehmen ist, fällt der Modellvergleich in allen drei Studierendengruppen (Lehramt Primarstufe, Sekundarstufe und Berufliche Schulen) durchgehend zugunsten des dreidimensionalen Modells aus.

[26] Dieser Abschnitt basiert zu Teilen auf Jahn et al. (in Druck).

Tabelle 12

Skalenkennwerte des ein-, zwei- und dreidimensionalen Modells professioneller Unterrichtswahrnehmung für die drei Studierendengruppen Lehramt Primarstufe, Sekundarstufe und Berufliche Schulen[27]

	1-dim	2-dim Modell	3-dim	1-dim	2-dim Modell	3-dim	1-dim	2-dim Modell	3-dim
	Primarstufe (n = 166)			*Sekundarstufe (n = 671)*			*Berufliche Schulen (n = 171)*		
Reliabilität	.95			.93			.93		
Beschreiben		.85	.85		.80	.80		.80	.80
Erklären			.84			.84			.78
Vorhersagen		.94	.89		.93	.88		.91	.85
Varianz	1.19			0.94			0.78		
Beschreiben		0.89	0.89		0.61	0.60		0.62	0.62
Erklären			1.09			1.00			0.68
Vorhersagen		1.54	2.31		1.36	2.02		0.94	1.39

Anmerkung. EAP/PV-Reliabilitäten und Varianzen basieren auf unidimensionalen Schätzern.

[27] adaptiert nach Jahn et al. (in Druck).

Tabelle 13

Modellvergleich des ein-, zwei- und dreidimensionalen Modells professioneller Unterrichtswahrnehmung in den drei Studierendengruppen Lehramt Primarstufe, Sekundarstufe und Berufliches Lehramt[28]

	Deviance	Parameter	ΔDeviance (*df*)	BIC
Primarstufe (n =166)				
dreidimensional	18018	112	-	18587
zweidimensional	18071	109	53** (3)	18625
eindimensional	18146	107	128** (5)	18690
Sekundarstufe (n = 671)				
dreidimensional	78698	112	-	79427
zweidimensional	78899	109	201** (3)	79608
eindimensional	79457	107	759** (5)	80153
Berufliche Schulen (n = 171)				
dreidimensional	19938	112	-	20514
zweidimensional	19974	109	36** (3)	20534
eindimensional	20000	107	62** (5)	20550

Anmerkungen. Baseline für den Modellvergleich ist das dreidimensionale Modell. *Deviance:* -2log[LR] der Modellschätzung. Δ*Deviance:* χ^2-verteilte Teststatistik des Likelihood-Quotienten-Tests, Differenz der geschätzten Parameter als Freiheitsgrade *(df).* ** $p < .01$.

Zusammenfassend ist zu konstatieren, dass die Prüfung der Struktur professioneller Unterrichtswahrnehmung für alle drei Studierendengruppen (Lehramt Primarstufe, Sekundarstufe und Berufliche Schulen) zu vergleichbaren Ergebnissen führt. Die Modellvergleiche fallen durchgehend zugunsten des dreidimensionalen Modells aus, das gleichzeitig auch die höchste Varianz aufweist. Dieses Ergebnis entspricht der theoretisch angenommenen Struktur professioneller Unterrichtswahrnehmung.

[28] adaptiert nach Jahn et al. (in Druck).

6.2.2 Vergleich der Itemschwierigkeiten

Hinsichtlich der Prüfung der Vergleichbarkeit der Itemschwierigkeiten des Instruments wurden für die drei Studierendengruppen (Lehramt für Primarstufe, Sekundarstufe und Berufliche Schulen) DIF-Analysen durchgeführt, bei denen Gruppenunterschiede in den Personenfähigkeiten kontrolliert wurden sowie der mittlere DIF über alle Items auf 0 festgelegt wurde. Der Chi-Quadrat-Test auf Parametergleichheit in diesen Analysen deutet auf differentielle Unterschiede in den Itemschwierigkeiten zwischen den drei Studierendengruppen hin ($\chi^2(206) = 6193.29$, $p < .05$). Die DIF-Werte rangieren über alle drei Studierendengruppen hinweg zwischen -0.90 und 1.02 (*SD* = .28). Ein deskriptiver Überblick über die DIF-Werte ist in Tabelle 14 dargestellt.

Tabelle 14

Deskriptive Statistik der DIF-Werte der drei Studierendengruppen Lehramt Primarstufe, Sekundarstufe und Berufliche Schulen (n = 106 Items)[29]

	M	*SD*	*Min*	*Max*
Primarstufe	0.00	0.34	-0.65	1.02
Sekundarstufe	0.00	0.20	-0.51	0.39
Berufliche Schulen	000	0.29	-0.90	0.77
Gesamt	0.00	0.28	-0.90	1.02

Anmerkung. DIF = Differential Item Functioning.

Gemäß der ETS-Klassifikation (Penfield & Camilli, 2007) weisen insgesamt 17 Items (in mindestens einer Studierendengruppe) moderaten DIF und neun Items (in mindestens einer Studierendengruppe) großen DIF auf. Dies entspricht 25 % aller Items und ist damit ein Indikator für substantielle DIF-Effekte auf Ebene des Messinstruments (vgl. Penfield & Algina, 2006). Nimmt man die Studierendengruppen einzeln in den Blick, fällt auf, dass in der Studierendengruppe Lehramt Sekundarstufe, die als Referenzgruppe dient, bis auf ein Item DIF zu vernachlässigen ist, wohingegen in der Gruppe Lehramt Primarstufe die größte Anzahl ($n = 21$) an Items mit DIF vorhanden ist. Allerdings variiert die Richtung des DIF in den Studierendengruppen Lehramt Primarstufe und Lehramt Berufliche Schulen unsystematisch. Dies

[29] adaptiert nach Jahn et al. (in Druck).

bedeutet, dass sowohl Items mit positiven Werten existieren, die schwieriger zu lösen sind, als auch Items mit negativen Werten, die leichter zu lösen sind. Lediglich für die Studierendengruppe Lehramt Primarstufe zeichnet sich eine Tendenz zu positiven DIF-Werten ab, die auf tendenziell höhere Itemschwierigkeiten hindeutet. Ein Überblick über die Verteilung der Items zur Erfassung professioneller Unterrichtswahrnehmung, die substantiellen DIF aufweisen, ist Tabelle 15 zu entnehmen.

Tabelle 15

Überblick über die Gesamtanzahl der Items mit substantiellen DIF, getrennt nach den drei Studierendengruppen sowie aufgeteilt nach moderatem und großem DIF sowie positiven und negativen Werten (professionelle Unterrichtswahrnehmung: n = 106 Items)

	Anzahl der Items				
	gesamt	moderater DIF	großer DIF	positive Werte	negative Werte
Primarstufe	21	13	8	15	6
Sekundarstufe	1	1	0	0	1
Berufliche Schulen	16	13	3	7	9

Anmerkung. DIF = Differential Item Functioning.

Dieses unsystematische Bild hinsichtlich der drei Studierendengruppen zeigt sich auch, wenn die Items mit substantiellem DIF differenziert nach den drei Aspekten Beschreiben, Erklären und Vorhersagen betrachtet werden (vgl.

Tabelle 16). Die Items mit substantiellem DIF verteilen sich auf alle drei Aspekte professioneller Unterrichtswahrnehmung und variieren unsystematisch.

Tabelle 16

Überblick über die Anzahl der Items mit substantiellen DIF, aufgeteilt nach moderatem und großem DIF sowie positiven und negativen Werten getrennt nach den drei Studierendengruppen (Beschreiben: n = 38, Erklären: n = 36, Vorhersagen: n = 32)[30]

	Anzahl der Items											
	moderater DIF			großer DIF			positive Werte			negative Werte		
	B	E	V	B	E	V	B	E	V	B	E	V
Primarstufe	5	2	6	4	3	1	5	4	6	4	1	1
Sekundarstufe	0	1	0	0	0	0	0	0	0	0	1	0
Berufliche Schulen	9	3	1	1	2	0	5	2	0	5	3	1

Anmerkungen. DIF = Differential Item Functioning. B = Beschreiben, E = Erklären, V = Vorhersagen.

In einer detaillierten Analyse der $n = 26$ Items mit substantiellem DIF (vgl. Tabelle 17) wird sichtbar, dass über alle Studierendengruppen hinweg tendenziell auf Ebene des Beschreibens die größte Anzahl an Items mit DIF auftritt. Insgesamt weisen $n = 15$ Items für eine Studierendengruppe und $n = 10$ Items für zwei Studierendengruppen DIF auf. Demnach sind es nicht immer dieselben Items, die für verschiedene Studierendengruppen DIF aufweisen. Es fällt auf, dass sich bei den Items, die für zwei Studierendengruppen substantiellen DIF aufweisen, die Richtung des DIF stets unterscheidet und über die beiden Gruppen hinweg variiert. Die Items sind folglich nicht konstant schwieriger oder leichter für die Studierendengruppen Lehramt Primarstufe und Berufliche Schulen, noch bevorzugen oder benachteiligen sie konstant eine Studierendengruppe. Die Befunde dieser detaillierten Itemanalyse unterstreichen die fehlende Systematik des auftretenden DIFs.

[30] adaptiert nach Jahn et al. (in Druck).

Tabelle 17

Auflistung der einzelnen Items mit substantiellem DIF

Items	Primarstufe	Sekundarstufe I + II	Berufliche Schulen
Beschreiben (n = 12)			
item_1	−		+
item_2	−		+
item_3	−		+
item_4			+
item_5			−
item_6			+
item_7	−		
item_8	++		−
item_9	++		−
item_10	++		−−
item_11	+		−
item_12	++		
Erklären (n = 7)			
item_13	++		−
item_14			−−
item_15	−−		++
item_16			+
item_17	+		
item_18	++		−
item_19		−	
Vorhersagen (n = 7)			
item_20	+		
item_21	+		
item_22	−		
item_23	+		
item_24	++		−
item_25	+		
item_26	+		

Anmerkungen. + = positiver, moderater DIF; ++ = positiver, großer DIF; − = negativer, moderater DIF; −− = negativer, großer DIF.

Berücksichtigt man für weitere Analysen den Aufbau des Tools Observer (vgl. Abschnitt 2.2.3.2.4), gemäß dem 54 Rating-Items mit jeweils vier verschiedenen Videoclips als unabhängige Item-Prompts eingesetzt werden, sind zwei Befunde anzumerken. Zum einen weist insgesamt lediglich eines der 54 Rating-Items durchgehend bei allen Einsätzen substantiellen DIF auf und zum anderen gibt es einen Videoclip, bei dem 58 % aller Items substantiellen DIF zeigen.

Zusammenfassend ist festzuhalten, dass DIF-Analysen zur Überprüfung der Vergleichbarkeit der Itemschwierigkeiten des Tools Observer substantielle Unterschiede in den Itemschwierigkeiten für die drei Studierendengruppen Lehramt Primarstufe, Sekundarstufe und Berufliche Schulen aufzeigen. Allerdings ist keine Systematik der DIF-Werte bezüglich Items, Richtung und Studierendengruppe erkennbar. Leichte Tendenzen zeigen sich lediglich dahingehend, dass zum einen die meisten Items für die Studierendengruppe Lehramt Primarstufe DIF aufweisen und tendenziell eher schwieriger sind, wohingegen DIF für die Studierendengruppe Lehramt Sekundarstufe, die als Referenzgruppe dient, bis auf ein Item zu vernachlässigen ist, und zum anderen, dass über alle Studierendengruppen hinweg auf Ebene des Beschreibens die meisten Items mit DIF auftreten.

6.2.3 Zusammenfassung der Ergebnisse

Die Prüfung der Kompetenzstruktur in den drei Studierendengruppen Lehramt Primarstufe, Sekundarstufe und Berufliche Schulen führt hypothesenkonform zu vergleichbaren Ergebnissen. Im Gegensatz dazu zeigen DIF-Analysen substantielle Unterschiede in den Itemschwierigkeiten der drei Studierendengruppen. Somit ist abschließend festzuhalten, dass die Struktur professioneller Unterrichtswahrnehmung in den drei Studierendengruppen (Lehramt Primarstufe, Sekundarstufe und Berufliche Schulen) vergleichbar ist, jedoch Einschränkungen hinsichtlich der Vergleichbarkeit der Itemschwierigkeiten bestehen.

6.3 Diskussion[31]

Im Rahmen eines übergreifenden Einsatzes im Large-Scale-Kontext soll das Tool Observer langfristig von Studierenden unterschiedlicher Lehramtsstudiengänge bearbeitet und dazu genutzt werden, Unterschiede in der professionellen Unterrichtswahrnehmung zwischen diesen Studierenden abzubilden. Im bisherigen Projektverlauf wurde das Instrument jedoch ausschließlich mit Lehramtsstudierenden der Sekundarstufe erprobt. Im Rahmen der dritten Forschungsfrage wurde deshalb der Validitätsaspekt *Generalisierbarkeit* über verschiedene Lehramtsstudiengänge hinweg überprüft. An einer Stichprobe von $N = 1008$ Studierenden, die sich auf die Lehramtsstudiengänge Lehramt für Primarstufe, Sekundarstufe und Berufliche Schulen verteilen, wurde untersucht, inwieweit die Struktur professioneller Unterrichtswahrnehmung sowie die Itemschwierigkeiten für diese drei Studierendengruppen vergleichbar sind. Im Folgenden werden zunächst zentrale Befunde zusammengefasst und inhaltlich diskutiert. Anschließend wird das methodische Vorgehen bewertet und weitere Implikationen werden dargestellt.

6.3.1 Zusammenfassung und Diskussion zentraler Befunde

Im Folgenden werden zunächst die zentralen Befunde zur Vergleichbarkeit der Kompetenzstruktur und anschließend zur Vergleichbarkeit der Itemschwierigkeiten zusammengefasst und aus inhaltlicher Perspektive diskutiert.

6.3.1.1 Vergleichbarkeit der Kompetenzstruktur

Zur Überprüfung der Vergleichbarkeit der Kompetenzstruktur professioneller Unterrichtswahrnehmung wurden für alle drei Studierendengruppen (Lehramt für Primarstufe, Sekundarstufe und Berufliche Schulen) getrennte Einzelskalierungen vorgenommen. Dabei wurde analog zum methodischen Vorgehen der ersten Forschungsfrage (vgl. Abschnitt 4.1.3.2) das theoretisch postulierte dreidimensionale Modell mit den drei Aspekten Beschreiben, Erklären und Vorhersagen gegen ein eindimensionales Modell (professionelle Unterrichtswahrnehmung als eine Gesamtfähigkeit) sowie ein zweidimensionales Modell (Beschreiben und Integrieren) geprüft. Die Modellvergleiche fallen hypothesenkonform durchgehend zugunsten des dreidimensionalen Modells aus, das gleichzeitig auch die höchste Varianz aufweist. Bei einem Ver-

[31] Dieser Abschnitt basiert zu Teilen auf Jahn et al. (in Druck).

gleich der drei Studierendengruppen fällt auf, dass die Modelle für die Gruppe der Studierenden Lehramt für Berufliche Schulen die niedrigsten Varianzen aufweisen. Eine mögliche Erklärung für diesen Befund könnte sein, dass diese Studierendengruppe mit zahlreichen Fächerkombinationsmöglichkeiten in sich sehr heterogen ist. Andererseits könnte dieser Befund darauf zurückzuführen sein, dass sich diese Gruppe durch die Berufserfahrung der meisten Studierenden stark von den anderen beiden Gruppen unterscheidet.

Der Befund, dass das dreidimensionale Modell in jeder der drei Studierendengruppen (Lehramt für Primarstufe, Sekundarstufe und Berufliche Schulen) die Daten am besten abbildet, entspricht der theoretisch angenommenen Struktur professioneller Unterrichtswahrnehmung. Darüber hinaus werden damit die Befunde aus der Skalierungsstudie (Seidel & Stürmer, in Druck) sowie der Scaling-up-Studie im Rahmen der ersten Forschungsfrage (vgl. Abschnitt 4.2) bestätigt. Damit konnte die Kompetenzstruktur für alle drei Studierendengruppen repliziert werden. Dadurch wird unterstrichen, dass es sich bei der professionellen Unterrichtswahrnehmung mit den Aspekten Beschreiben, Erklären und Vorhersagen um eine Fähigkeit handelt, die Studierende verschiedener Lehramtsstudiengänge zeigen.

6.3.1.2 Vergleichbarkeit der Itemschwierigkeiten

Zur Überprüfung der Vergleichbarkeit der Itemschwierigkeiten wurden mit Blick auf die drei Studierendengruppen (Lehramt für Primarstufe, Sekundarstufe und Berufliche Schulen) DIF-Analysen durchgeführt. Als Referenzgruppe diente die Gruppe der Studierenden Lehramt für Sekundarstufe, da das Instrument Videoclips aus der Sekundarstufe enthält und im bisherigen Projektverlauf für diese Zielgruppe erprobt wurde. Es zeigen sich substantielle Unterschiede in den Itemschwierigkeiten für die drei Studierendengruppen. Insgesamt weisen 25 % aller Items substantiellen DIF auf. Dies entspricht der kritischen Grenze, ab der auch auf Ebene des Instruments von substantiellen DIF-Effekten ausgegangen werden muss (vgl. Penfield & Algina, 2006). Auf eine mögliche Erklärung für diese DIF-Effekte wird im übernächsten Absatz eingegangen. Die Items mit substantiellem DIF verteilen sich in ähnlichem Umfang auf die Studierendengruppen Lehramt für Primarstufe und Lehramt für Berufliche Schulen mit einer leichten Tendenz zur Gruppe Lehramt für Primarstufe. In der Studierendengruppe Lehramt für Sekundarstufe, die als Referenzgruppe herangezogen wurde, weist dennoch ein Item substantiellen DIF auf. Dies könnte darauf zurückzuführen sein, dass dieses Item in allen drei Studierendengruppen nur selten gelöst wurde (6.6 – 16.3 % Lösungswahrscheinlichkeit).

In Bezug auf die Richtung des DIF in den Studierendengruppen Lehramt für Primarstufe und Lehramt für Berufliche Schulen zeigt sich ein unsystematisches Bild. In beiden Gruppen existieren sowohl Items mit positiven Werten, die schwieriger zu lösen sind, als auch Items mit negativen Werten, die leichter zu lösen sind. Lediglich für die Studierendengruppe Lehramt Primarstufe zeichnet sich eine Tendenz zu positiven DIF-Werten ab, die auf tendenziell höhere Itemschwierigkeiten hindeutet. Eine detaillierte Analyse der 26 Items mit substantiellem DIF unterstreicht die fehlende Systematik der DIF-Werte. Die Items verteilen sich auf alle drei Aspekte professioneller Unterrichtswahrnehmung, mit der größten Anzahl auf Ebene des Beschreibens, und variieren dabei unsystematisch in der Richtung der DIF-Werte. Darüber hinaus sind es nicht immer dieselben Items, die für verschiedene Studierendengruppen substantiellen DIF aufweisen. Weisen Items für zwei Studierendengruppen substantiellen DIF auf, sind die Items wiederum nicht für beide Gruppen schwieriger oder leichter und bevorzugen oder benachteiligen auch nicht konstant eine der zwei Gruppen.

Trotz der fehlenden Systematik der DIF-Werte bedeuten die substantiellen Unterschiede in den Itemschwierigkeiten, dass das Instrument in der aktuellen Form nicht dazu geeignet ist, systematische Unterschiede zwischen Studierenden verschiedener Lehramtsstudiengänge vergleichend zu testen. Bei der Suche nach möglichen Erklärungen für diesen Befund stellt der Aufbau des Instruments einen entscheidenden Anknüpfungspunkt dar (vgl. Abschnitt 2.2.3.2.4). Im Tool Observer werden 54 Rating-Items mit vier verschiedenen Videoclips, die Unterricht aus der Sekundarstufe zeigen, als item-unabhängige Prompts eingesetzt, um professionelle Unterrichtswahrnehmung vor dem Hintergrund von Zielorientierung, Lernbegleitung und Lernatmosphäre zu erfassen. Mit Blick auf diesen Aufbau sind zwei Befunde besonders zu beachten. Zum einen weist insgesamt nur eines der 54 Rating-Items durchgehend bei allen Einsätzen substantiellen DIF auf und zum anderen gibt es einen Videoclip, bei dem 58 % aller Items substantiellen DIF zeigen. Diese Befunde könnten ein Hinweis darauf sein, dass eine Erklärung für die unterschiedlichen Itemschwierigkeiten weniger in den Rating-Items begründet ist (z. B. Formulierung), sondern eher in der entsprechenden Kontextualisierung (im Clip gezeigter Unterricht). Es könnte sein, dass es Studierenden für Lehramt Primarstufe und Lehramt Berufliche Schulen schwerfällt, ihr Wissen über Zielorientierung, Lernbegleitung und Lernatmosphäre auf Unterrichtssituationen anzuwenden, die Unterricht aus der Sekundarstufe zeigen und damit nicht ihrem späteren Handlungsfeld entsprechen. Denn der Erwerb von Wissen über effektives Lehren und Lernen ist Teil eines jeden Lehramtsstudiums (Terhart, 2009), da dieses Wissen eine Voraussetzung für die fächer- und schulartübergreifen-

de Gestaltung von Lernumgebungen darstellt (Voss et al., 2011). Der Einwand, dass sich die Studierenden in der vorliegenden Arbeit zu Beginn ihres Lehramtsstudiums befanden und bis zum Zeitpunkt der Datenerhebung eventuell noch keine entsprechende Lehrveranstaltung besuchten, ist im Hinblick auf die Vergleichbarkeit der Itemschwierigkeiten nicht relevant. Bei der gewählten Methode zur DIF-Analyse, IRT-Skalierung durch ein Multifacetten-Modell, werden Gruppenunterschiede in den Personenfähigkeiten kontrolliert.

6.3.2 Methodische Überlegungen

Im Rahmen der Überprüfung der dritten Forschungsfrage konnte gezeigt werden, dass das Tool Observer die Struktur professioneller Unterrichtswahrnehmung für verschiedene Studierendengruppen (Lehramt für Primarstufe, Sekundarstufe und Berufliche Schulen) vergleichbar erfasst. Im Gegensatz dazu erwiesen sich die Itemschwierigkeiten für die verschiedenen Studierendengruppen als nicht vergleichbar. Im Hinblick auf die Interpretation der Befunde sollten jedoch folgende zwei methodische Einschränkungen bedingt durch die Zusammensetzung der Stichprobe beachtet werden: (1) heterogene Stichprobengrößen der einzelnen Studierendengruppen und (2) unausgewogene Verteilung auf verschiedene Hochschulstandorte.

Erstens weisen die drei Studierendengruppen deutlich unterschiedliche Stichprobengrößen auf. Die Gruppe Lehramt Sekundarstufe umfasst ungefähr viermal so viele Studierende wie die Gruppen Lehramt Primarstufe und Lehramt Berufliche Schulen. Dadurch werden Verzerrungen im Rückschluss auf die Zielpopulation begünstigt (Field, 2009).

Zweitens sind in der Untersuchungsstichprobe die Lehramtsstudiengänge mit den Hochschulstandorten konfundiert. Zum einen bieten nicht alle Standorte alle Lehramtsstudiengänge an. Zum anderen setzt sich die Gruppe Lehramt für Berufliche Studien vornehmlich aus Studierenden zweier Standorte zusammen. Infolgedessen ist keine getrennte Prüfung von Effekten der Lehramtsstudiengänge und Hochschulstandorte möglich. Diese Verzerrung der Stichprobe schränkt die Generalisierbarkeit der Befunde auf die Zielpopulation ein (vgl. Kunter & Klusmann, 2010). Deshalb sollten die Befunde der vorliegenden Arbeit an einer Stichprobe, in der Lehramtsstudiengänge und Hochschulstandorte nicht konfundiert sind, überprüft werden.

6.3.3 Implikationen

Trotz der diskutierten methodischen Einschränkungen wurde im Rahmen der Überprüfung der dritten Forschungsfrage Evidenz dafür generiert, in welchem Rahmen die Messergebnisse des Tools Observer über verschiedene Lehramtsstudiengänge hinweg valide interpretiert werden können und wo auf Basis der Befunde der vorliegenden Arbeit Grenzen bestehen. Das Instrument stellt einen geeigneten Zugang dar, um die Struktur professioneller Unterrichtswahrnehmung innerhalb der Studierendengruppen Lehramt Primarstufe, Sekundarstufe und Berufliche Schulen zu erfassen. Daher kann es innerhalb dieser Studierendengruppen als Diagnoseinstrument eingesetzt werden. Allerdings sprechen die vorliegenden Befunde dagegen, das Instrument dazu zu nutzen, um systematische Unterschiede in der professionellen Unterrichtswahrnehmung zwischen Studierenden verschiedener Lehramtsstudiengänge abzubilden. Denn es ist nicht gewährleistet, dass gemessene Unterschiede in der professionellen Unterrichtswahrnehmung zwingend auf Unterschiede in den Personenfähigkeiten zurückzuführen sind und nicht etwa durch Eigenschaften des Instruments hervorgerufen werden.

Daraus ergeben sich weiterführende Implikationen für Forschung und Lehre. In weiteren Studien sollte überprüft werden, inwieweit professionelle Unterrichtswahrnehmung über verschiedene Lehramtsstudiengänge hinweg erfasst werden kann. Eine Möglichkeit stellt die Integration von Videoclips, die Unterricht aus der Primarstufe oder aus Beruflichen Schulen zeigen, dar. Eine Studie sollte so angelegt sein, um bei Verwendung derselben Rating-Items Effekte der Videoclips hinsichtlich Übereinstimmung mit dem späteren Handlungsfeld systematisch überprüfen zu können (vgl. Abschnitt 8.3.1). Wenn durch weitere Studien die Vermutung untermauert wird, dass es Studierenden schwerfällt, ihr Wissen auf Unterrichtssituationen anzuwenden, die nicht ihrem späteren Handlungsfeld entsprechen, ist dieser Befund von direkter Relevanz für die universitäre Lehrerbildung. Es stellt sich die Frage, inwieweit Wissen über effektives Lehren und Lernen und pädagogisch-psychologisches Wissen generell im Rahmen universitärer Lehrerbildung über den Kontext von Schularten hinweg auf einer Metaebene aufgebaut werden kann. Nachdem professionelle Unterrichtswahrnehmung als Indikator für die Anwendung dieses Wissens (Seidel & Stürmer, in Druck) per Definition (vgl. Abschnitt 2.1.1.1) kontextspezifisch ist, muss diese durch Erfahrungen und Lernen in relevanten domänen-spezifischen Situationen erworben werden (vgl. Koeppen et al., 2008). Es scheint allerdings angebracht, den Kontext nicht zu eng zu fassen, verschiedene Kontexte als Lerngelegenheit anzubieten und die Studierenden dafür zu sensibilisieren, dass das Wissen um effektives Lehren und Lernen fächer-, stufen- und schulartübergreifend relevant ist. Ansonsten

könnte der Transfer des in einem spezifischen Kontext erworbenen Wissens bei der Anwendung auf unbekannte Unterrichtssituationen gefährdet sein.

7 ÜBERPRÜFUNG DES VALIDITÄTSASPEKTS EXTERNALITÄT

Die vierte Forschungsfrage zielt in zwei Teilfragen auf die Überprüfung des Validitätsaspekts *Externalität* ab. Als externale Kriterien werden Variablen verwendet, die typischerweise im Rahmen von Verfahren zur Eignungsfeststellung als Prädiktoren für Studien- und Berufserfolg erfasst werden. (vgl. Abschnitt 2.3.2.4.4). In der ersten Teilfrage wird untersucht, inwieweit sich Lehramtsstudierende in Subgruppen hinsichtlich pädagogischen Interesses, Interesse an den Bildungswissenschaften, lernbegleitungsorientiertem Lehrbegriff und dem Persönlichkeitsmerkmal Verträglichkeit gruppieren lassen. In der zweiten Teilfrage wird überprüft, inwieweit ein Zusammenhang zwischen der Zugehörigkeit zu diesen Profilen und der professionellen Unterrichtswahrnehmung Lehramtsstudierender besteht. Im Folgenden wird zunächst das methodische Vorgehen beschrieben. Anschließend werden die Ergebnisse dargestellt und diskutiert.

7.1 Methodisches Vorgehen

Im Rahmen des methodischen Vorgehens wird die Zusammensetzung der Stichprobe erläutert, auf die eingesetzten Messinstrumente eingegangen und es werden die Methoden zur Datenauswertung ausführlich beschrieben.

7.1.1 Stichprobe

Die Stichprobe zur Überprüfung der vierten Forschungsfrage stellt eine Substichprobe der Scaling-up-Stichprobe (vgl. Abschnitt 4.1.1) dar. Sie setzt sich aus denjenigen Lehramtsstudierenden zusammen, bei denen durch die freiwillige Angabe des Identifikationscodes die Daten zur professionellen Unterrichtswahrnehmung aus der Scaling-up-Studie den ersten beiden Befragungswellen des Panels zum Lehramtsstudium (PaLea; Bauer et al., 2010) zugeordnet werden konnten. Die Untersuchungsstichprobe besteht aus $N = 159$ Lehramtsstudierenden (74.7 % weiblich) unterschiedlicher Studienfächer und Lehramtsstudiengänge, die zu Beginn ihres ersten Semesters an der Eingangsbefragung und am Ende ihres ersten Semesters an der zweiten Befragungswelle von PaLea teilnahmen sowie im zweiten (69.2 %) oder dritten Semester (30.8 %) das Tool Observer im Rahmen der Scaling-up-Studie (vgl. Abschnitt 4.1.1) bearbeiteten. Ein Überblick über die deskriptive Statistik der Untersuchungsstichprobe ist in

Tabelle 18 dargestellt. Insgesamt sind zwölf der 13 PaLea-Universitätsstandorte vertreten. Dies entspricht keiner aktiven Selektion, sondern kam zufällig zustande. Die genaue Verteilung der $N = 159$ Lehramtsstudierenden auf die einzelnen Universitätsstandorte ist Tabelle 19 zu entnehmen.

Tabelle 18

Deskriptive Statistik der Untersuchungsstichprobe

	Untersuchungsstichprobe ($N = 159$)	
	M	*SD*
Alter	21.39	2.79
	Anteil (in %)	
Geschlecht weiblich	74.7	
Lehramtsstudienrichtung		
Berufliche Schulen	42.8	
Sekundarstufe II	35.2	
Sekundarstufe I	10.0	
Primarstufe	11.9	
Fächerkombination		
Nur Naturwissenschaften	38.6	
Nur Geistes- und Sozialwissenschaften	26.6	
Gemischt	33.5	

Tabelle 19

Verteilung der Lehramtsstudierenden auf die einzelnen Universitäten für die Untersuchungsstichprobe (N = 159) im Vergleich zur PaLea-Gesamtstichprobe (N = 4468)

	Universitätsstandorte												
	B	D	E	F	G	H	I	J	K	L	M	N	P
Untersuchungsstichprobe (N = 159)													
Anteil (in %)	1.9	0.0	6.3	8.2	3.1	5.0	2.5	5.7	1.9	47.8	14.5	1.9	1.3
PaLea-Gesamtstichprobe (N = 4468)													
Anteil (in %)	9.3	3.2	7.4	8.3	2.1	8.3	7.9	6.2	14.3	4.9	6.4	11.5	10.3

7.1.2 Messinstrumente

Nachfolgend wird beschrieben, wie die relevanten Variablen im Rahmen der vierten Forschungsfrage nach der externalen Validität des Tools Observer operationalisiert sind. Zunächst werden im Hinblick auf die erste Teilfrage nach der Profilbildung die Variablen dargestellt, die in der PaLea-Studieneingangsbefragung (kognitive Grundfähigkeiten, pädagogisches Interesse, subjektive Lehrbegriff und das Persönlichkeitsmerkmal Verträglichkeit) sowie der zweiten PaLea-Befragungswelle am Ende des ersten Semesters (Interesse an Bildungswissenschaften) erfasst wurden. Anschließend wird im Hinblick auf die zweite Teilfrage nach dem Zusammenhang mit der professionellen Unterrichtswahrnehmung auf die Erfassung eben dieser in der Scaling-up-Studie eingegangen.

7.1.2.1 Kognitive Grundfähigkeiten

Zur Erfassung der kognitiven Grundfähigkeiten wurde der Subtest „Figurenanalogien“ aus dem kognitiven Fähigkeitstest (KFT; Heller & Perleth, 2000) verwendet, anhand dessen schlussfolgerndes Denken erfasst wird. Dieser Subtest korreliert hoch mit dem Gesamttest und stellt damit einen geeigneten Indikator für allgemeine kognitive Fähigkeiten dar (Retelsdorf & Möller, 2012). Die Studierenden haben acht Minuten Zeit, um 25 Items zu lösen. Es wird ein Summenscore gebildet. Dieser erreicht eine zufriedenstellende Reliabilität (Cronbach’s $\alpha = .66$).

7.1.2.2 Pädagogisches Interesse

Pädagogisches Interesse wurde mit der gleichnamigen Subskala des Fragebogens zur Erfassung der Motivation für das Lehramtsstudium (FEMOLA; Pohlmann & Möller, 2010) erfasst. Alle Items werden eingeleitet mit „Ich habe das Lehramtsstudium gewählt, weil…". Insgesamt werden sieben Items (z. B. „… ich gerne mit Kindern und Jugendlichen arbeite.") auf einer vierstufigen Likert-Skala von ‚1' trifft überhaupt nicht zu bis ‚4' trifft völlig zu eingeschätzt. Die Subskala Pädagogisches Interesse weist eine hohe Reliabilität auf (Cronbach's $\alpha = .81$).

7.1.2.3 Interesse an Bildungswissenschaften

Zur Erfassung des Interesses an Bildungswissenschaften wurde im Kontext von PaLea (Bauer et al., 2010) in Anlehnung an den Fragebogen zur Erfassung von Studieninteresse (Schiefele, Krapp, Wild & Winteler, 1993) eine Skala mit drei Items (z. B. „Die Inhalte dieses Studienbereichs entsprechen meinen persönlichen Neigungen") entwickelt. Damit wird das Interesse an Bildungswissenschaften auf einer vierstufigen Likert-Skala (‚1' trifft nicht zu/‚4' trifft zu) hoch reliabel erfasst (Cronbach's $\alpha = .91$).

7.1.2.4 Lernbegleitungsorientierter Lehrbegriff

In der vorliegenden Arbeit zur Erfassung des lernbegleitungsorientierten Lehrbegriffs wurde eine Subskala aus PaLea eingesetzt (Bauer et al., 2010, März). Sie erfasst prototypische Konzepte von Unterrichten über sieben Begriffe (z. B. „Interesse wecken"), die auf einer vierstufigen Likert-Skala danach bewertet werden, wie ähnlich sie der eigenen Vorstellung von Unterrichten sind (‚1' nicht ähnlich/ ‚4' sehr ähnlich). Die Subskala lernbegleitungsorientierter Lehrbegriff zeigt eine zufriedenstellende Reliabilität (Cronbach's $\alpha = .73$).

7.1.2.5 Persönlichkeitsmerkmal Verträglichkeit

Das Persönlichkeitsmerkmal Verträglichkeit wurde mit einer Subskala der Kurzversion des Fragebogens zur Erfassung der Big Five (Herzberg & Brähler, 2006) erhoben. Dazu werden zwei Adjektive (z. B. „verständnisvoll"), die dem Persönlichkeitsmerkmal Verträglichkeit zugeordnet werden können, auf einer vierstufigen Likert-Skala (‚1' stimmt gar nicht/ ‚4' stimmt genau) eingeschätzt. Die Subskala Verträglichkeit erreicht eine zufriedenstellende Reliabilität (Cronbach's $\alpha = .73$).

7.1.2.6 Professionelle Unterrichtswahrnehmung

Professionelle Unterrichtswahrnehmung wird mit dem videobasierten Online-Tool Observer (Seidel et al., 2010b) erfasst. Eine detaillierte Beschreibung des Instruments findet sich in Abschnitt 2.2.3.2.4. Aus der Einschätzung von insgesamt 216 Rating-Items resultieren reliable Scores für professionelle Unterrichtswahrnehmung (Cronbach's $\alpha = .95$) und die drei Aspekte Beschreiben (Cronbach's $\alpha = .78$), Erklären (Cronbach's $\alpha = .85$) und Vorhersagen (Cronbach's $\alpha = .91$). Diese entsprechen jeweils der prozentualen Übereinstimmung mit der Expertennorm und bewegen sich zwischen 0 und 1.

7.1.3 Auswertungsmethoden

Im Folgenden wird zunächst auf die Überprüfung der Stichprobe und den Umgang mit fehlenden Werten eingegangen. Anschließend wird das methodische Vorgehen bei der Bildung von Profilen und der Untersuchung des Zusammenhangs zwischen diesen Profilen und der professionellen Unterrichtswahrnehmung dargestellt.

7.1.3.1 Überprüfung der Stichprobe

Nachdem die Untersuchungsstichprobe sowohl eine Substichprobe der Scaling-up-Stichprobe als auch eine Substichprobe der PaLea-Gesamtstichprobe darstellt (vgl. Abschnitt 7.1.1), wird in einem ersten Analyseschritt überprüft, inwieweit die Untersuchungsstichprobe hinsichtlich der für die vierte Forschungsfrage relevanten Variablen mit diesen beiden Stichproben vergleichbar ist. Mit Blick auf die Vergleichbarkeit mit der Scaling-up-Stichprobe werden die drei Aspekte professioneller Unterrichtswahrnehmung Beschreiben, Erklären und Vorhersagen sowie Geschlecht und Alter als zusätzliche Kontrollvariablen untersucht. Ebenso werden für den Vergleich mit der PaLea-Gesamtstichprobe neben den Variablen für die Profilbildung (pädagogisches Interesse, Interesse an Bildungswissenschaften, lernbegleitungsorientierter Lehrbegriff und Verträglichkeit) weitere Kontrollvariablen (Verteilung auf die verschiedenen Universitätsstandorte, Geschlecht, Alter und kognitive Grundfähigkeiten) herangezogen. Eine derart sorgfältige Überprüfung möglicher Verzerrungen soll dazu beitragen, Informationen darüber zu gewinnen, in welchem Ausmaß die Ergebnisse dieser Studie auf die Zielpopulation übertragbar sind (Kunter & Klusmann, 2010).

7.1.3.2 Umgang mit fehlenden Werten

Für die Variable Interesse an Bildungswissenschaften liegen bei 53.8 % und für die Variable Verträglichkeit bei 8.2 % der Lehramtsstudierenden keine Angaben vor. Für alle anderen Variablen, die im Rahmen der vierten Forschungsfrage relevant sind, existieren keine fehlenden Werte. Zum Umgang mit den fehlenden Werten wurde eine multiple Imputation vorgenommen. Diese Methode entspricht dem „state of the art“, da sie älteren Methoden (z. B. Fallausschluss oder Imputation des Mittelwerts) hinsichtlich Präzision und Teststärke der nachfolgenden Analysen überlegen ist (vgl. Schafer & Graham, 2002). Bei der multiplen Imputation wird das Problem fehlender Werte vor den Analysen zur Datenauswertung gelöst, indem die fehlenden Werte durch $m > 1$ Sets simulierter imputierter Werte ersetzt werden, die in m unterschiedlichen vollständigen Datensätzen resultieren (Collins, Schafer & Kam, 2001). Multiple Imputation hat sich auch bei hohen Anteilen fehlender Werte als effizientes und robustes Verfahren erwiesen (vgl. Graham, 2009). In das Imputationsmodell wurden neben allen analysierten Variablen zwei Hilfsvariablen aufgenommen, aber für weitere Analysen nicht berücksichtigt: Vorfreude aufs Studium (Bauer et al., 2010), die mit acht Rating-Items auf einer 3-stufigen Likert-Skala reliabel eingeschätzt wird (Cronbach's $\alpha = .72$), und berufliche Selbstwirksamkeit (adaptiert nach Schaufeli, Martínez, Marques Pinto, Salanova & Bakker, 2002), die mit drei Rating-Items auf einer 3-stufigen Likert-Skala reliabel eingeschätzt wird (Cronbach's $\alpha = .70$). Gemäß aktuellen Empfehlungen aus der Methodenliteratur (van Buuren, 2012) orientierte sich die Anzahl imputierter Datensätze an der Prozentzahl fehlender Werte. Dementsprechend wurden mit der Software Mplus 7.11 (Muthén & Muthén, 1998-2013) $m = 50$ vollständige Datensätze mit den Variablen Interesse an Bildungswissenschaften, pädagogisches Interesse, lernbegleitungsorientierter Lernbegriff, Verträglichkeit sowie professionelle Unterrichtswahrnehmungen mit ihren drei Aspekten erstellt. Die im Ergebnisteil berichteten Analysen stellen kombinierte Befunde aus den 50 Datensätzen dar.

7.1.3.3 Profilbildung

Zur Überprüfung der ersten Teilfrage wurden latente Profilanalysen (LPA; vgl. Pastor et al., 2007) mit der Software Mplus 7.11 (Muthén & Muthén, 1998-2013) basierend auf den Mittelwerten der folgenden Skalen durchgeführt: pädagogisches Interesse, Interesse an Bildungswissenschaften, lernbegleitungsorientierter Lehrbegriff und Verträglichkeit. Für diese Variablen wurde zuvor eine z-Standardisierung vorgenommen, um die Interpretation zu er-

leichtern. Latente Profilanalysen stellen eine geeignete Methode dar, um Personen aufgrund von qualitativen Unterschieden in Mustern mehrerer latenter Variablen zu clustern (Lubke & Muthén, 2005). Dabei bietet dieses modell-basierte Verfahren den Vorteil, verschiedene Profil-Lösungen anhand der jeweiligen Modell-Fits vergleichen zu können (Pastor et al., 2007). Für eine Übersicht über weitere Vorteile dieser Methode gegenüber herkömmlichen Methoden wie Clusteranalysen siehe Pastor und Kollegen (2007). Ziel der latenten Profilanalyse war es, eine möglichst inhaltlich interpretierbare und sparsame Klassenlösung zu erzielen. Zur Beurteilung der Anzahl latenter Profile wurden folgende Kriterien herangezogen (vgl. Geiser, 2009): Aikakes Informationskriterium (AIC), Bayessches Informationskriterium (BIC) und das um die Stichprobengröße angepasste Bayessche Informationskriterium (aBIC). Aufgrund der multiplen Imputation konnte kein Lo-Mendell-Rubin-Likelihood-Ratio-Test (Lo, Mendell & Rubin, 2001) berechnet werden, um zu vergleichen, inwieweit ein zusätzliches Profil zu einer statistisch signifikanten Verbesserung des Modell-Fits beiträgt. Als zusätzliche Entscheidungskriterien wurden die inhaltliche Interpretierbarkeit der Klassenlösungen, die Größe der einzelnen Profile (keine Klassengrößen < 5%) sowie die Differenzierbarkeit zwischen den einzelnen Profilen beurteilt (vgl. Lubke et al., 2005).

7.1.3.4 Zusammenhang mit professioneller Unterrichtswahrnehmung

Zur Überprüfung des Zusammenhangs zwischen den einzelnen Profilen und der professionellen Unterrichtswahrnehmung im Rahmen der zweiten Teilfragestellung wurden mit der Software Mplus 7.11 (Muthén & Muthén, 1998-2013) hierarchische Modelle mit manifesten Variablen geschätzt. Der Zusammenhang mit der Gesamtfähigkeit professionelle Unterrichtswahrnehmung einerseits und der Zusammenhang mit den drei interkorrelierenden Aspekten Beschreiben, Erklären und Vorhersagen andererseits wurde in zwei getrennten Analyseschritten untersucht. Der Zugang über Strukturgleichungsmodelle bietet gegenüber multiplen Regressionsanalysen unter anderem den Vorteil, mehrere abhängige Variablen simultan aufnehmen und dabei die Interkorrelationen dieser Variablen berücksichtigen zu können.

7.2 Ergebnisse

Ziel der vierten Forschungsfrage ist die Überprüfung des Validitätsaspekts *Externalität.* Zunächst wird die Stichprobe hinsichtlich ihrer Vergleichbarkeit mit der Scaling-up-Stichprobe sowie der PaLea-Gesamtstichprobe überprüft. Anschließend wird die deskriptive Statistik dargestellt. Dann folgen die Ergebnisse der ersten Teilfrage nach der Einordnung Lehramtsstudierender in Subgruppen mit bestimmten Profilen sowie der zweiten Teilfrage nach dem Zusammenhang zwischen der Zugehörigkeit zu diesen Profilen und der professionellen Unterrichtswahrnehmung Lehramtsstudierender.

7.2.1 Überprüfung der Stichprobe

Im Folgenden wird der Vergleich der Stichprobe mit der Scaling-up-Stichprobe sowie der PaLea-Gesamtstichprobe dargestellt.

7.2.1.1 Vergleich mit der Scaling-up-Stichprobe

Einen Überblick über die deskriptive Statistik der Untersuchungsstichprobe im Vergleich zur Scaling-up-Stichprobe bietet Tabelle 20. Die Untersuchungsstichprobe wurde mit der Scaling-up-Stichprobe mittels Chi-Quadrat-Tests, t-Test und einfaktorieller multivariater Varianzanalyse verglichen. Hinsichtlich der Geschlechterverteilung zeigen sich signifikante Unterschiede ($\chi^2(1) = 9.69$, $p < .01$, Cramer's $\nu = .09$), die aufgrund der geringen Effektstärke jedoch zu vernachlässigen sind. Bezüglich des Alters der Studierenden unterscheiden sich die beiden Stichproben signifikant ($t(1184) = 2.02$, $p = .04$, Cohen's $d = 0.17$). Dieser Unterschied ist aufgrund der geringen Effektstärke jedoch ebenfalls zu vernachlässigen. Im Hinblick auf die professionelle Unterrichtswahrnehmung der Studierenden gibt es keine signifikanten Unterschiede für Beschreiben, Erklären und Vorhersagen ($F(3, 1184) = 2.47$, $p = .06$, $\eta_p^2 = .01$).

Tabelle 20

Deskriptive Statistik der Untersuchungsstichprobe im Vergleich zur Scaling-up-Stichprobe

	Untersuchungsstichprobe (N = 159)		Scaling-up-Stichprobe (N = 1029)	
	Anteil (in %)		Anteil (in %)	
Geschlecht weiblich	74.68		61.4	
	M	*SD*	*M*	*SD*
Alter	22.02	2.62	22.56	3.19
Beschreiben	.46	.15	.44	.17
Erklären	.36	.18	.36	.18
Vorhersagen	.42	.23	.43	.24

Anmerkung. Werte für Beschreiben, Erklären und Vorhersagen entsprechen der prozentualen Übereinstimmung mit der Expertennorm.

7.2.1.2 Vergleich mit der PaLea-Gesamtstichprobe

Die Untersuchungsstichprobe wurde mit der PaLea-Gesamtstichprobe mittels Chi-Quadrat-Tests und t-Tests verglichen. Wie Tabelle 19 zu entnehmen ist, verteilen sich die Studierenden der Untersuchungsstichprobe signifikant unterschiedlich ($\chi^2(12) = 527.85$, $p < .001$, Cramer's $\nu = .34$) auf die einzelnen Universitäten. Bezüglich der Geschlechterverteilung unterscheiden sich die beiden Stichproben ($\chi^2(1) = 1.30$, $p = .26$, Cramer's $\nu = .02$) nicht signifikant. Auch hinsichtlich des Alters der Studierenden zeigen sich keine signifikanten Unterschiede ($t(4512) = -.69$, $p = .49$, Cohen's $d = -0.06$). Des Weiteren unterscheidet sich die Untersuchungsstichprobe nicht signifikant von der PaLea-Gesamtstichprobe in Bezug auf die kognitiven Grundfähigkeiten der Studierenden ($t(3450) = -.07$, $p = .95$, Cohen's $d = -0.01$). Ebenso gibt es im Hinblick auf die Variablen Interesse an Bildungswissenschaften ($t(1027) = -1.10$, $p = .27$, Cohen's $d = -0.09$), pädagogisches Interesse ($t(4500) = -.42$, $p = .67$, Cohen's $d = -0.03$), lernbegleitungsorientierter Lehrbegriff ($t(4564) = .88$, $p = .38$, Cohen's $d = 0.07$) und Verträglichkeit ($t(4119) = 1.54$, $p = .16$, Cohen's $d = 0.12$) keine signifikanten Unterschiede. Zusammenfassend ist damit festzuhalten, dass sich die Untersuchungsstichprobe lediglich in der Verteilung der Studierenden auf die Universitäten, aber weder in den Kontrollvariablen Geschlecht, Alter und kognitive Grundfähigkeiten noch in den Variablen für die Profilbildung (Interesse an Bildungswissenschaften, pädagogisches Interesse, lernbegleitungsorientierter Lernbegriff und Verträglichkeit) von der PaLea-Gesamtstichprobe unter-

scheidet. Eine Gegenüberstellung der deskriptiven Statistik beider Stichproben ist in Tabelle 21 zu finden.

Tabelle 21

Deskriptive Statistik der Untersuchungsstichprobe im Vergleich zur PaLea-Gesamtstichprobe

	Untersuchungsstichprobe (N = 159)		PaLea-Gesamtstichprobe (N = 4468)	
	Anteil (in %)		Anteil (in %)	
Geschlecht weiblich	74.68		70.48	
	M	*SD*	*M*	*SD*
Alter	21.39	2.79	21.20	3.37
Pädagogisches Interesse	3.47	0.43	3.46	0.44
Interesse an Bildungswissenschaften	2.72	0.73	2.63	0.67
Lernbegleitungsorientierter Lehrbegriff	3.40	0.39	3.42	0.40
Verträglichkeit	3.54	0.47	3.53	0.47
Kognitive Grundfähigkeiten	20.72	2.97	20.70	3.35

7.2.1.3 Weitere Überprüfung der Stichprobe

Nachdem sich die Untersuchungsstichprobe zu 47.8 % aus Studierenden des Standorts L zusammensetzt, wird weiter überprüft, inwieweit sich diese Studierenden von Studierenden anderer Universitätsstandorte unterscheiden. Dazu wurden t-Tests und Chi-Quadrat-Tests berechnet. Bezüglich der Geschlechterverteilung gibt es keine signifikanten Unterschiede ($\chi^2(1) = 1.20$, $p = .27$, Cramer's $\nu = .09$). Auch hinsichtlich des Alters ($t(156) = -1.52$, $p = .13$, Cohen's $d = 0.24$) und der kognitiven Grundfähigkeiten ($t(118) = 1.18$, $p = .24$, Cohen's $d = 0.19$) zeigen sich keine signifikanten Unterschiede zwischen den Studierendengruppen. Des Weiteren unterscheiden sich Studierende des Standorts L nicht signifikant von Studierenden anderer Universitätsstandorte in Bezug auf das Interesse an Bildungswissenschaften ($t(72) = -.78$, $p = .44$, Cohen's $d = -0.12$) und der Verträglichkeit ($t(142) = -.03$, $p = .97$, Cohen's $d = -0.00$). Jedoch weisen Studierende des Standorts L signifikant niedrigere Werte bei den Variablen pädagogisches Interesse ($t(125) = 3.46$, $p < .01$, Cohen's $d = 0.55$) und lernbe-

gleitungsorientierter Lehrbegriff ($t(157) = 3.94$, $p < .01$, Cohen's $d = 0.63$) auf. Die deskriptive Statistik der beiden Studierendengruppen im Vergleich ist Tabelle 22 zu entnehmen.

	M	*SD*	(1)	(2)	(3)	(4)

Tabelle 22

Deskriptive Statistik zu Messzeitpunkt I für Studierende des Standorts L im Vergleich zu Studierenden der anderen Universitätsstandorte

	Studierende des Standorts L (n = 76)		Studierende anderer Universitätsstandorte (n = 83)	
	Anteil (in %)		Anteil (in %)	
Geschlecht weiblich	78.67		71.08	
	M	*SD*	*M*	*SD*
Alter	21.75	2.49	21.07	3.01
Pädagogisches Interesse	3.36	0.51	3.59	0.32
Interesse an Bildungswissenschaften	2.79	0.66	2.65	0.79
Lernbegleitungsorientierter Lehrbegriff	3.27	0.40	3.51	0.35
Verträglichkeit	3.53	0.50	3.53	0.45
Kognitive Grundfähigkeiten	20.34	2.99	20.99	2.95

7.2.2 Deskriptive Statistik der Stichprobe

Einen Überblick über die deskriptive Statistik der für die Profilbildung relevanten Variablen sowie professioneller Unterrichtswahrnehmung und ihrer drei Aspekte bietet

Pädagogisches Interesse (1)	3.47	0.43				
Interesse Bildungswissenschaften (2)	2.68	0.73	.31*			
Lernbegleitungsorientierter Lehrbegriff (3)	3.40	0.40	.34*	.13		
Verträglichkeit (4)	3.54	0.47	.33*	.07	.19*	
Professionelle Unterrichtswahrnehmung (5)	.42	.17	.17*	.17*	.27*	.28*
Beschreiben (6)	.46	.15	.05	.16	.27*	.17*
Erklären (7)	.36	.18	-.04	.12	.17*	.24*
Vorhersagen (8)	.42	.23	-.05	.10	.18*	.26*

	M	*SD*	(1)	(2)	(3)	(4)
Pädagogisches Interesse (1)	3.47	0.43				
Interesse Bildungswissenschaften (2)	2.68	0.73	.31*			
Lernbegleitungsorientierter Lehrbegriff (3)	3.40	0.40	.34*	.13		
Verträglichkeit (4)	3.54	0.47	.33*	.07	.19*	
Professionelle Unterrichtswahrnehmung (5)	.42	.17	.17*	.17*	.27*	.28*

Tabelle 23. Die Lehramtsstudierenden erreichen auf den Skalen pädagogisches Interesse, lernbegleitungsorientierter Lehrbegriff und Verträglichkeit durchschnittlich hohe Werte bei gleichzeitig niedrigen Standardabweichungen ($SD < .47$). Im Gegensatz dazu ist das Interesse an Bildungswissenschaften mit $M = 2.68$ mittelmäßig ausgeprägt und weist eine höhere Standardabweichung ($SD = .73$) auf. Hinsichtlich professioneller Unterrichtswahrnehmung zeigen sich mittlere Übereinstimmungen mit der Expertennorm, die durchschnittlich für Beschreiben am höchsten ($M = .46$, $SD = .15$) und Erklären am niedrigsten ($M = .36$, $SD = .18$) ausfallen. Zusätzlich enthält

Beschreiben (6)	.46	.15	.05	.16	.27*	.17*
Erklären (7)	.36	.18	-.04	.12	.17*	.24*
Vorhersagen (8)	.42	.23	-.05	.10	.18*	.26*

Tabelle 23 die Reliabilität der für die Profilbildung relevanten Variablen und professioneller Unterrichtswahrnehmung mit ihren drei Aspekten sowie einen Überblick über die Interkorrelationen dieser Variablen.

Tabelle 23

Deskriptive Statistik und Korrelationen zwischen den Skalen

	M	*SD*	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Pädagogisches Interesse (1)	3.47	0.43							
Interesse Bildungswissenschaften (2)	2.68	0.73	.31*						
Lernbegleitungsorientierter Lehrbegriff (3)	3.40	0.40	.34*	.13					
Verträglichkeit (4)	3.54	0.47	.33*	.07	.19*				
Professionelle Unterrichts-wahrnehmung (5)	.42	.17	.17*	.17*	.27*	.28*			
Beschreiben (6)	.46	.15	.05	.16	.27*	.17*	.91*		
Erklären (7)	.36	.18	-.04	.12	.17*	.24*	.95*	.80*	
Vorhersagen (8)	.42	.23	-.05	.10	.18*	.26*	.96*	.79*	.88*

Anmerkungen. Werte aus *m* = 50 imputierten Datensätzen kombiniert. Werte für professionelle Unterrichtswahrnehmung entsprechen der prozentualen Übereinstimmung mit der Expertennorm.

7.2.3 Profile Lehramtsstudierender

Im Rahmen der latenten Profilanalyse wurden Modelle mit 1 – 5 latenten Profilen spezifiziert. Ein Überblick über die Modell-Fit-Indizes der fünf Profil-Lösungen sowie die Anzahl einzelner Profile mit weniger als 5 % der Untersuchungsstichprobe ist Tabelle 24 zu entnehmen.

Tabelle 24

Modell-Fit-Indizes für verschiedene Profil-Lösungen sowie die Anzahl einzelner Profile < 5 % der Untersuchungsstichprobe

Profile	Parameter	AIC	BIC	aBIC	< 5 %
1	8	1.805.526	1.830.026	1.804.703	0
2	13	1.731.385	1.771.199	1.730.048	0
3	18	1.638.647	1.693.774	1.636.795	0
4	23	1.632.623	1.703.062	1.630.256	1
5	34	1.326.993	1.431.121	1.323.495	2

Anmerkungen. N = 159 Lehramtsstudierende. m = 50 imputierte Datensätze. AIC = Aikakes Informationskriterium; BIC = Bayessches Informationskriterium; aBIC = stichproben-angepasstes Bayessches Informationskriterium.

Alle drei Informationskriterien (AIC, BIC und aBIC) nehmen bis zur 3-Profil-Lösung stetig ab. Von der 3- zur 4-Profil-Lösung verändern sich die Informationskriterien nur minimal. Eine 5-Profil-Lösung weist zwar die niedrigsten Informationskriterien auf, aber gleichzeitig auch zwei Profile, die weniger als 5 % der Lehramtsstudierenden repräsentieren. Vergleicht man die z-standardisierten Mittelwerte der Profile bei einer 3- und 4-Profil-Lösung, fällt Folgendes auf: In der 4-Profil-Lösung repräsentiert ein Profil lediglich 0.5 % aller Lehramtsstudierenden und stellt eine Extremgruppe eines Profils der 3-Profil-Lösung dar. Basierend auf diesen statistischen und inhaltlichen Entscheidungskriterien und unter Berücksichtigung des Parsimonitätsprinzips (vgl. Collins & Lanza., 2010) ist die 3-Profil-Lösung zu favorisieren. Die N = 159 Lehramtsstudierenden verteilen sich wie folgt auf die drei latenten Profile: Profil 1 (34.9 %), Profil 2 (20.9 %) und Profil 3 (44.2 %). Die z-standardisierten Mittelwerte der drei Profile sind in Abbildung 6 dargestellt. Lehramtsstudierende im Profil 1 weisen unterdurchschnittliche Werte vor allem für pädagogisches In-

teresse und Verträglichkeit auf, wohingegen Lehramtsstudierende im Profil 3 überdurchschnittliche Werte vor allem auf diesen beiden Variablen aufweisen. Lehramtsstudierende im Profil 2 zeichnen sich durch durchschnittliche Ausprägungen auf allen vier Variablen aus. Es fällt allerdings auf, dass sich die einzelnen Profile lediglich quantitativ, vor allem in der Ausprägung bezüglich Verträglichkeit, unterscheiden, jedoch keine qualitativen Unterschiede aufweisen, die sich graphisch in überschneidenden Linien ausdrücken.

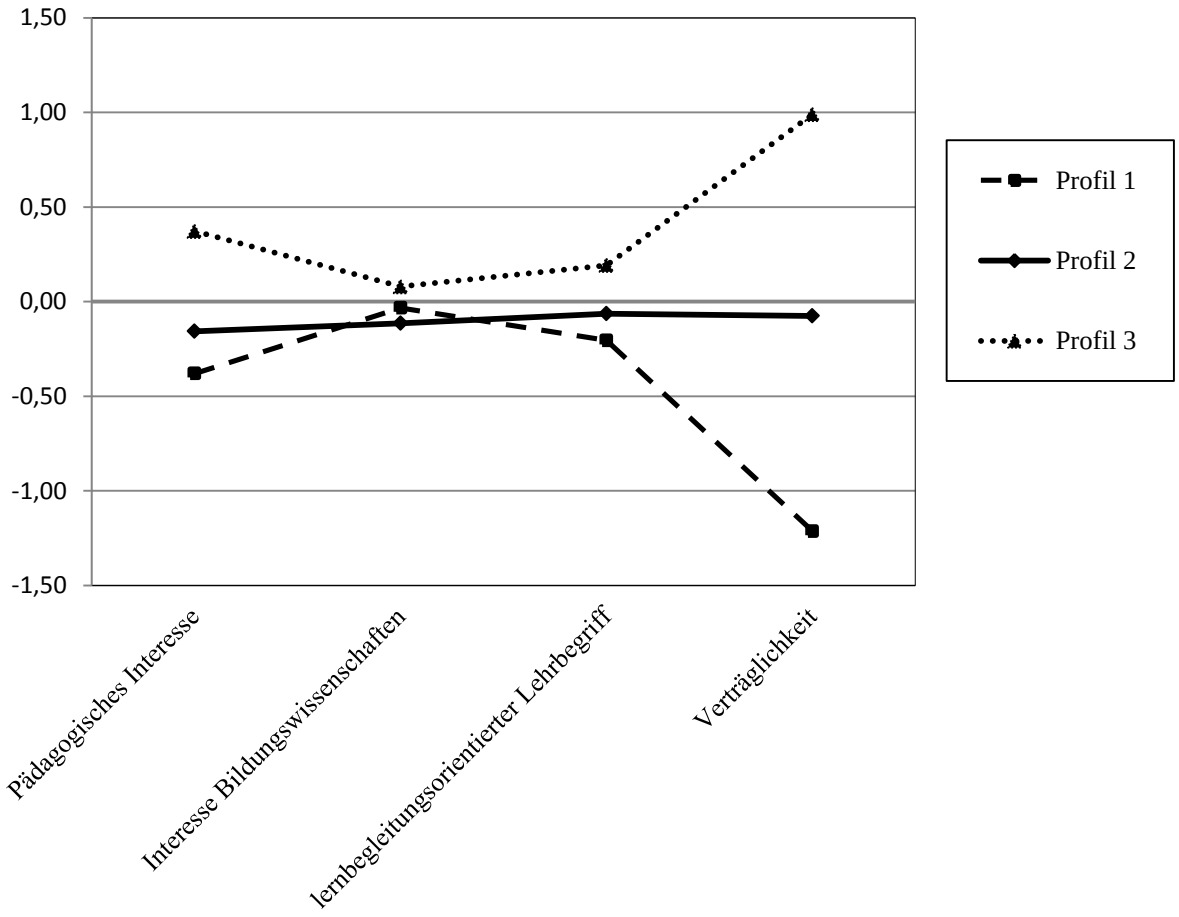

Abbildung 6. z-standardisierte Mittelwerte für pädagogisches Interesse, Interesse an Bildungswissenschaften, lernbegleitungsorientierter Lehrbegriff und Verträglichkeit, differenziert nach den drei Profilen.

Damit ist zu konstatieren, dass die latenten Profilanalysen mit den Variablen pädagogisches Interesse, Interesse an Bildungswissenschaften, lernbegleitungsorientierter Lehrbegriff und Verträglichkeit zu keinen latenten Profilen mit qualitativ unterschiedlichen Mustern über die betrachteten Variablen hinweg führen. Die qualitative Differenzierungsfähigkeit ist jedoch per Definition eine Voraussetzung für latente Klassen oder Profile (vgl. Gollwitzer, 2012). Demzufolge werden die weiteren Analysen zur Untersuchung des Zusammenhangs mit professioneller Unterrichtswahrnehmung auf Basis der einzelnen Variablen durchgeführt (vgl. van Eye & Spiel, 2010). In zwei getrennten Analyseschritten wird der Zusammenhang der manifesten Variablen pädagogisches Interesse, Interesse an Bildungswissenschaften, lernbegleitungsorientierter Lehrbegriff und Verträglichkeit mit der Gesamtfähigkeit professionelle Unterrichtswahrnehmung sowie den drei interkorrelierenden Aspekten Beschreiben, Erklären und Vorhersagen mit Hilfe von hierarchischen Modellen untersucht.

7.2.4 Zusammenhang mit professioneller Unterrichtswahrnehmung

In einem ersten Analyseschritt wurde der Zusammenhang mit der Gesamtfähigkeit professionelle Unterrichtswahrnehmung untersucht. Die Ergebnisse der hierarchischen Regression zur Vorhersage professioneller Unterrichtswahrnehmung durch die einzelnen Variablen pädagogisches Interesse, Interesse an Bildungswissenschaften, lernbegleitungsorientierter Lehrbegriff und Verträglichkeit sind in Tabelle 25 dargestellt. Pädagogisches Interesse (ß = -.02, p = .84) und Interesse an Bildungswissenschaften (ß = .13, p = .17) sagen professionelle Unterrichtswahrnehmung nicht signifikant vorher, sobald die Variablen lernbegleitungsorientierte Lehrbegriff und Verträglichkeit zusätzlich in das Modell aufgenommen werden. Die Variablen lernbegleitungsorientierter Lehrbegriff (ß = .21, p = .01) und Verträglichkeit (ß = .24, p < .001) erweisen sich jedoch als signifikante Prädiktoren für professionelle Unterrichtswahrnehmung. Modell 4 erklärt mit 15 % am meisten Varianz.

Tabelle 25

Hierarchische Regression zur Vorhersage professioneller Unterrichtswahrnehmung

Variablen	*B*	*SEB*	ß	R^2
Modell 1				
Pädagogisches Interesse	.07	.03	.17*	.03
Modell 2				
Pädagogisches Interesse	.05	.03	.14	.03
Interesse an Bildungswissenschaften	.03	.03	.13	
Modell 3				
Pädagogisches Interesse	.02	.03	.06	
Interesse an Bildungswissenschaften	.10	.04	.12*	.10
Lernbegleitungsorientierter Lehrbegriff	.10	.04	.23*	
Modell 4				
Pädagogisches Interesse	- .01	.04	-.02	
Interesse an Bildungswissenschaften	.03	.02	.13	
Lernbegleitungsorientierter Lehrbegriff	.09	.04	.21*	.15
Verträglichkeit	.09	.03	.24*	

Anmerkungen. $N = 159$. *B* = unstandardisierter Regressionskoeffizient; *SEB* = Standardfehler von *B*; ß = standardisierter Regressionskoeffizient (stdyx); R^2 = Determinationskoeffizient. * $p < .05$.

In einem zweiten Analyseschritt wurde der Zusammenhang mit den drei interkorrelierenden Aspekten professioneller Unterrichtswahrnehmung untersucht. Die hierarchische Regression zur Vorhersage von Beschreiben (vgl.

Tabelle 26), Erklären (vgl. Tabelle 27) und Vorhersagen (vgl. Tabelle 28) zeigt ein ähnliches Bild wie für die Gesamtfähigkeit professionelle Unterrichtswahrnehmung. Der lernbegleitungsorientierte Lehrbegriff und Verträglichkeit stellen signifikante Prädiktoren dar. Allerdings fällt die Varianzaufklärung mit jeweils 12 % für Erklären und Vorhersagen im Vergleich zur Gesamtfähigkeit professionelle Unterrichtswahrnehmung etwas geringer und für Beschreiben mit 17 % etwas höher aus.

Tabelle 26

Hierarchische Regression zur Vorhersage des Aspekts Beschreiben

Variablen	*B*	*SEB*	ß	R^2
Modell 1				
Pädagogisches Interesse	.08	.03	.24*	.06
Modell 2				
Pädagogisches Interesse	.07	.03	.19*	.08
Interesse an Bildungswissenschaften	.03	.02	.16	
Modell 3				
Pädagogisches Interesse	.03	.03	.10	.15
Interesse an Bildungswissenschaften	.03	.02	.15	
Lernbegleitungsorientierter Lehrbegriff	.11	.03	.27*	
Modell 4				
Pädagogisches Interesse	.02	.03	.05	.17
Interesse an Bildungswissenschaften	.03	.02	.16	
Lernbegleitungsorientierter Lehrbegriff	.10	.03	.26*	
Verträglichkeit	.05	.03	.16*	

Anmerkungen. $N = 159$. *B* = unstandardisierter Regressionskoeffizient; *SEB* = Standardfehler von *B*; ß = standardisierter Regressionskoeffizient (stdyx); R^2 = Determinationskoeffizient. * $p < .05$.

Tabelle 27

Hierarchische Regression zur Vorhersage des Aspekts Erklären

Variablen	*B*	*SEB*	ß	R^2
Modell 1				
Pädagogisches Interesse	.05	.03	.13	.02
Modell 2				
Pädagogisches Interesse	.04	.04	.10	
Interesse an Bildungswissenschaften	.03	.03	.11	.03
Modell 3				
Pädagogisches Interesse	.01	.04	.04	
Interesse an Bildungswissenschaften	.03	.02	.11	.06
Lernbegleitungsorientierter Lehrbegriff	.08	.04	.19*	
Modell 4				
Pädagogisches Interesse	-.02	.04	-.04	
Interesse an Bildungswissenschaften	.03	.02	.12	
Lernbegleitungsorientierter Lehrbegriff	.07	.04	.17*	.12
Verträglichkeit	.09	.03	.24*	

Anmerkungen. $N = 159$. *B* = unstandardisierter Regressionskoeffizient; *SEB* = Standardfehler von *B*; ß = standardisierter Regressionskoeffizient (stdyx); R^2 = Determinationskoeffizient. * $p < .05$.

Tabelle 28

Hierarchische Regression zur Vorhersage des Aspekts Vorhersagen

Variablen	*B*	*SEB*	ß	R^2
Modell 1				
Pädagogisches Interesse	.07	.04	.13	.02
Modell 2				
Pädagogisches Interesse	.05	.05	.10	.03
Interesse an Bildungswissenschaften	.03	.03	.09	
Modell 3				
Pädagogisches Interesse	.02	.05	.03	.06
Interesse an Bildungswissenschaften	.03	.03	.09	
Lernbegleitungsorientierter Lehrbegriff	.12	.05	.20*	
Modell 4				
Pädagogisches Interesse	-.03	.05	-.05	.12
Interesse an Bildungswissenschaften	.03	.03	.10	
Lernbegleitungsorientierter Lehrbegriff	.10	.05	.18*	
Verträglichkeit	.13	.04	.26*	

Anmerkungen. $N = 159$. *B* = unstandardisierter Regressionskoeffizient; *SEB* = Standardfehler von *B*; ß = standardisierter Regressionskoeffizient (stdyx); R^2 = Determinationskoeffizient. * $p < .05$.

Zusammenfassend kann festgehalten werden, dass die Variablen lernbegleitungsorientierter Lehrbegriff und Verträglichkeit signifikante Prädiktoren für professionelle Unterrichtswahrnehmung und ihre Aspekte Beschreiben, Erklären und Vorhersagen darstellen.

7.2.5 Zusammenfassung der Ergebnisse

Die latenten Profilanalysen mit den Variablen pädagogisches Interesse, Interesse an Bildungswissenschaften, lernbegleitungsorientierter Lehrbegriff und Verträglichkeit im Rahmen der ersten Teilfrage führen zu keinen qualitativ unterschiedlichen latenten Profilen. Folglich ist es nicht möglich, Lehramtsstudierende in Subgruppen mit bestimmten Profilen einzuordnen. Die hierarchischen Regressionsanalysen auf Variablenebene im Rahmen der zweiten Teilfrage zeigen, dass die Variablen lernbegleitungsorientierter Lehrbegriff und Verträglichkeit signifikante Prädiktoren für professionelle Unterrichtswahrnehmung und ihre drei Aspekte darstellen. Demzufolge gibt es einen Zusammenhang zwischen einem lernbegleitungsorientiertem Lehrbegriff und dem Persönlichkeitsmerkmal Verträglichkeit einerseits und der professionellen Unterrichtswahrnehmung Lehramtsstudierender andererseits.

7.3 Diskussion

Mit Blick auf einen großflächigen Einsatz des Tools Observer zur Erfassung professioneller Unterrichtswahrnehmung im Verlauf der universitären Lehrerbildung muss der Einsatz des Instruments über das Abbilden von integrierten Wissensstrukturen hinaus gerechtfertigt werden. Dafür ist ein Zusammenhang zum späteren professionellen Handeln im Unterricht von Bedeutung. Eine weitere Möglichkeit eines großflächigen Einsatzes des Instruments stellt die Nutzung zur Selbstexploration im Rahmen der Studienwahl dar. Dafür ist es erforderlich, dass ein Zusammenhang mit dem Studienerfolg besteht. Deshalb wurde im Rahmen der vierten Forschungsfrage untersucht, inwieweit professionelle Unterrichtswahrnehmung Lehramtsstudierender mit Variablen, die typischerweise als Prädiktoren für Studien- und Berufserfolg herangezogen werden, zusammenhängt. Damit wird der Validitätsaspekt *Externalität* geprüft. In der ersten Teilfrage wurde untersucht, inwieweit sich Lehramtsstudierende in Subgruppen hinsichtlich pädagogischen Interesses, Interesses an den Bildungswissenschaften, lernbegleitungsorientierten Lehrbegriffs und Verträglichkeit gruppieren lassen. In der zweiten Teilfrage wurde überprüft, inwieweit ein Zusammenhang zwischen der Zugehörigkeit zu diesen Profilen und der professionellen Unterrichtswahrnehmung Lehramtsstudierender besteht. Im Folgenden werden zunächst zentrale Befunde zusammengefasst und inhaltlich diskutiert. Anschließend wird das methodische Vorgehen bewertet und weitere Implikationen werden dargestellt.

7.3.1 Zusammenfassung und inhaltliche Diskussion zentraler Befunde

Zur Überprüfung der ersten Teilfragestellung wurde ein personenzentrierter Ansatz gewählt. Es wurde eine latente Profilanalyse mit den Variablen pädagogisches Interesse, Interesse an Bildungswissenschaften, lernbegleitungsorientierter Lehrbegriff und Verträglichkeit durchgeführt. Entgegen der Hypothese konnten keine qualitativ unterschiedlichen Profile identifiziert werden. Dieser Befund deutet darauf hin, dass sich die Lehramtsstudierenden zwar quantitativ in den einzelnen Variablen unterscheiden, aber keine heterogenen Subgruppen mit qualitativ unterschiedlichen Profilen existieren. Dementsprechend kann durch variablenzentrierte Analysen der Stichprobe auf die Gesamtpopulation geschlossen werden (vgl. Pastor et al., 2007). Dieser Befund widerspricht den Ergebnissen anderer Studien mit Lehramtsstudierenden, bei denen heterogene Subgruppen identifiziert wurden. Allerdings wurden bei diesen Studien andere Variablen zur Profilbildung herangezogen,

nämlich Studienwahlmotivation (Billich-Knapp et al., 2012) und Einstellung zur Berufswahl (Watt & Richardson, 2008).

Für die zweite Teilfragestellung wurden basierend auf den Mittelwerten der Variablen pädagogisches Interesse, Interesse an Bildungswissenschaften, lernbegleitungsorientierter Lehrbegriff und Verträglichkeit hierarchische Modelle geschätzt. Der Zusammenhang mit professioneller Unterrichtswahrnehmung und den drei Aspekten Beschreiben, Erklären und Vorhersagen wurde in zwei getrennten Analyseschritten untersucht. Entgegen der Hypothese stellen nicht alle vier Variablen, sondern lediglich ein lernbegleitungsorientierter Lehrbegriff und das Persönlichkeitsmerkmal Verträglichkeit signifikante Prädiktoren für professionelle Unterrichtswahrnehmung dar. Diese zwei Variablen erklären zusammen 15 % Varianz. Für die Aspekte Beschreiben, Erklären und Vorhersagen zeigt sich ein ähnliches Bild. Allerdings fällt die Varianzaufklärung mit jeweils 12 % für Erklären und Vorhersagen im Vergleich etwas geringer und für Beschreiben mit 17 % etwas höher aus. Das Ausbleiben des Zusammenhangs zwischen pädagogischem Interesse und professioneller Unterrichtswahrnehmung könnte darauf zurückzuführen sein, dass pädagogisches Interesse sehr allgemein definiert ist und sich auf die generelle Tätigkeit mit Kindern und Jugendlichen bezieht. Dies entspricht nicht der sehr spezifischen Ebene, auf der professionelle Unterrichtswahrnehmung mit dem Wissensfokus Zielorientierung, Lernbegleitung und Lernatmosphäre erhoben wird. Das Interesse an Bildungswissenschaften ist zwar im Vergleich zum pädagogischen Interesse enger gefasst, jedoch ist es fraglich, inwieweit die Studierenden zum Zeitpunkt der Datenerhebung, am Ende ihres ersten Semesters, bereits eine konkrete Vorstellung von Inhalten der Bildungswissenschaften hatten. Die große Anzahl der fehlenden Angaben (53.8 %) deutet darauf hin, dass es den Studierenden schwer fiel, die Items zu beantworten. Folglich sollte der Zusammenhang zwischen Interesse an Bildungswissenschaften und professioneller Unterrichtswahrnehmung zu einem fortgeschritteneren Zeitpunkt im Lehramtsstudium erneut überprüft werden. Zwar konnte in einer Studie mit einer Teilstichprobe der PaLea-Gesamtstichprobe von $N = 1169$ Studierenden (Rösler et al., 2013) bis zum Ende des vierten Semesters im Mittel keine Veränderung des Interesses an Bildungswissenschaften nachgewiesen werden, jedoch zeigte sich bei Studierenden für Lehramt Gymnasium ein Zuwachs über die Zeit. Diese Befunde sprechen dafür, bei der Überprüfung des Zusammenhangs zwischen Interesse an Bildungswissenschaften und professioneller Unterrichtswahrnehmung den Lehramtsstudiengang in das Modell aufzunehmen.

Die Variablen lernbegleitungsorientierter Lehrbegriff und Verträglichkeit erklären zusammen 15 % Varianz professioneller Unterrichtswahrnehmung. In Bezug auf einen lernbegleitungsorientierten Lehrbegriff entspricht dieser hypothesenkonforme Befund Ergebnis-

	M	*SD*	(1)	(2)	(3)	(4)

sen aus der Forschung zu Überzeugungen, die zeigen, dass Überzeugungen einen Filter darstellen, durch den die Umwelt wahrgenommen und interpretiert wird (vgl. Pajares, 1992). Hinsichtlich Verträglichkeit wird die Hypothese bestätigt, dass dieses Persönlichkeitsmerkmal positiv mit professioneller Unterrichtswahrnehmung zusammenhängt. Im Vergleich zu anderen Persönlichkeitsmerkmalen liegen zu Verträglichkeit die wenigsten Einzelstudien vor und die Befundlage bezüglich Studien- und Berufserfolg ist uneinheitlich (vgl. Foerster, 2008). Die Ergebnisse der vorliegenden Arbeit deuten darauf hin, dass dieses Persönlichkeitsmerkmal spezifisch für den Prozess der Wahrnehmung und Interpretation von Unterricht relevant ist. Die gemeinsame prädiktive Kraft der Variablen lernbegleitungsorientierter Lehrbegriff und Verträglichkeit ist angesichts der Befunde von Stürmer und Kollegen (eingereicht) durchaus als relevant einzuschätzen. In dieser Studie werden 24 % Varianz der professionellen Unterrichtswahrnehmung durch die Anzahl der besuchten Lehrveranstaltungen zum Thema “Lehren und Lernen” sowie das Interesse an derartigen Inhalten aufgeklärt (Stürmer et al., eingereicht). Es sollte in einer weiteren Studie geprüft werden, ob und in welchem Ausmaß die Variablen lernbegleitungsorientierter Lehrbegriff und Verträglichkeit zusätzlich zur Varianzaufklärung beitragen.

Der Befund, dass die Varianzaufklärung für den Aspekt Beschreiben vergleichsweise am höchsten ausfällt, könnte darauf zurückzuführen sein, dass dieser Aspekt bei den Studierenden der vorliegenden Stichprobe, die sich zu Beginn ihres Lehramtsstudiums befinden, bereits am stärksten ausgeprägt ist (vgl.

Pädagogisches Interesse (1)	3.47	0.43				
Interesse Bildungswissenschaften (2)	2.68	0.73	.31*			
Lernbegleitungsorientierter Lehrbegriff (3)	3.40	0.40	.34*	.13		
Verträglichkeit (4)	3.54	0.47	.33*	.07	.19*	
Professionelle Unterrichts-wahrnehmung (5)	.42	.17	.17*	.17*	.27*	.28*
Beschreiben (6)	.46	.15	.05	.16	.27*	.17*
Erklären (7)	.36	.18	-.04	.12	.17*	.24*
Vorhersagen (8)	.42	.23	-.05	.10	.18*	.26*

Tabelle 23). Allerdings zeigte sich in der Studie von Stürmer und Kollegen (eingereicht), die ebenfalls mit Studierenden im zweiten oder dritten Semester durchgeführt wurde, die größte Varianzaufklärung für den Aspekt Erklären. Diese unterschiedlichen Befunde weisen auf differenzielle Effekte verschiedener Variablen auf einzelne Aspekte professioneller Unterrichtswahrnehmung hin. Dies sollte in weiteren Studien systematisch geprüft werden.

7.3.2 Methodische Überlegungen

Im Rahmen der Überprüfung der vierten Forschungsfrage konnten keine Subgruppen mit unterschiedlichen Profilen für die Variablen pädagogisches Interesse, Interesse an Bildungswissenschaften, lernbegleitungsorientierter Lehrbegriff und Verträglichkeit identifiziert werden. Allerdings zeigte sich in der zweiten Teilfrage auf Variablenebene, dass ein lernbegleitungsorientierter Lehrbegriff und Verträglichkeit signifikante Prädiktoren für professionelle Unterrichtswahrnehmung und ihre Aspekte Beschreiben, Erklären und Vorhersagen darstellen. Im Hinblick auf die Interpretation der Befunde muss jedoch auf folgende zwei methodische Einschränkungen hingewiesen werden: (1) geringe Varianz in den Skalen und (2) unausgewogene Zusammensetzung der Stichprobe.

Der erste Punkt betrifft die geringe Varianz in den Skalen der Variablen pädagogisches Interesse, lernbegleitungsorientierter Lehrbegriff und Verträglichkeit. Ein vergleichbares Bild zeigt sich in der PaLea-Gesamtstichprobe (vgl. Abschnitt 7.2.1.2), was gegen eine Verzerrung der Untersuchungsstichprobe spricht. Die durchweg sehr hohen Mittelwerte in beiden Stichproben könnten darauf hindeuten, dass die Studierenden sozial erwünscht antworten, da sie sich bereits für ein Lehramtsstudium entschieden haben. Aufgrund geringer Varianzen werden potentielle Zusammenhänge mit anderen Variablen unterschätzt (Furr & Bacharach, 2008). Eine Möglichkeit, um eventuell mehr Varianz zu erhalten, wäre eine Analyse auf Item- statt auf Skalenebene, da durch das Bilden von Skalenmittelwerten Information verloren geht. Ein derartiges Vorgehen war in der vorliegenden Studie aufgrund der geringen Stichprobengröße nicht möglich. Deshalb sollte eine Studie mit einer ausreichend großen Stichprobe durchgeführt werden, um eine latente Profilanalyse und gegebenenfalls eine hierarchische Regression auf Basis von einzelnen Items zu berechnen.

Der zweite Punkt betrifft die Zusammensetzung der Stichprobe. Zum einen können aufgrund der ungleichen Verteilung auf verschiedene Lehramtsstudiengänge Effekte des Lehramtsstudiengangs nicht systematisch überprüft werden. Zum anderen stammen 47.8 % der Studierenden aus einem Hochschulstandort. Diese Studierenden zeigen niedrigeres pädagogisches Interesse und einen weniger ausgeprägten lernbegleitungsorientierten Lehrbegriff auf. Deshalb sollten Effekte des Hochschulstandorts systematisch überprüft werden. Dies ist durch das vorliegende Design allerdings nicht möglich und sollte demnach in einer weiteren Studie untersucht werden.

7.3.3 Implikationen

Trotz der diskutierten methodischen Einschränkungen wurde im Rahmen der Überprüfung der vierten Forschungsfrage in einem gewissen Ausmaß Evidenz für den Validitätsaspekt *Externalität* gefunden. Werden die Befunde, dass professionelle Unterrichtswahrnehmung mit einem lernbegleitungsorientierten Lehrbegriff sowie dem Persönlichkeitsmerkmal Verträglichkeit zusammenhängt, durch weitere Studien gestützt, ergeben sich daraus drei praktische Implikationen für die Lehrerbildung. Zum einen sollte die Vorstellung von Lehren im Rahmen der universitären Lehrerbildung thematisiert und der Nutzen eines lernbegleitungsorientierten Unterrichtens aufgezeigt werden. Denn es gilt zu bedenken, dass die Studierenden bereits mit sehr konkreten und stabilen Vorstellungen von Lehren und Lernen an die Universität kommen, die im Verlauf ihrer Schulzeit geprägt wurden und nicht zwingend einer Gestaltung von Lerngelegenheiten zur Anregung individueller Lernprozesse entsprechen (Reusser et al., 2011). Um zukünftige Lehrpersonen auszubilden, die nicht den Unterricht ihrer eigenen Schulzeit unreflektiert kopieren, sollte in einem Lehramtsstudium Gelegenheit zur angeleiteten Reflexion der eigenen Vorstellung von Lehren geschaffen werden. Zum anderen deutet der Zusammenhang zwischen professioneller Unterrichtswahrnehmung und dem Persönlichkeitsmerkmal Verträglichkeit darauf hin, dass es Studierende gibt, die aufgrund ihrer individuellen Voraussetzungen einen erhöhten Beratungs- und Förderbedarf haben. Vor diesem Hintergrund gewinnt das Angebot zusätzlicher Lerngelegenheiten an Bedeutung. Darüber hinaus ist der Befund zu beachten, dass 53.8 % der Studierenden am Ende des ersten Semesters das Interesse an Bildungswissenschaften nicht eingeschätzt haben. Dies könnte auf ein mangelndes Bewusstsein für die Relevanz der bildungswissenschaftlichen Studieninhalte hinweisen. Dieser Eindruck entspricht den Ergebnissen anderer Studien (z. B. Rösler et al., 2013) und verdeutlicht, dass es notwendig ist, die Bedeutung bildungswissenschaftlicher Studieninhalte für den Professionalisierungsprozess zukünftiger Lehrpersonen herauszustellen.

Im Hinblick auf mögliche Implikationen muss auch kritisch diskutiert werden, dass nur zwei der vier ausgewählten Prädiktoren für Studien- und Berufserfolg einen Zusammenhang mit professioneller Unterrichtswahrnehmung aufweisen. Dieser Befund kann auf zwei Arten interpretiert werden. Die erste Interpretation bezieht sich auf das Instrument und die zweite Interpretation auf die verwendeten Prädiktoren. Bezogen auf das Instrument weist die begrenzte Evidenz im Hinblick auf den Validitätsaspekt *Externalität* darauf hin, dass die Ergebnisse des Tools Observer im Hinblick auf einen Zusammenhang mit Stu-

dien- und Berufserfolg mit Zurückhaltung interpretiert werden sollten. Die vorliegenden Befunde werfen allerdings auch die Frage auf, inwiefern die ausgewählten Variablen angemessene externale Kriterien für die Validität der Interpretation der Messergebnisse darstellen. Die Auswahl und die Erfassung dieser Kriterien sind für die Überprüfung des Validitätsaspekts *Externalität* von zentraler Bedeutung (AERA et al., 1999). Diesbezüglich müssen zwei Punkte kritisch diskutiert werden. Zum einen ist es neben einem theoretisch vermuteten Zusammenhang der Kriterien mit der erfassten Kompetenz essentiell, dass die Kriterien selbst valide interpretiert werden können (Furr & Bacharach, 2008). Interessensskalen, wie sie in der vorliegenden Arbeit verwendet wurden, werden typischerweise im Rahmen von Self-Assessments zur Abklärung der Studien- und Berufseignung von angehenden Lehramtsstudierenden eingesetzt (vgl. Herlt & Schaarschmidt, 2007; Nieskens & Müller, 2009). Im Hinblick auf den Zusammenhang dieser Self-Assessments mit dem tatsächlichen Studien- und Berufserfolg Studierender existiert jedoch noch keine ausreichende Evidenzgrundlage (Köller et al., 2012). Die Befunde der vorliegenden Arbeit sprechen dafür, die Validität dieser Self-Assessments weiter zu prüfen. Zum anderen gilt es kritisch zu hinterfragen, ob grundsätzlich Prädiktoren für Studien- und Berufserfolg, unabhängig von ihrer Validität, geeignete Kriterien darstellen, um die *Externalität* eines Instruments zur Erfassung professioneller Unterrichtswahrnehmung zu überprüfen. Im Kontext der Kompetenzerfassung von (zukünftigen) Lehrpersonen ist letzten Endes die praktische Relevanz für das Handeln entscheidend (Kunter & Klusmann, 2010). Wie in Abschnitt 2.2.1 ausgeführt, stellt die professionelle Unterrichtswahrnehmung eine wichtige Voraussetzung für das professionelle Handeln im Unterricht dar. Deshalb sollte das unterrichtliche Handeln als externales Kriterium herangezogen werden. Dies wird aktuell im Rahmen der dritten Phase des Projekts verwirklicht (vgl. Seidel et al., eingereicht). Um eine standardisierte Erfassung zu ermöglichen, wurden nach dem Konzept der „approximations of practice“ (Grossman et al., 2009) Micro-Teaching-Events entwickelt, die komplexitätsreduzierte, aber authentische Beispiele von Unterricht darstellen.

8 GESAMTDISKUSSION

Gemäß dem dieser Arbeit zugrunde liegenden übergreifenden Konzept von Validität (vgl. AERA et al., 1999; Messick, 1995; vgl. Abschnitt 2.3.2.1.2) soll die Überprüfung der Validität nicht anhand einzelner Aspekte wie eine Checkliste abgearbeitet werden (Gorin, 2007). Vielmehr wird im Folgenden basierend auf den verschiedenen Evidenzgrundlagen eine übergreifende Validitätsaussage bezüglich des Einsatzes des Tools Observer im Large-Scale-Kontext synthetisiert. Im Anschluss daran wird der in der vorliegenden Arbeit gewählte Ansatz zur Validierung des Instruments beurteilt. Abschließend wird ein Ausblick auf anschließende Forschungsfragen skizziert.

8.1 Übergreifende Validitätsaussage bezüglich des Einsatzes des Tools Observer im Large-Scale-Kontext

Die vorliegende Arbeit leistet einen substantiellen Beitrag zur Validierung des Tools Observer und zeigt Möglichkeiten und Grenzen eines Einsatzes im Large-Scale-Kontext auf. In vier Forschungsfragen wurden die Validitätsaspekte (1) *Struktur,* (2) G*eneralisierbarkeit* über verschiedene Erhebungsbedingungen, (3) *Generalisierbarkeit* über verschiedene Lehramtsstudiengänge hinweg und (4) *Externalität* überprüft. Die Befunde zeigen, dass das Instrument in einer großen und heterogenen Stichprobe die Struktur professioneller Unterrichtswahrnehmung mit den drei Aspekten Beschreiben, Erklären und Vorhersagen empirisch abbildet und über verschiedene Erhebungsbedingungen hinweg stabil eingesetzt werden kann. Im Hinblick auf den Validitätsaspekt *Generalisierbarkeit* über verschiedene Lehramtsstudiengänge hinweg weisen die Befunde darauf hin, dass das Instrument innerhalb der untersuchten Studiengänge Lehramt für Primarstufe, Sekundarstufe und Berufliche Schulen die Struktur professioneller Unterrichtswahrnehmung stabil erfasst. Jedoch ist aktuell von einem Einsatz zum Vergleich verschiedener Lehramtsstudiengänge abzuraten, da sich substantielle Unterschiede in den Itemschwierigkeiten zeigten. Bezogen auf den Validitätsaspekt *Externalität* konnte ein gewisser Zusammenhang zwischen professioneller Unterrichtswahrnehmung mit Prädiktoren für den Studien- und Berufserfolg nachgewiesen werden, der jedoch noch weiterer Prüfung bedarf (vgl. Abschnitt 7.3.3).

Übergreifend ist zu konstatieren, dass mit dem Tool Observer ein Instrument vorliegt, dass professionelle Unterrichtswahrnehmung im Large-Scale-Kontext für Studierende unterschiedlicher Lehramtsstudiengänge valide erfasst. Damit kann das Instrument über die Beurteilung des individuellen Lernerfolgs Studierender hinaus in die universitäre Lehrerbildung integriert werden, um Aussagen auf systemischer Ebene zu treffen (z. B. standortübergreifende Wirksamkeit der Überprüfung von Lehrveranstaltungen). Gleichzeitig wird der Forderung entsprochen, im Large-Scale-Kontext bisher dominierende Selbsteinschätzungen im Fragebogenformat (Frey, 2006) um ein Instrument mit einer verhaltensnahen Kompetenzerfassung zu ergänzen (Kunter & Klusmann, 2010; Seidel et al., 2010a; Seidel & Stürmer, in Druck). Durch die Berücksichtigung des situativen Kontexts mithilfe der Videoclips ist ein größerer Zusammenhang zwischen der erfassten und der tatsächlichen Kompetenz zu erwarten als bei Selbsteinschätzungen im Fragebogenformat (vgl. Bledow & Frese, 2009; Darling-Hammond, 2006).

8.2 Beurteilung des gewählten Ansatzes zur Validierung

Für die vorliegende Arbeit wurde ein Konzept von Validität und ein Ansatz zur Validierung gewählt, der in Nordamerika entwickelt wurde und im deutschen Sprachraum bisher kaum rezipiert wird (vgl. Frey, 2014, März). Basierend auf einem übergreifenden Konzept von Validität, nach dem die Validität der Interpretation und Nutzung der Messergebnisse anhand verschiedener Evidenzgrundlagen beurteilt wird (vgl. AERA et al., 1999; Messick, 1995), wurde in der vorliegenden Arbeit ein integriertes Modell von Validität erarbeitet, das die verschiedenen Differenzierungen von Evidenzen (AERA et al. versus 1999; Messick, 1995) ineinander überführt. Für den Validierungsprozess des Tools Observer im Hinblick auf einen Einsatz im Large-Scale-Kontext wurde dieses Modell mit dem argumentbasierten Ansatz zur Validierung (Kane, 1992, 2001, 2013a) verbunden. Nach diesem Ansatz werden zunächst Anforderungen, die sich aus der intendierten Interpretation und Nutzung der Messergebnisse ergeben, spezifiziert und anschließend gezielt Evidenz dafür generiert. Darüber hinaus wurde der Validierungsprozess des Tools Observer anhand des generellen Modells bildungswissenschaftlichen Testens (Hattie et al., 1999; adaptiert nach Zumbo, 2007) ausdifferenziert, um verschiedene Anhaltspunkte zu erhalten, worauf sich einzelne Anforderungen beziehen können.

Im Rahmen der vorliegenden Arbeit hat sich das gewählte Vorgehen als gewinnbringend erwiesen. Gerade in Anbetracht der Tatsache, dass eine Menge an Instrumenten zur Kompetenzerfassung frei zur Verfügung gestellt wird, ist es wichtig ein, Messinstrument nicht grundsätzlich als valide einzuschätzen. Vielmehr sollte die Validität für jeden Kontext neu beurteilt werden. Dafür ist die Spezifikation der Anforderungen nützlich, die sich aus der intendierten Interpretation und Nutzung der Messergebnisse ergeben und somit stets implizit vorhanden sind, aber nicht immer explizit und damit überprüfbar gemacht werden. Die vorliegende Arbeit stellt ein Beispiel dafür dar, wie die Validität eines Instruments mit Blick auf eine bestimmte Nutzung der Messergebnisse durch die gezielte Überprüfung spezifischer Anforderungen ökonomisch überprüft werden kann.

Abschließend ist zu beachten, dass die Überprüfung der Validität des Tools Observer mit der vorliegenden Arbeit nicht abgeschlossen ist. Denn bei der Validierung handelt es sich um einen kontinuierlichen Prozess (Gorin, 2007). Zum einen folgen aus neuen Einsatzmöglichkeiten zusätzliche Anforderungen an die Interpretation und Nutzung der Messergebnisse. Sollte das Instrument beispielswiese im Rahmen von High-Stake-Testungen eingesetzt werden (z. B. zur Zulassung zum Lehramtsstudium), müsste der Validitätsaspekt *Konsequenzen* (vgl. Abschnitt 2.3.2.1.2), der die Folgen der Benutzung des Messinstruments fokussiert (AERA et al., 1999; Messick, 1995), überprüft werden. Zum anderen gilt zu berücksichtigen, dass, auch wenn Evidenz für die Erfüllung bestimmter Anforderungen vorliegt, dies kein Beweis dafür ist, dass diese Anforderungen tatsächlich gewährleistet sind (Borsboom & Markus, 2013). Dieser Einwand entspricht dem wissenschaftlichen Grundprinzip der Falsifikation (vgl. Popper, 1962), nachdem eine Hypothese nicht verifiziert werden kann. Deshalb ist es im Rahmen der Validierung eines Messinstruments entscheidend, die bestmöglich verfügbare Evidenz zu generieren und die Validität ständig weiter zu überprüfen (Kane, 2013b).

8.3 Ausblick auf anschließende Forschungsfragen

Im Folgenden werden ausgewählte Anschlussfragestellungen, die über die Replikation der Befunde der vorliegenden Arbeit hinausgehen, skizziert. Diese beziehen sich auf mögliche Erweiterungen des Instruments und eine multi-methodale Erfassung professioneller Unterrichtswahrnehmung.

8.3.1 Mögliche Erweiterungen des Tools Observer

Bei der Überprüfung der Externalität im Rahmen der vierten Forschungsfrage zeigt sich lediglich für zwei der vier Prädiktoren für Studien- und Berufserfolg ein Zusammenhang mit professioneller Unterrichtswahrnehmung. Wie in Abschnitt 7.3.3 erläutert, kann dies sowohl auf die Auswahl der Prädiktoren als auch auf das Instrument zurückgeführt werden. Im Hinblick auf das Instrument könnte der geringer als erwartet ausfallende Zusammenhang mit den Prädiktoren darauf hindeuten, dass die Referenz für die Qualität der Bearbeitung des Instruments nicht ausreichend berufsbezogen ist. Die Qualität der Bearbeitung wird über die Übereinstimmung der Einschätzungen der Rating-Items mit einer Expertennorm ermittelt. Die aktuell verwendete Norm entspricht der konsensvalidierten Einschätzung von drei Experten der Unterrichts- und Hochschulforschung mit 100 bis 400 Stunden Erfahrung in der Beobachtung von Unterricht (vgl. Abschnitt 2.2.3.2.4). Die Befunde der vierten Forschungsfrage sollten zum Anlass genommen werden, die aktuelle Expertennorm zu überprüfen. Dazu sollte eine zusätzliche Expertennorm erstellt und systematisch auf Unterschiede zur bisherigen Expertennorm untersucht werden. Grundsätzlich stellt die Auswahl geeigneter Experten eine große Herausforderung dar (Berliner, 2001). Es muss festgelegt werden, wofür Experten gesucht sind und anhand welcher Kriterien diese ausgewählt werden. Mit Blick auf einen hohen Berufsbezug könnten erfahrene Lehrpersonen als Experten dienen. Für die Auswahl von Lehrpersonen mit ausreichender Expertise werden häufig folgende Kriterien herangezogen: drei bis fünf Jahre Erfahrung im relevanten Unterrichtskontext, entsprechende Zertifikate, Empfehlungen von Kollegen und Vorgesetzten sowie Hinweise auf positive Effekte des Unterrichtens auf die Leistung der Schülerinnen und Schüler (Palmer, Stough, Burdenski & Gonzales, 2005). Für die professionelle Unterrichtswahrnehmung scheint nicht nur die Erfahrung im Unterrichten, sondern vor allem im Beobachten von Unterricht ausschlaggebend zu sein. Dies entspricht z. B. dem Tätigkeitsbereich von Seminarlehrerinnen und Seminarlehrern, die im Rahmen der Ausbildung von Referendarinnen und Referendaren häufig Unterricht beobachten und beurteilen. Folglich könnten für eine Expertennorm aus Lehrpersonen drei Seminarlehrinnen und -lehrer ausgewählt werden, die bereits drei bis fünf Jahre in dieser Funktion tätig sind und von Kollegen und Vorgesetzten als Experten in der Beobachtung von Unterricht empfohlen werden. Durch eine systematische Prüfung beider Expertennormen können mögliche Unterschiede zwischen diesen sowie den jeweils erzielten Übereinstimmungen der Studierenden identifiziert werden.

Darüber hinaus könnte das Instrument durch die Integration zusätzlicher Videoclips erweitert werden, um den Einfluss des Unterrichtskontexts auf die Erfassung professioneller Unterrichtswahrnehmung gezielt zu untersuchen. Wie in Abschnitt6.3.1.2 erläutert, könnte der Unterschied in den Itemschwierigkeiten für Studierende Lehramt für Primarstufe und Berufliche Schulen auf die Diskrepanz zwischen dem im Videoclip gezeigten Unterricht und dem späteren Handlungsfeld der Studierenden zurückzuführen sein. Deshalb sollten Videoclips in das Instrument eingebunden werden, die vergleichbare Unterrichtssequenzen aus Primarstufe, Sekundarstufe und Beruflichen Schulen zeigen. Dabei ist zu beachten, die Gesamtanzahl von sechs Videoclips nicht zu überschreiten, um eine gewissenhafte Bearbeitung zu gewährleisten und die Teilnehmenden nicht zu überlasten. In einem ersten Schritt sollten die Itemschwierigkeiten für Studierende der drei Lehramtsstudiengänge verglichen werden. In einem zweiten Schritt sollte die Bearbeitung einer Version des Instruments mit adaptierten Videoclips auf weitere Zielgruppen mit mehr Expertise ausgeweitet werden, die in der Primarstufe, Sekundarstufe und in Beruflichen Schulen unterrichten. Mögliche Zielgruppen mit größerer Expertise wären Referendarinnen und Referendare, Lehrpersonen, Seminarlehrerinnen und Seminarlehrer sowie Schulleiterinnen und Schulleiter. Dadurch könnte überprüft werden, ob der Einfluss des Unterrichtskontexts auf die Erfassung professioneller Unterrichtswahrnehmung bei Novizen und Experten vergleichbar ist. Nachdem der Kontext für den Erwerb von Expertise eine entscheidende Rolle spielt (vgl. Ericsson et al., 1993), bleibt offen, inwieweit die Itemschwierigkeiten für diese Expertengruppen unabhängig vom Unterrichtskontext vergleichbar sind. Erste Befunde aus dem Projekt BilWiss-Beruf zeigen, dass die Struktur professioneller Unterrichtswahrnehmung auch in der Gruppe der Referendarinnen und Referendare grundsätzlich empirisch abgebildet werden kann und die Itemschwierigkeiten mit denen der Lehramtsstudierenden vergleichbar sind (Stürmer & Seidel, eingereicht). Damit sind die Voraussetzungen erfüllt, an einer Stichprobe mit Referendarinnen und Referendaren der Primarstufe, Sekundarstufe und Beruflichen Schulen mit adaptierten Videoclips den Einfluss der Passung zwischen Unterrichtskontext im Videoclip und eigenem Handlungsfeld unter Einbezug der Expertise systematisch zu untersuchen.

8.3.2 Multi-methodale Erfassung professioneller Unterrichtswahrnehmung

Nach der erfolgreichen Prüfung des Tools Observer als Instrument, das professionelle Unterrichtswahrnehmung im Large-Scale-Kontext valide erfasst, sollte in einem nächsten Schritt eine Kombination mit anderen Erfassungsmethoden angestrebt werden. Dies entspricht der Forderung nach einer multi-methodalen Kompetenzerfassung, um der Komplexität professioneller Kompetenz von Lehrpersonen gerecht zu werden und mögliche Einschränkungen einzelner Erhebungsverfahren auszugleichen (vgl. Darling-Hammond, 2006; Maag Merki & Werner, 2006). Zur Erfassung professioneller Unterrichtswahrnehmung könnte das standardisierte Tool Observer mit offenen Fragen, inwieweit die Lehrperson die in den Videoclips repräsentierten lernwirksamen Unterrichtskomponenten umgesetzt hat, kombiniert werden. Im Vergleich zu geschlossenen Antwortformaten wird so die Wissensanwendung ohne spezifische Prompts erfasst. Darüber hinaus wäre der Einsatz eines Eye-Trackers bei der Bearbeitung des Instruments denkbar. Die aufgezeichneten Augenbewegungen während des Ansehens der Videoclips könnten Anhaltspunkte dafür liefern, inwieweit bestimmte Aktivitäten der Lehrperson sowie der Schülerinnen und Schüler tatsächlich wahrgenommen wurden. Eine Kombination des Eye-Trackers mit Laut-Denken-Protokollen könnte zusätzlich Begründungen dafür liefern, warum bestimmte Personen fokussiert werden. Die genannten Erfassungsmethoden erscheinen für einen Einsatz im Large-Scale-Kontext jedoch zu zeitaufwendig und kostenintensiv. Allerdings wäre es möglich, ausgewählte Prozessdaten, die während der Bearbeitung aufgezeichnet werden, für die Kompetenzerfassung zu nutzen. Daten bezüglich der Zeit, die für das Ansehen eines Videoclips oder die Bearbeitung der Rating-Items verwendet wird, könnten Aufschluss darüber geben, wie gewissenhaft das Instrument bearbeitet wurde.

Im Bereich der multi-methodalen Kompetenzerfassung besteht noch erheblicher Forschungsbedarf (Maag Merki & Werner, 2006). Deshalb sollte empirisch überprüft werden, wie die Erfassung professioneller Unterrichtswahrnehmung durch das Tool Observer sinnvoll ergänzt werden kann, ohne die Kompetenzerfassung durch das Instrument zu beeinflussen.

9 LITERATURVERZEICHNIS

Abel, J. (2010). Veränderung der Berufs- und Wissenschaftsorientierung. In J. Abel (Hrsg.), *Wirkt Lehrerbildung? Antworten aus der empirischen Forschung* (S. 25–33). Münster: Waxmann.

Adams, R. J. & Carstensen, C. (2002). Scaling outcomes. In R. J. Adams & M. Wu (Eds.), *PISA 2000 Technical Report* (pp. 149–162). Paris: OECD.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Hillsdale: Lawrence Erlbaum Associates.

Artelt, C. & Schneider, W. (2011). Herausforderungen und Möglichkeiten der Diagnose und Modellierung von Kompetenzen und ihrer Entwicklung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 43,* 167-172.

Bauer, J. & Drechsel, B. (2010, März). *Subjektive Begriffe von "Lehren" im Lehramtsstudium: Erste Befunde mit einem neuen Instrument.* Vortrag auf dem 22. Kongress der Deutschen Gesellschaft für Erziehungswissenschaft (DGfE), Mainz.

Bauer, J., Diercks, U., Retelsdorf, J., Sporer, T., Zimmermann, F., Köller, O. et al. (2011). Spannungsfeld Polyvalenz in der Lehrerbildung: Wie polyvalent sind Lehramtsstudiengänge und was bedeutet dies für die Berufswahlsicherheit der Studierenden? *Zeitschrift für Erziehungswissenschaft, 14,* 629-649.

Bauer, J., Drechsel, B., Retelsdorf, J., Sporer, T., Rösler, L., Prenzel, M. et al. (2010). Panel zum Lehramtsstudium – PaLea: Entwicklungsverläufe zukünftiger Lehrkräfte im Kontext der Reform der Lehrerbildung. *Beiträge zur Hochschulforschung, 32,* 34-55.

Baumert, J. & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft, 9,* 469-520.

Baumert, J., Stanat, P. & Demmrich, A. (2001). PISA 2000: Untersuchungsgegenstand, theoretische Grundlagen und Durchführung der Studie. In J. Baumert, E. Klieme, M.

Neubrand, M. Prenzel, U. Schiefele, W. Schneider et al. (Hrsg.), *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich.* Opladen: Leske + Budrich.

Beck, E., Baer, M., Guldimann, T., Bischoff, S., Brühwiler, C., Müller, P. et al. (2008). *Adaptive Lehrkompetenz: Analyse von Struktur, Veränderbarkeit und Wirkung handlungssteuernden Lehrerwissens*. Münster: Waxmann.

Bergman, L. R. & Andersson, H. (2010). The person and the variable in the developmental psychology. *Zeitschrift für Psychologie/Journal of Psychology, 218,* 155-165.

Berliner, D. C. (2001). Learning about and learning from expert teachers. *International Journal of Educational Research, 35,* 463-482.

Billich-Knapp, M., Künsting, J. & Lipowsky, F. (2012). Profile der Studienwahlmotivation bei Grundschullehramtsstudierenden. *Zeitschrift für Pädagogik, 58,* 696-719.

Birnbaum, M. H. (2004). Human research and data collection via the internet. *Annual Review of Psychology, 55,* 803-832.

Bledow, R. & Frese, M. (2009). A situational judgment test of personal initiative and its relationship to performance. *Personnel Psychology, 62,* 229-258.

Blömeke, S. (2004). Empirische Befunde zur Wirksamkeit von Lehrerbildung. In S. Blömeke, P. Reinhold, G. Tulodziecki & J. Wildt (Hrsg.), *Handbuch Lehrerbildung* (S. 59–91). Bad Heilbrunn/Obb: Klinkhardt.

Blömeke, S. (2009). Ausbildungs- und Berufserfolg im Lehramtsstudium im Vergleich zum Diplom-Studium: Zur prognostischen Validität kognitiver und psycho-motivationaler Auswahlkriterien. *Zeitschrift für Erziehungswissenschaft, 12,* 82-110.

Blömeke, S., Kaiser, G., Lehmann, R., König, J., Döhrmann, M., Buchholtz, C. et al. (2009). TEDS-M: Messung von Lehrerkompetenzen im internationalen Vergleich. In Mulder R, O. Zlatkin-Troitschanskaia, K. Beck, R. Nickolaus & D. Sembill (Hrsg.), *Professionalität von Lehrenden: Zum Stand der Forschung* (S. 181–210). Weinheim: Beltz.

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2. Vol.). Mahwah, NJ.: Lawrence Erlbaum Associates Publishers.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher, 33,* 3-15.

Borsboom, D. & Markus, K. A. (2013). Truth and evidence in validity theory. *Journal of Educational Measurement, 50,* 110-114.

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation: Für Human- und Sozialwissenschaftler* (4. Aufl.). Heidelberg: Springer.

Bromme, R. (1992). Expertenstudien mit Lehrern. In R. Bromme (Hrsg.), *Der Lehrer als Experte: Zur Psychologie des professionellen Wissens* (S. 52–72). Bern: Hans Huber.

Brookhart, S. M. & Freeman, D. J. (1992). Characteristics of entering teacher candidates. *Review of Educational Research, 62,* 37-60.

Bruder, S., Keller, S., Klug, J. & Schmitz, B. (2011). Ein Vergleich situativer Methoden zur Erfassung der Beratungskompetenz von Lehrkräften. *Unterrichtswissenschaft, 39,* 123-137.

Brunner, M., Kunter, M., Krauss, S., Klusmann, U., Baumert, J., Blum, W. et al. (2006). Die professionelle Kompetenz von Mathematiklehrkräften: Konzeptualisierung, Erfassung und Bedeutung für den Unterricht: Eine Zwischenbilanz des COACTIV-Projekts. In M. Prenzel & L. Allolio-Näcke (Hrsg.), *Untersuchungen zur Bildungsqualität von Schule: Abschlussbericht des DFG-Schwerpunktprogramms* (S. 54–82). Münster: Waxmann.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3. Aufl.). Psychologie. München, Boston: Pearson Studium.

Camilli, G. & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage Publications.

Carter, K., Cushing, K., Sabers, D., Stein, P. & Berliner, D. C. (1988). Expert-novice differences in perceiving and processing visual classroom information. *Journal of Teacher Education, 39,* 25-31.

Cochran-Smith, M. (2003). Assessing assessment in teacher education. *Journal of Teacher Education, 54,* 187-191.

Collins, L. M. & Lanza, S. T. (2010). *Latent class and latent transition analysis for the social, behavioral, and health sciences*. Hoboken, NJ: Wiley.

Collins, L. M., Schafer, J. L. & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6,* 330-351.

Cronbach, L. J. (1980). Selection theory for a political world. *Public Personnel Management,9,* 37-50.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Czerwenka, K. & Nölle, K. (2011). Forschung zur ersten Phase der Lehrerbildung. In E. Terhart, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (S. 362–380). Münster: Waxmann.

Darling-Hammond, L. (2006). Assessing teacher education: The usefulness of multiple measures for assessing program outcomes. *Journal of Teacher Education, 57,* 120–138.

de Ayala, R. J. (2009). *The theory and practice of item response theory: Methodology in the social sciences.* New York: Guilford Press.

Deci, E. L. & Ryan, R. M. (1993). Die Selbstbestimmungstheorie der Motivation und ihre Bedeutung für die Pädagogik. *Zeitschrift für Pädagogik, 39,* 223–238.

DeMars, C. E. (2011). An analytic comparison of effect sizes for differential item functioning. *Applied Measurement in Education, 24,* 189–209.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualization and measures. *Educational Researcher, 38,* 181–199.

Drechsel, B. (2001). *Subjektive Lernbegriffe und Interesse am Thema Lernen bei angehenden Lehrpersonen*. Münster: Waxmann.

Embretson, S. E. & Reise, S. (2000). *Psychometric methods: Item response theory for psychologists. Multivariate applications.* Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Ericsson, K. A., Krampe, R. Th. & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100,* 363–406.

Field, A. & Hole, G. (2003). *How to design and report experiments*. London: Sage.

Field, A. (2009). *Discovering statistics using SPSS* (3. Vol.). Introducing statistical methods. Los Angeles: Sage Publications.

Fischer, G. H. (1995). Derivations of the Rasch model. In G. H. Fischer & I. W. Molenaar (Hrsg.), *Rasch models. Foundations, recent developments, and applications* (S. 15–38). New York: Springer.

Foerster, F. (2008). *Personale Voraussetzungen von Grundschullehramtsstudierenden. Eine Untersuchung zur prognostischen Relevanz von Persönlichkeitsmerkmalen für den Studien- und Berufserfolg*. Münster: Waxmann.

Frey, A. (2006). Methoden und Instrumente zur Diagnose beruflicher Kompetenzen von Lehrkräften – eine erste Standortbestimmung. In C. Allemann-Ghionda (Hrsg.), *Kompetenzen und Kompetenzentwicklung von Lehrerinnen und Lehrern.* 51. Beiheft der Zeitschrift für Pädagogik (S. 30–46). Weinheim: Beltz.

Frey, A. (2014, März). *Validität: Internationaler Forschungsstand und Umsetzung in Deutschland:* Vortrag auf der 2. Tagung der Gesellschaft für Empirische Bildungsforschung (GEBF), Frankfurt.

Frey, A. & Jung, C. (2011). Kompetenzmodelle und Standards in Lehrerbildung und Lehrerberuf. In E. Terhart, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (S. 540–572). Münster: Waxmann.

Furr, R. M. & Bacharach, V. R. (2008). *Psychometrics: An introduction*. Los Angeles: Sage Publications.

Geiser, C. (2009). *Datenanalyse mit Mplus: Eine anwendungsorientierte Einführung* (1. Aufl.). Lehrbuch. Wiesbaden: VS Verlag für Sozialwissenschaften.

Gold, A. & Souvignier, E. (2008). Prognose der Studierfähigkeit. Ergebnisse aus Längsschnittanalysen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 37,* 214–222.

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,* 141–155.

Gollwitzer, M. (2012). Latent-Class-Analysis. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Aufl., S. 279–306). Berlin, Heidelberg: Springer.

Goodwin, C. (1994). Professional vision. *American Anthropologist, 96,* 606–633.

Gorin, S. J. (2007). Reconsidering issues in validity theory. *Educational Researcher, 36,* 456–462.

Graham, J. W. (2009). Missing data analysis: Making it work in the real word. *Annual Review of Psychology, 60,* 549–576.

Grossman, P., Compton, C., Igra, D., Ronfeldt M, Shahan, E. & Williamson, P. W. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record, 119,* 2055–2100.

Hartig, J. (2008). Kompetenzen als Ergebnisse von Bildungsprozessen. In N. Jude, J. Hartig & E. Klieme (Hrsg.), *Kompetenzerfassung in pädagogischen Handlungsfeldern: Theorien, Konzepte und Methoden* (S. 15–25). Berlin: Bundesministerium für Bildung und Forschung.

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

Hartig, J., Klieme, E. & Leutner, D. (Hrsg.). (2008). *Assessment of competencies in educational contexts*. Cambridge, MA: Hogrefe & Huber Publishers.

Hattie, J., Jaeger, R. M. & Bond, L. (1999). Persistent methodological questions in educational testing. *Review of Research in Education, 40,* 393–446.

Heckhausen, H. (1989). *Motivation und Handeln*. Berlin: Springer.

Heller, K. A. & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision*. Göttingen: Beltz.

Herlt, S. & Schaarschmidt, U. (2007). Fit für den Lehrerberuf?! Ein Selbsterkundungsverfahren für Interessenten am Lehramtsstudium. In U. Schaarschmidt & U. Kieschke (Hrsg.), *Gerüstet für den Schulalltag: Psychologische Unterstützungsangebote für Lehrerinnen und Lehrer* (S. 157–181). Weinheim, Basel: Beltz.

Herzberg, P. Y. & Brähler, E. (2006). Assessing the big-five personality domains via short forms: A cautionary note and a proposal. *European Journal of Psychological Assessment, 22,* 139–148.

Humphreys, M. S. & Revelle, W. (1984). Personality, Motivation and Performance: A theory of the relationship between individual differences and information processing. *Psychological Review, 91,* 153–184.

Jahn, G., Prenzel, M., Stürmer, K. & Seidel, T. (2011). Varianten einer computergestützten Erhebung von Lehrerkompetenzen: Untersuchungen zu Anwendungen des Tools „Observer“. *Unterrichtswissenschaft, 39,* 136–153.

Jahn, G., Stürmer, K., Seidel, T. & Prenzel, M. (in Druck). Professionelle Unterrichtswahrnehmung von Lehramtsstudierenden: Eine Scaling-up-Studie des Observe-Projekts. *Zeitschrift für Entwicklungspsychologie und pädagogische Psychologie*.

Jude, N. & Klieme, E. (2008). Einleitung. In N. Jude, J. Hartig & E. Klieme (Hrsg.), *Kompetenzerfassung in pädagogischen Handlungsfeldern. Theorien, Konzepte und Methoden* (S. 9–13). Berlin: Bundesministerium für Bildung und Forschung.

Jurecka, A. (2008). Introduction to the computer-based assessment of competencies. In J. Hartig, E. Klieme & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 193–213). Cambridge, MA: Hogrefe & Huber Publishers.

Kane, M. T. (1992). An argument-based approach to validity. *Quantitative Methods in Psychology, 112,* 527–535.

Kane, M. T. (1994). Validating interpretive arguments for licensure and certification examinations. *Evaluation and the Health Professions, 19,* 133–159.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38,* 319–342.

Kane, M. T. (2013a). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50,* 1–73.

Kane, M. T. (2013b). Validation as a pragmatic, scientific activity. *Journal of Educational Measurement, 50,* 115–122.

Kersting, N. B. (2008). Using video clips of mathematics classroom instruction as item prompts to measure teachers' knowledge of teaching mathematics. *Educational and Psychological Measurement, 68,* 845–861.

Kersting, N. B., Givvin, K. B., Sotelo, F. L. & Stigler, J. W. (2010). Teachers' analyses of classroom video predict student learning of mathematics: Further explorations of a novel measure of teacher knowledge. *Journal of Teacher Education, 61,* 162–181.

Kersting, N. B., Givvin, K., Thompson, B. J., Santagata, R. & Stigler, J. W. (2012). Measuring usable knowledge: Teachers' analyses of mathematics classroom videos predict

teaching quality and student learning. *American Educational Research Journal, 49,* 568–589.

Keuffer, J. (2010). Reform der Lehrerbildung und kein Ende? Eine Standortbestimmung. *Zeitschrift Erziehungswissenschaft, 21,* 51–67.

Kingston, N. (2007). Future challenges to psychometrics: Validity, validity, validity. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics. Psychometrics* (pp. 1111–1112). North Holland: Elsevier Science B.V..

Kleickmann, T. & Anders, Y. (2011). Lernen an der Universität. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV* (S. 305–315). Münster: Waxmann.

Kleickmann, T., Richter, D., Kunter, M., Elsner, J., Besser, M., Krauss, S. et al. (2013). Teachers' content knowledge and pedagogical content knowledge: The role of structural differences in teacher education. *Journal of Teacher Education, 64,* 90–106.

Klieme, E. & Hartig, J. (2007). Kompetenzkonzepte in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs. *Zeitschrift für Erziehungswissenschaft, Sonderheft 8,* 11–29.

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

Klieme, E. & Leutner, D. (2006b). *Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen: Schwerpunktprogrammantrag an die DFG*. Frankfurt a.M.: Deutsches Institut für Internationale Pädagogische Forschung.

Klieme, E., Hartig, J. & Rauch, D. (2008). The concept of competence in educational contexts. In E. Klieme, D. Leutner & J. Hartig (Eds.), *Assessment of competencies in educational contexts* (pp. 3–22). Toronto: Hogrefe & Huber Publishers.

Klusmann, U., Trautwein, U., Lüdtke, O., Kunter, M. & Baumert, J. (2009). Eingangsvoraussetzungen beim Studienbeginn: Werden Lehramtskandidaten unterschätzt? *Zeitschrift für Pädagogische Psychologie, 23,* 265–278.

Koeppen, C., Hartig, J., Klieme, E. & Leutner, D. (2008). Current issues in competence modeling and assessment. *Zeitschrift für Psychologie/Journal of Psychology, 216,* 61–73.

Köller, M., Klusmann, U., Retelsdorf, J. & Möller, J. (2012). Geeignet für den Lehrerberuf? Self-Assessments auf dem Prüfstand. *Unterrichtswissenschaft, 40,* 121–139.

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

Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2004). *Standards für die Lehrerbildung: Bildungswissenschaften (Beschluss der KMK vom 16.12.2004).* Bonn: KMK. Zugriff am 14.07.2014 unter http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung-Bildungswissenschaften.pdf.

Krapp, A. & Prenzel, M. (2011). Research on Interest in Science: Theories, methods, and findings. *International Journal of Science Education, 33,* 27–50.

Kunina-Habenicht, O., Lohse-Bossenz, H., Kunter, M., Dicke, T., Förster, D., Gößling, J. et al. (2012). Welche bildungswissenschaftlichen Inhalte sind wichtig in der Lehrerbildung? Ergebnisse einer Delphi-Studie, *15,* 649–682.

Künsting, J. & Lipowsky, F. (2011). Studienwahlmotivation und Persönlichkeitseigenschaften als Prädiktoren für Zufriedenheit und Strategienutzung im Lehramtsstudium, *Zeitschrift für Pädagogische Psychologie, 25,* 105–114.

Kunter, M. & Klusmann, U. (2010). Kompetenzmessung von Lehrkräften - Methodische Herausforderungen. *Unterrichtswissenschaft, 38,* 68–86.

Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S. & Neubrand, M. (Hrsg.). (2011). *Professionelle Kompetenz von Lehrkräften: Ergebnisse des Forschungsprogramms COACTIV*. Münster: Waxmann.

Kunter, M., Kleickmann, T., Klusmann, U. & Richter, D. (2011). Die Entwicklung professioneller Kompetenz 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. 55–68). Münster: Waxmann.

Leutner, D., Fleischer, J., Wirth, J., Greiff, S. & Funke Joachim. (2012). Analytische und dynamische Problemlösekompetenz im Lichte internationaler Schulleistungsvergleichsstudien: Untersuchungen zur Dimensionalität. *Psychologische Rundschau, 63,* 34–42.

Leutner, D., Hartig, J. & Jude, N. (2008). Measuring competencies: Introduction to concepts and questions of assessment in education. In J. Hartig, E. Klieme & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 177–192). Cambridge, MA: Hogrefe & Huber Publishers.

Lienert, G. Adolf & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Aufl.). Weinheim: Beltz.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7,* 328.

Lo, Y., Mendell, N. R. & Rubin, D. R. (2001). Testing the number of components in a normal mixture. *Biometrika, 88,* 767–778.

Lubke, G. H. & Muthén, B. (2005). Investigating population heterogenity with factor mixture models. *Psychological Methods, 10,* 21–39.

Maag Merki, K. & Werner, S. (2011). Erfassung und Bewertung professioneller Kompetenz von Lehrpersonen. In E. Terhart, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (S. 573–591). Münster: Waxmann.

McCrae, R. R. & Costa, P. T. (2008). The five-factor theory of personality. In O. P. John, R. W. Robins & L. A. Pervin (Eds.), *Handbook of personality. Theory and research* (3. Vol., pp. 159–181). New York: Guilford Press.

Messick, S. (1975). The standard problem: Meanings and values in measurement and evaluation. *American Psychologist, 30,* 955–966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35,* 1012–1027.

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

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research, 45,* 35–44.

Messner, H. & Reusser, K. (2000). Berufliches Lernen als lebenslanger Prozess. *Beiträge zur Lehrerbildung, 18,* 277–294.

Molenaar, I. W. (1995). Some background for item response theory and the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 3–14). New York: Springer.

Monahan, P. O., McHorney, C. A., Stump, T. E. & Perkins, A. J. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics, 32,* 92–109.

Muthén, L. & Muthén, B. (1998-2013). *Mplus Version 7.11*. Los Angeles, CA: Muthén & Muthén.

Nieskens, B. & Müller, F. (2009). Soll ich Lehrerin werden? Web-basierte Selbsterkundung persönlicher Voraussetzungen und Interessen. *Erziehung und Unterricht, 159,* 41–49.

O'Neill, K. A. & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255–279). Hillsdale: Lawrence Erlbaum Associates.

Organisation for Economic Co-Operation and Development (2014). *PISA 2012 Results: Creative problem solving: Students' skills in tackling real-life problems (Vol. 5),* PISA, OECD Publishing. Zugriff am 14.07.2014 unter http://dx.doi.org/10.1787/9789264208070-en.

Oser, F. (1997). Standards in der Lehrerbildung. Teil 2: Wie werden Standards in der schweizerischen Lehrerbildung erworben? Erste empirische Ergebnisse. *Beiträge zur Lehrerbildung, 15,* 210–228.

Oser, F., Heinzer, S. & Salzmann, P. (2010). Die Messung der Qualität von professionellen Kompetenzprofilen von Lehrpersonen mit Hilfe der Einschätzung von Filmvignetten: Chancen und Grenzen des advokatorischen Ansatzes. *Unterrichtswissenschaft, 38,* 5–28.

Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research, 62,* 307–333.

Palmer, D. J., Stough, L. M., Burdenski, T. K. & Gonzales, M. (2005). Identifying teacher expertise: An examination of researchers' decision making. *Educational Psychologist, 40,* 13–25.

Pastor, D. A., Barron, K., Miller, B. J. & Davis, S. L. (2007). A latent profile analysis of college students' achievement goal orientation. *Contemporary Educational Psychology, 32,* 8-47.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal for Personality and Social Psychology, 46,* 598–609.

Pellegrino, J. W., Chudowsky, N. & Glaser R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington DC: National Academic Press.

Penfield, R. D. & Algina, J. (2006). A generalized DIF effect variance estimator for measuring unsigned differential test functioning in mixed formats tests. *Journal of Educational Measurement, 43,* 295–312.

Penfield, R. D. & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (pp. 125–167). North Holland: Elsevier Science B.V..

Plöger, W. & Scholl, D. (2014). Analysekompetenz von Lehrpersonen - Modellierung und Messung. *Zeitschrift für Erziehungswissenschaft, 17,* 85–112.

Pohlmann, B. & Möller, J. (2010). Fragebogen zur Erfassung der Motivation für die Wahl des Lehramtsstudiums. *Zeitschrift für Pädagogische Psychologie, 24,* 73–84.

Popper, K. R. (1962). *Conjecture and refutation: The growth of scientific knowledge*. New York: Basic Books.

Prenzel, M. (2009). Von der Unterrichtsforschung zur Exzellenz in der Lehrerbildung. *Beiträge zur Lehrerbildung, 27,* 327–345.

Prenzel, M. (2013). Initiativen und Perspektiven zur Weiterentwicklung der Lehrerbildung. In W. Benz, J. Kohler & K. Landfried (Hrsg.), *Handbuch Qualität in Studium und Lehre: Evaluation nutzen - Akkreditierung sichern - Profile schärfen* (S. 1–21). Stuttgart: Raabe Verlag.

Prenzel, M., Gogoling, I. & Krüger, H.-H. (2007). Editorial. *Zeitschrift Erziehungswissenschaft, Sonderheft 8,* 5–8.

Prenzel, M., Reiss, K. & Seidel, T. (2011). Lehrerbildung an der TUM School of Education. *Erziehungswissenschaft, 22,* 47–56.

Prenzel, M., Seidel, T. & Drechsel, B. (2004). Autonomie in Wissensprozessen. In G. Reinmann & H. Mandl (Hrsg.), *Psychologie des Wissensmanagements: Perspektiven, Theorien und Methoden* (S. 102–113). Göttingen: Hogrefe.

Priemer, B. (2004). Logfile Analysen: Möglichkeiten und Grenzen ihrer Nutzung bei Untersuchungen zur Mensch-Maschine-Interaktion. *Medienpädagogik, 1,* 1–23.

Rauch, D. & Hartig, J. (2012). Interpretation von Testwerten in der IRT. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Aufl., S. 240-259). Berlin, Heidelberg: Springer.

Retelsdorf, J. & Möller, J. (2012). Grundschule oder Gymnasium? Zur Motivation ein Lehramt zu studieren. *Zeitschrift für Pädagogische Psychologie, 26,* 5–17.

Reusser, K., Pauli, C. & Elmer, A. (2011). Berufsbezogene Überzeugungen von Lehrerinnen und Lehrern. In E. Terhart, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (S. 478–495). Münster: Waxmann.

Rösler, L., Zimmermann, F., Bauer, J., Möller, J. & Köller, M. (2013). Interessieren sich Lehramtsstudierende für bildungswissenschaftliche Studieninhalte. *Zeitschrift für Pädagogik, 59,* 24–41.

Rost, J. (1999). Was ist aus dem Rasch-Modell geworden? *Psychologische Rundschau, 50,* 140–156.

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* (2. Aufl.). Psychologie Lehrbuch. Bern: Huber.

Rothland, M. (2011). Warum entscheiden sich Studierende für den Lehrberuf? Interessen, Orientierungen und Berufswahlmotive angehender Lehrkräfte im Spiegel empirischer Forschung. In E. Terhart, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (S. 268–295). Münster: Waxmann.

Rozendaal, J. S., Minnaert, A. & Boekaerts, M. (2003). Motivation and self-regulated learning in secondary vocational education: Information processing type and gender differences. *Learning and Individual Differences, 13,* 273–289.

Rutheman, U. (2004). Die Psychologie in der Lehrerbildung zwischen Berufsfeld- und Wissenschaftsorientierung. *Beiträge zur Lehrerbildung, 22,* 353–361.

Santagata, R. & Angelici, G. (2010). Studying the impact of the lesson analysis framework on pre-service teachers' ability to reflect on videos of classroom teaching. *Journal of Teacher Education, 61,* 339–349.

Santagata, R. & Guarino, J. (2011). Using video to teach future teachers to learn from teaching. *ZDM The International Journal of Mathematics Education, 43,* 133–145.

Santagata, R., Zannoni, C. & Stigler, J. W. (2007). The role of lesson analysis in pre-service teacher education: An empirical investigation of teacher learning from a virtual video-based field experiment. *Journal of Mathematics Teacher Education, 10,* 124–140.

Schaefers, C. (2002). Forschung zur Lehrerausbildung in Deutschland - eine bilanzierende Übersicht der neueren empirischen Studien. *Schweizerische Zeitschrift für Bildungswissenschaften, 24,* 65–88.

Schafer, J. L. & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177.

Schäfer, S. & Seidel, T. (akzeptiert). Noticing and reasoning of teaching and learning components by pre-service teachers. *Journal of Educational Research Online*.

Schaufeli, W. B., Martínez, I. M., Marques Pinto, A., Salanova, M. & Bakker, A. B. (2002). Burnout and engagement in university students: A cross-national study. *Journal of Cross-Cultural Psychology, 33,* 464–481.

Schiefele, U., Krapp, A., Wild, K.-P. & Winteler, A. (1993). Der "Fragebogen zum Studieninteresse" (FSI). *Diagnostica, 39,* 335–351.

Schommer, M. (1993). Epistemological development and academic performance among secondary students. *Journal of Educational Psychology, 85,* 406-411.

Seidel, T. (2003). *Lehr-Lernskripts im Unterricht*. Münster: Waxmann.

Seidel, T. & Prenzel, M. (2007). Wie Lehrpersonen Unterricht wahrnehmen und einschätzen - Erfassung pädagogisch-psychologischer Kompetenzen bei Lehrpersonen mit Hilfe von Videosequenzen. *Zeitschrift Erziehungswissenschaft, Sonderheft 8,* 201–218.

Seidel, T. & Prenzel, M. (2008). Assessment in Large-Scale Studies. In J. Hartig, E. Klieme & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 279–304). Cambridge MA: Hogrefe & Huber Publishers.

Seidel, T. & Prenzel, M. (2011). *Observe III – Differenzielle Untersuchungen zur prädiktiven Validität der Professionellen Unterrichtswahrnehmung als pädagogisch-psychologische Kompetenz von Lehramtsstudierenden.* Projektantrag im Rahmen des Schwerpunktprogramms "Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen" an die Deutsche Forschungsgemeinschaft. München: Technische Universität München.

Seidel, T. & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77,* 454–499.

Seidel, T. & Stürmer, K. (in Druck). Modeling and measuring the structure of professional vision in pre-service teachers. *American Educational Research Journal,* doi: 10.3102/0002831214531321.

Seidel, T., Blomberg, G. & Stürmer, K. (2010a). OBSERVE - Validierung eines videobasierten Instruments zur Erfassung der professionellen Wahrnehmung von Unterricht. In E. Klieme & M. Kenk (Hrsg.), *Kompetenzmodellierung: Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes. 56. Beiheft der Zeitschrift für Pädagogik* (S. 296–307). Weinheim: Beltz.

Seidel, T., Blomberg, G. & Stürmer, K., (2010b). *OBSERVER - Videobasiertes Tool zur Diagnose pädagogisch-psychologischer Kompetenzen bei Lehrpersonen* [Software]. München: TUM School of Education/ TU München. Zugriff am 14.07.2014 unter http://ww3.unipark.de/uc/Observer_Demo/kv/.

Seidel, T., Stürmer, K., Prenzel, M., Jahn, G. & Schäfer, S. (eingereicht). Investigating preservice teachers' professional vision within university-based teacher education.

Shavelson, R. J. (2010). On the measurement of competency. *Empirical Research in Vocational Education and Training, 2,* 41–63.

Shavelson, R. J. (2012). Assessing business-planning competence using the collegiate learning assessment as a prototype. *Empirical Research in Vocational Education and Training, 4,* 77–90.

Shavelson, R. J. & Ruiz-Primo (2005). On the psychometrics of assessing science understanding. In J. J. Mintzes, J. H. Wandersee & J. D. Novak (Eds.), *Assessing science understanding: A human constructivist view* (pp. 303–341). Burlington, MA: Elsevier Academic Press.

Sherin, M. (2007). The development of teachers' professional vision in video clubs. In R. Goldman, P. B. Baron & S. J. Derry (Eds.), *Video research in the learning sciences* (pp. 383–395). Mahwah, NJ: Lawrence Erlbaum Associates.

Sherin, M. G. (2002). When teaching becomes learning. *Cognition and Instruction, 20,* 119–150.

Sherin, M. G. & Han, S. Y. (2004). Teacher learning in the context of a video club. *Teaching and Teacher Education, 20,* 163–183.

Sherin, M. G. & van Es, E. A. (2009). Effects of video club participation on teachers' professional vision. *Journal of Teacher Education, 60,* 20–37.

Sherin, M. G., Jacobs, V. R. & Philipp, R. A. (2011). *Mathematics teacher noticing: Seeing through teachers' eyes*. New York: Routledge.

Shulman, L. S. (1987). Knowledge and teaching: Foundation of the new reform. *Harvard Educational Review, 57,* 1–23.

Spinath, B. (2012). Beiträge der Pädagogischen Psychologie zur Professionalisierung von Lehrerinnen und Lehrern: Diskussion zum Themenschwerpunkt. *Zeitschrift für Pädagogische Psychologie, 26,* 307–312.

Star, J. R. & Strickland, S. K. (2008). Learning to observe: Using video to improve preservice mathematics teachers' ability to notice. *Journal of Mathematics Teacher Education, 11,* 107–125.

Statistisches Bundesamt (2013). *Bildung und Kultur: Allgemeinbildende Schulen. Korrigierte Fassung vom 13. Februar 2014.* Fachserie 1, Reihe 11, Wiesbaden.

Sterba, S. K. & Bauer, D. J. (2010). Matching method with theory in person-oriented developmental psychopathology research. *Development and Psychopathology, 22,* 239–254.

Stürmer, K., Könings, K. D. & Seidel, T. (eingereicht). Pre-service teachers' capacities for professional vision within university-based teacher education.

Stürmer, K., Könings, K. D. & Seidel, T. (2013). Development of declarative knowledge and professional vision in teacher education: The effect of courses in teaching and learning. *British Journal of Educational Psychology, 83,* 467–483.

Stürmer, K. & Seidel, T. (eingereicht). Assessing professional vision in teacher candidates: Approaches to validate the observer extended research tool.

Stürmer, K., Seidel, T. & Blomberg, G. (2010). „Observe“ – Inhaltliche Validierung eines videogestützten Instruments zur Erfassung professioneller Wahrnehmung mittels „Laut-Denken“ Protokollen von Lehramtsstudierenden. In B. Schwarz, P. Nenninger & R. S. Jäger (Hrsg.), *Erziehungswissenschaftliche Forschung - nachhaltige Bildung. Beiträge zur 5. DGfE-Sektionstagung "Empirische Bildungsforschung"/AEPF-KBBB im Frühjahr 2009* (S. 170–177). Landau: Verlag Empirische Pädagogik.

Südkamp, A., Möller, J. & Pohlmann, B. (2008). Der simulierte Klassenraum: Eine experimentelle Untersuchung zur diagnostischen Kompetenz. *Zeitschrift für Pädagogische Psychologie, 22,* 261–278.

Syring, M., Reuschling, A., Bohl, T., Kleinknecht, M., Kuntze, S. & Rehm, M. (2013). Classroom-Management lehren und lernen: Zur Bedeutung des Konzepts im Unterricht und dessen Vermittlung in fallbasierten Seminaren. In R. Arnold, C. Gómez Tutor & C. Menzer (Hrsg.), *Didaktik im Fokus* (S. 75–91). Baltmannsweiler: Schneider Hohengehren.

Terhart, E. (2007). Erfassung und Beurteilung der beruflichen Kompetenz von Lehrkräften. In M. Lüders & J. Wissinger (Hrsg.), *Forschung zur Lehrerbildung: Kompetenzentwicklung und Programmevaluation* (S. 37–62). Münster: Waxmann.

Terhart, E. (2008). Die Lehrerbildung. In K. S. Cortina, J. Baumert, A. Leschinsky, K., U. Meyer & L. Trommer (Hrsg.), *Das Bildungswesen in der Bundesrepublik Deutschland: Strukturen und Entwicklungen im Überblick* (S. 745–772). Reinbek: Rowohlt.

Terhart, E. (2009). Erste Phase: Lehrerbildung an der Universität. In O. Zlatkin-Troitschanskaia (Hrsg.), *Lehrprofessionalität. Bedingungen, Genese, Wirkungen und ihre Messung* (S. 425–437). Weinheim, Basel: Beltz.

Terhart, E. (Hrsg.). (2000). *Perspektiven der Lehrerbildung in Deutschland: Abschlussbericht der von der Kultusministerkonferenz eingesetzten Kommission*. Weinheim, Basel: Beltz.

Tittle, C. K. (2006). Assessment of teacher learning and development. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2. Vol., pp. 953–980). Mahwah, NJ: Erlbaum.

Tönjes, B., Dickhäuser, O. & Kröner, S. (2008). Berufliche Zielorientierungen und wahrgenommener Leistungsmangel bei Lehrkräften. *Zeitschrift für Pädagogische Psychologie, 22,* 151–160.

Trapmann, S., Hell, B., Weigand, S. & Schuler, H. (2007). Die Validität von Schulnoten zur Vorhersage des Studienerfolgs - eine Metaanalyse. *Zeitschrift für Pädagogische Psychologie, 21,* 11–27.

Treiblmaier, H. (2011). Datenqualität und Validität bei Online-Befragungen. *der markt. Journal für Marketing, 50,* 3–18.

van Buuren, S. (2012). *Flexible imputation of missing data.* Boca Raton, FL: CRC Press.

van Es, E. A. (2009). Participants' roles in the context of a video club. *The Journal of the Learning Sciences, 18,* 100–137.

van Es, E. A. & Sherin, M. G. (2002). Learning to notice: Scaffolding new teachers' interpretations of classroom interactions. *Journal of Technology and Teacher Education, 10,* 571-596.

van Es, E. A. & Sherin, M. G. (2008). Mathematics teachers' "learning to notice" in the context of a video club. *Teaching and Teacher Education, 24,* 244–276.

van Eye, A. (2006). Variablen- und personenzentrierte Forschung. In A. Ittel & H. Merkens (Hrsg.), *Veränderungsmessung und Längsschnittstudien in der empirischen Erziehungswissenschaft* (S. 9–26). Wiesbaden: VS Verlag für Sozialwissenschaften.

van Eye, A. & Spiel C. (2010). Conducting person-oriented research. *Zeitschrift für Psychologie/Journal of Psychology, 218,* 151–154.

Voss, T., Kunter, M. & Baumert, J. (2011). Assessing teacher candidates' general pedagogical and psychological knowledge: Test construction and validation. *Journal of Educational Psychology, 103*, 952–969.

Wang, W.-C. (2000). The simultaneous factorial analysis of differential item functioning. *Methods of Psychological Research Online, 5*, 57–76.

Watt, H. M. G. & Richardson, P. W. (2008). Motivations, perceptions, and aspirations concerning teaching as a career for different types of beginning teachers. *Learning and Instruction, 18,* 408–428.

Weinert, F. E. (2002). Vergleichende Leistungsmessung in Schulen - eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen*. 2. Aufl. (S. 17–31). Weinheim, Basel: Beltz.

Weinert, F. E. (Hrsg.). (2002). *Leistungsmessungen in Schulen* (2. Aufl.). Weinheim, Basel: Beltz.

Weinert, F. E. (2001). Concept of competence: A conceptual clarification. In D. S. Rychen & L. H. Salganik (Eds.), *Defining and selecting key competencies* (pp. 45–65). Seattle/Toronto/Bern/Göttingen: Hogrefe & Huber Publishers.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Winter, M. (2008). Studienstrukturreform in der universitären Lehramtsausbildung: Zum Stand der Umstrukturierung des Lehrerstudiums und zum Studienmodell Sachsen-Anhalts. *Beiträge zur Hochschulforschung, 30,* 82–111.

Wissenschaftsrat. (2006). *Empfehlungen zur künftigen Rolle der Universitäten im Wissenschaftssystem*.

Wright, B. D. (1996). Sample size again. *Rasch Measurement Transactions, 9,* 468.

Wu, M. & Adams, R. J. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Melbourne: Educational Measurement Solutions.

Wu, M., Adams, R. J. & Wilson, M. (1998) Acer ConQuest. Generalised item response modelling software [computer software]. Camberwell, Australia: ACER.

Zieky, M. (1993). Practical questions in the use DIF-statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale: Lawrence Erlbaum Associates.

Zumbach, J. & Reimann, P. (2001). Interpunktionsmanagement und Strukturierung zur Unterstützung komplexer Kooperation und Kollaboration in synchronen Lernumgebungen. In M. Beißwenger (Hrsg.), *Chat-Kommunikation: Sprache, Interaktion, Sozialität & Identität in synchroner computervermittelter Kommunikation. Perspektiven auf ein interdisziplinäres Forschungsfeld.* Stuttgart: Ibidem.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics. Psychometrics* (pp. 45–79). North Holland: Elsevier Science B.V..